Multiway Calibration in 3D QSAR

Applications to Dopamine Receptor Ligands

RIJKSUNIVERSITEIT GRONINGEN

Multiway Calibration in 3D QSAR
Applications to Dopamine Receptor Ligands

PROEFSCHRIFT

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. van der Woude,
in het openbaar te verdedigen op
vrijdag 27 maart 1998
des namiddags te 2.45 uur

door

Jonas Nilsson

geboren op 9 juni 1967
te Kalmar, Zweden

Promotores

Prof. dr. H.V. Wikström
Prof. dr. A.K. Smilde

ISBN 90 367 0863 x
NUGI 746

“Också jag uppskattar vackra rum, gåslever, gammal konjak och samtal med bildade personer. Men jag vill inte jämt kunna njuta av dessa goda ting. Man ska slita ont dessemellan. Den som aldrig är riktigt trött eller riktigt hungrig har gått miste om något väsentligt; han vet inte hur gott också kolbullar kan smaka eller hur obeskrivlig skönt det kan kännas att få sträcka ut sig på hårda marken, krypa in i sovsäcken och somna.”

Sten Selander (ur “Lappland – några sommarströvtåg”)

Till,
mamma och pappa

Promotiecommissie

Prof. dr. T. Liljefors
Prof. dr. S. Clementi
Prof. dr. J.M.F. ten Berge

Paranimfen

Johann Rollén
Göran Svensson

This project has been carried out within the framework of GUIDE (Groningen Utrecht Institute for Drug Exploration)

Cover Design: A projection of the fifth dimension.
Printing: ponsen & loijen bv, Wageningen

An electronic version of this thesis is available on the World Wide Web:
http://www.ub.rug.nl/eldoc/dis/science/j.nilsson

Table of Contents

Chapter 1

Introduction to the Medicinal Chemistry of Schizophrenia

1.1 Schizophrenia 1
1.2 Molecular Biology of Dopamine Receptors 6
1.3 Computer-Assisted Molecular Design 10
1.4 Quantitative Structure-Activity Relationships (QSAR) 20
1.5 References 27

Chapter 2

Introduction to Chemometrics and Statistics

2.1 Introduction 35
2.2 Data Pretreatment 35
2.3 Multiple Linear Regression (MLR) 36
2.4 Multicollinearity 37
2.5 Principal Component Regression (PCR) 38
2.6 Partial Least Squares Regression (PLS) 39
2.7 PARAFAC and Tucker Decomposition 41
2.8 Multilinear PLS Regression (N-PLS) 43
2.9 Multiway Principal Covariate Regression (PCovR) 46
2.10 Model Validation 48
2.11 Variable Selection 49
2.12 References 51

Chapter 3

Design, Synthesis and QSAR of a Series *trans*-1,2,3,4,4a,5,6,10b-Octahydrobenzo[*f*]quinolines with Dopaminergic Affinity

3.1 Introduction 53
3.2 Computational Chemistry 54
3.3 Chemistry 59
3.4 *In vitro* Pharmacology 60
3.5 Results and Discussions 60
3.6 Conclusions 64
3.7 Experimental Section 65
3.8 References 73

Chapter 4

A GRID/GOLPE 3D QSAR Study on a Set of Benzamides and Naphthamides

4.1 Introduction 75
4.2 Molecular Descriptors Generated in the GRID Program 76
4.3 Data Pretreatment 78
4.4 Statistical Tools...79
4.5 D-Optimal Preselection of Variables 80
4.6 Variable Selection Following a FFD Procedure...80
4.7 Results and Discussion 81
4.8 Conclusions...91
4.9 References...91

Chapter 5

Multilinear PLS Analysis with Application to 3D QSAR

5.1 Introduction 93
5.2 Theory 95
5.3 Results 99
5.4 Discussion...103
5.5 Conclusions...106
5.6 Matlab Code for Regression Coefficients 106
5.7 References...108

Chapter 6

A Multiway 3D QSAR Analysis of a Series of (*S*)-N-[(1-Ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides

6.1 Introduction 109
6.2 Theory and Methods..111
6.3 Results 115
6.4 Discussion...122
6.5 Conclusions...124
6.6 References...125

Chapter 7

Multiway Simultaneous Two-Block Analysis with Applications to 3D QSAR

7.1 Introduction 127
7.2 Theory and Methods..128
7.3 PCovR/Tucker Analysis..132
7.4 PCovR/PARAFAC Analysis 135
7.5 Discussion...139
7.6 Conclusions...141

7.7 References....142

Chapter 8

Future Perspectives

Multiway Analysis in Medicinal Chemistry

8.1 Introduction....143
8.2 Example One: Neuropharmacology with Microdialysis....143
8.3 Example Two: Combinatorial Chemistry....145
8.4 Example Three: Neuropharmacology....146
8.5 Conclusions....148
8.6 References....148

Summary....151
Samenvatting....155
Curriculum Vitae....158
Acknowledgements....159

Introduction to the Medicinal Chemistry of Schizophrenia

1

1.1 Schizophrenia

Schizophrenia is one of the most common psychiatric disorders and approximately 1 % of the worlds population suffer from severe symptoms occupying more than half of the beds in psychiatric clinics. Schizophrenia is distributed over the whole population independent of sex, location, social class or color of the skin. A schizophrenic patient is frequently described as a person with a “Dr Jekyll and Mr. Hyde” personality, but the diagnosis of schizophrenia is more complex than that. Generally, the symptoms are divided into two classes: positive (reality distortion) and negative symptoms (psycho-motor poverty syndrome).[1] Each patient is different and could suffer from more or less of positive or negative symptoms. The positive symptoms, *e.g.*, delusions, hallucinations, grandiosity, excitement, hostility and disorganization are more easily identified, as compared to the negative symptoms. Examples of negative symptoms are apathy, attentional impairment, affective blunting, asociality, poverty of speech and anhedonis that may be difficult to distinguish from either depression or side effects caused by medication with typical antipsychotic drugs.[2,3]

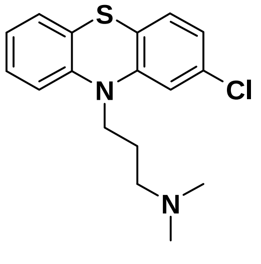
Approximately half of the schizophrenia patients will experience periods with severe depression during the course of their illness. Consequently, the detection of depressive symptoms[4] is very important, since, about 7 %–10 % of all schizophrenia patients commit suicide.[5]

: Treatment of Schizophrenia

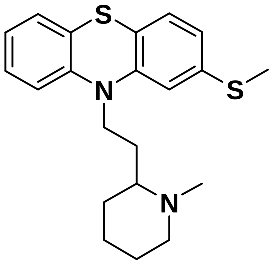
The diagnosis of schizophrenia is just as complex as the medication to suppress the symptoms. There is no real cure against schizophrenia and most patients are bound to medication for the rest of their lifes. The drug of choice is more often a trade-off between clinical efficacy and EPS (Extra Pyramidal Syndrome) or other side-effects. EPS are the side-effects, *e.g.*, major movement disorders elicited by typical antipsychotic drugs. In general, low-potency drugs, *e.g.*, chlorpromazine (**1**) or thioridazine (**2**), are more sedative and hypotensive than high-potency drugs, *e.g.*, fluphenazine (**3**) and haloperidol (**4**) which, in turn, produce more EPS than low-potency agents.

Thus, patients that are highly agitated and excited may be better off with a drug as chlorpromazine. On the contrary, if there is no need for sedation and no history of unusual sensitivity to EPS, high-potency drugs as haloperidol (**4**) or fluphenazine (**3**), are most likely prescribed.

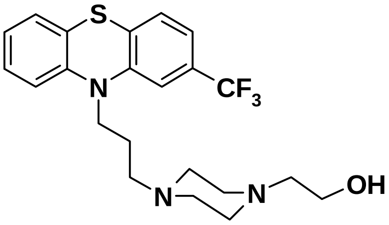
Recently, risperidone (**5**) at fixed doses of 2, 6 and 16 mg/day, has been reported to have higher efficacy and elicit fewer EPS than haloperidol (**4**).[6,7] It was found that risperidone,[7] at daily doses of 6 mg, was more effective than haloperidol and placebo against both negative and positive subscales of PANSS (Positive and Negative Syndrome Scale).[8] At higher doses, no advantage of risperidone over haloperidol was demonstrated. In the same investigation,[7] it was shown that risperidone can suppress TD (Tardive Dyskinesia) but whether it was superior to other typical neuroleptics was not clear.



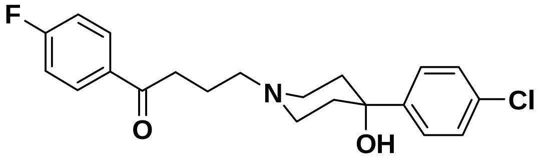
1 chlorpromazine



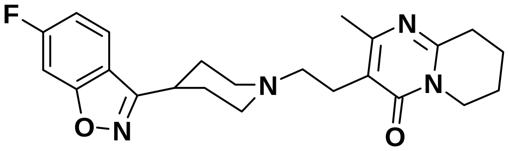
2 thioridazine



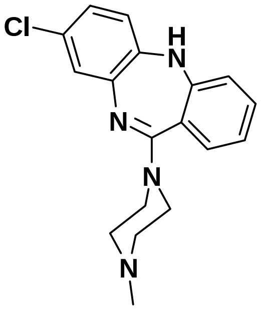
3 fluphenazine



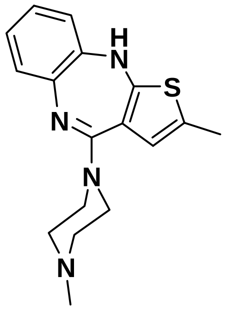
4 haloperidol



5 risperidone



6 clozapine



7 olanzapine, LY170053

Just as risperidone (**5**), clozapine (**6**) belongs to the new generation of antipsychotics often classified as atypical. An atypical antipsychotic drug produce, by definition, fewer EPS than typical antipsychotics and clozapine is the compound used as reference for new antipsychotics. The advantages of clozapine (**6**) over classical antipsychotics are manifold: a) it is effective in the treatment of both positive and negative symptoms[9]; b) it is more effective in treatment-refractory patients[10,11] and c) it produces fewer EPS.[9,12] However, clozapine is also an example of what often is referred to as a "dirty drug" since it has affinity to a large number of different receptors (see Table 1.1). As a consequence, side-effects like hypersalivation (may be due to peripheral drug actions)[13] and weight gain[14] must be considered. In addition, agranulocytosis,[15] a potentially fatal blood disorder has been observed in patients medicated with clozapine.

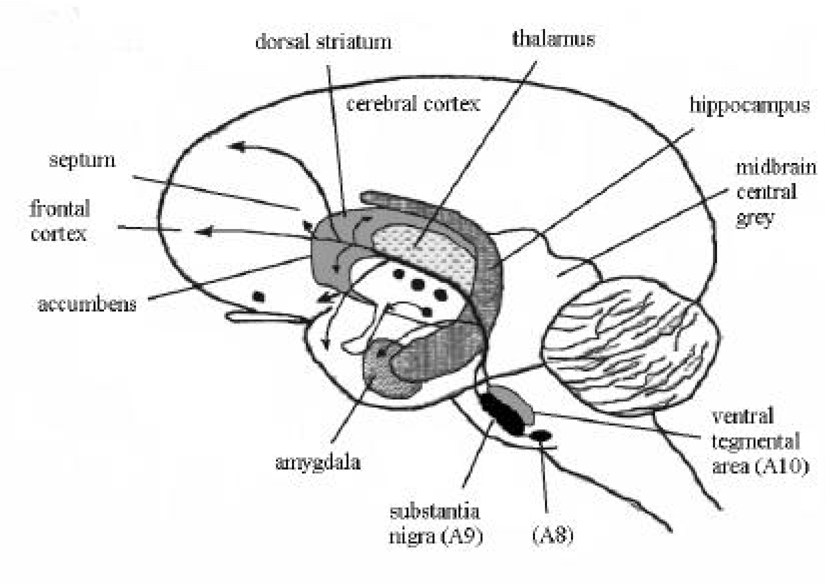


Figure 1.1 The *nigrostriatal* and the *mesolimbic* systems. The presynaptic part of the *nigrostriatal* system originates in A9 and A8 and its axons terminates, mainly, in the forebrain. The *mesolimbic* system, runs parallell to the *nigrostriatal* system, and originates in the VTA (A10) with projections to a number of areas, *e.g.*, *accumbens*, *septum* and *cerebral cortex* (*frontal*, *cingulate* and *enthorinal*).

Olanzapine (**7**, LY170053)[16,17] is a novel atypical antipsychotic with similar binding profile as clozapine (Table 1.1). In studies of schizophrenic patients,[18] olanzapine was demonstrated to be effective in the treatment of both negative and positive symptoms with few EPS. Additionally, the potency of olanzapine in reversing the effects of *d*-amphetamine was greater, as compared to clozapine.[16] (The reversal of the inhibitory effects of *d*-amphetamine on A10 cells [Figure 1.1] has been hypothesized to be predictive of clinical antipsychotic efficacy).[19] These findings are consistent with the fact that olanzapine is a DA D_2 antagonist *in vivo*,[17] and is more potent at DA D_2 receptors[17] *in vitro*, as compared to clozapine (Table 1.1).

Table 1.1 Receptor affinity for a selection of drugs used in the treatment of schizophrenia. All values are K_i (nM).

receptor	1 chlorpromazine	2 thioridazine	3 fluphenazine	4 haloperidol	5 risperidone	6 clozapine	7 olanzapine
D_1	32	32	5	10	75	85	31
D_{2L}	3	8	1	0.5	1.5	60-150	11
D_{2S}				0.5	1.5	35	11
D_3	~4	2		2	7	300	
D_4	34	31		2	7	9-54	27
D_5				27		35-400	
$5HT_{1A}$	3635	643		1714	16	875	>10000
$5HT_{2A}$	7	48	52	74	0.6	8	5
$5HT_{2C}$	12	60	295	5755	16	12	23
$5HT_6$	4	7	17	>5000	425	4	2.5
$5HT_7$	21	70	8	263	1.4	6	104
M_1	25	3		1700	>3000	2	2
M_2	150	14		2500	>3000	21	18
M_3	67	15		>3000	>3000	13	25
M_4	40	9		2700	>3000	12	13
M_5	42	13		1800	>3000	3.7	
H_1				3630	>10000	6	7
α_1	~5	~7		46	>10000	7	19
α_2				360	2904	8	228
β_1				>10000	>10000	>10000	>10000

Binding data are obtained from the following references:17,20-28

The Dopamine Hypothesis of Schizophrenia

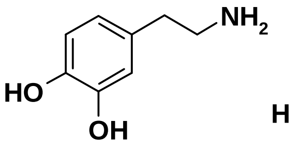
Extensive research during the last decades, have given rise to many different hypotheses over the pathophysiology of schizophrenia. In 1963 Carlsson and Lindquist[29] reported that neuroleptics, *e.g.*, chlorpromazine (**1**), increase dopamine (**8**) and noradrenaline (**9**) turnover in rat brain, and postulated that this effect is caused by blockade of catecholamine receptors. These findings form the basis of the DA hypothesis of schizophrenia. Accordingly, it was found that DA agonists can induce psychosis similar to acute paranoid schizophrenia[30,31] and that neuroleptics inhibit dopaminergic activity.[32-34] Based on the fact that dopamine-mimetic drugs elicit hallucinations, and that other neuroleptics cause rigidity, Van Rossum[35] suggested that schizophrenia may be caused by overactivity in certain dopaminergic pathways.[36] As further support of the DA hypothesis, the clinical doses of neuroleptics and antipsychotics were found to correlate very well with their ability to block DA D_2 receptors.[37,38] Furthermore, the correlations between the clinical efficacy of neuroleptic drugs and the *in vitro* binding affinity of the muscarinic cholinergic, histaminergic (H_1), serotonergic (5-HT_2) and α_1 receptors, were poor.[39]

There are two major dopaminergic neuronal systems that project in the forebrain: the *nigrostriatal* (A9) and the *mesolimbic* (A10) systems (Figure 1.1). Parkinsonism is a consequence of degeneration of neuronal pathways in the A9 system, and EPS induced by treatment with typical antipsychotics is caused by blockade of dopamine receptors in the same system. Consequently, it was postulated that the symptoms of schizophrenia originated from hyperactivity in the *mesolimbic* dopaminergic systems (A10).[34,40]

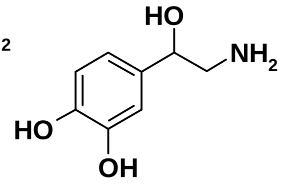
In agreement with the DA hypothesis is the clinical observation that patients with Parkinson's disease do not develop schizophrenia.[38]

: **Serotonin Hypothesis of Schizophrenia**

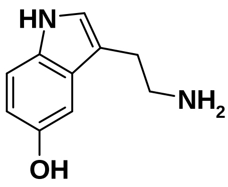
The first indications that serotonin (**10**, 5-HT) might be involved in the pathophysiology of schizophrenia came with the discovery that certain ergots (*e.g.*, lysergic acid diethylamide (**11**)), with structural resemblance to 5-HT, were hallucinogenic and induced many of the symptoms of schizophrenia.[41]



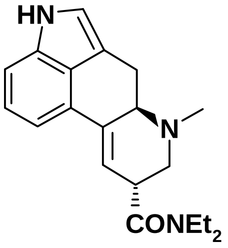
8 dopamine; DA



9 noradrenaline; NA



10 serotonin; 5-HT



11 (+)-LSD

Today, several atypical antipsychotic drugs (*e.g.*, clozapine (**6**), olanzapine (**7**) and risperidone (**5**)) with affinity to one or several serotonin receptor subtypes (Table 1.1) are known. Clozapine for example, the prototypical atypical antipsychotic drug, has been shown to have high affinity towards, at least, four different serotonin receptors including 5-HT_{2A}, 5-HT_{2C}, 5-HT_6 and 5-HT_7.[20,24,26,42] Meltzer *et al.*,[26] however, showed that most putative atypical antipsychotic drugs could be classified by their 5-HT_{2A}/D_2 affinity ratios.

It has been found that full or partial 5-HT_{1A} agonists reverse catalepsy in rat.[43,44] Catalepsy in the rat is predictive for extrapyramidal side-effects in man.[45]

: **Muscarinic Hyperactivity in Schizophrenia**

Tandon *et al.*[46] suggested that hyperactivity of muscarinic cholinergic receptors had a role in the pathogenesis of negative symptoms of schizophrenia. Their observations showed that unmedicated schizophrenics displayed symptoms, (*e.g.*, reduced pain perception, hyper-salivation and increased water intake), which resemble a muscarinic receptor hyperactive state. Occasionally, anti-cholinergic drugs have been reported effective in treating negative symptoms of schizophrenia.[47,48] It was not clear, however, whether the negative symptoms of schizophrenia or the neuroleptic induced EPS, were reduced.

Since the atypical antipsychotic drug clozapine (**6**), has high affinity towards all five muscarinic receptors (Table 1.1), one may predict that the degree of atypicality is related to its cholinergic activity. However, Bolden *et al.*[27] could not find any clear pattern in their investigation. Taken together, according to the investigations of Boldens *et al.*[27] and others,[11,49,50] it is not clear whether anticholinergic activity is essential for an atypical antipsychotic drug or not.

The Noradrenaline Receptor

The relationship between noradrenaline (**9**, NA) and schizophrenia was first studied by Stein *et al.*[51] in 1971. However, significant and reproducible research established a relationship between NA levels in the limbic forebrain and the intensity of the schizophrenic symptoms first in 1990.[52] Recently Breier *et al.*[53] demonstrated a direct correlation between the ability of clozapine (**6**) to elevate plasma NA levels with its ability to improve positive symptoms of schizophrenia. As yet, no selective drugs towards the α_1, α_2 or β receptors with high efficacy in man have been reported.[54]

1.2 Molecular Biology of Dopamine Receptors

Until now, five central dopamine receptor subtypes have been discovered, distributed with the highest concentrations in the *putamen*, caudate nucleus and the *nucleus accumbens*. As can be seen in Figure 1.2, the different dopamine receptor subtypes are not homogeneously distributed in the brain. Instead each subtype is concentrated in specific small areas. Generally, three major dopamine pathways[38,55] are discussed: the *nigrostriatal*, the *mesolimbic* and the *tuberoinfundibular* pathways. The first two pathways (Figure 1.1) control voluntary movement and regulate emotional behavior, respectively. The tuberoinfundibular pathway regulates the secretion of prolactin from the pituitary,[56] and is thus, not mentioned in the context of schizophrenia. The *nigrostriatal* pathway (A9) has cellbodies in the *substantia nigra*, with long axons projecting in the corpus striatum (Figure 1.1). The abundance of DA D_1 and D_2 receptors[57-60] in the *nigrostriatal* system is high while the presence of DA D_3 receptors is very low.[55,61] Instead, high concentration of mRNA for the DA D_3 receptors are found in the limbic areas[55,61,62] (Figure 1.2), suggesting DA D_3 receptors to be involved in emotional and cognitive disorders (*e.g.*, schizophrenia). The *mesolimbic* neuronal pathway (A10) has cellbodies in the *ventral tegmental* area (VTA) of the brainstem, with cells projecting in the *limbic* system (Figure 1.1).

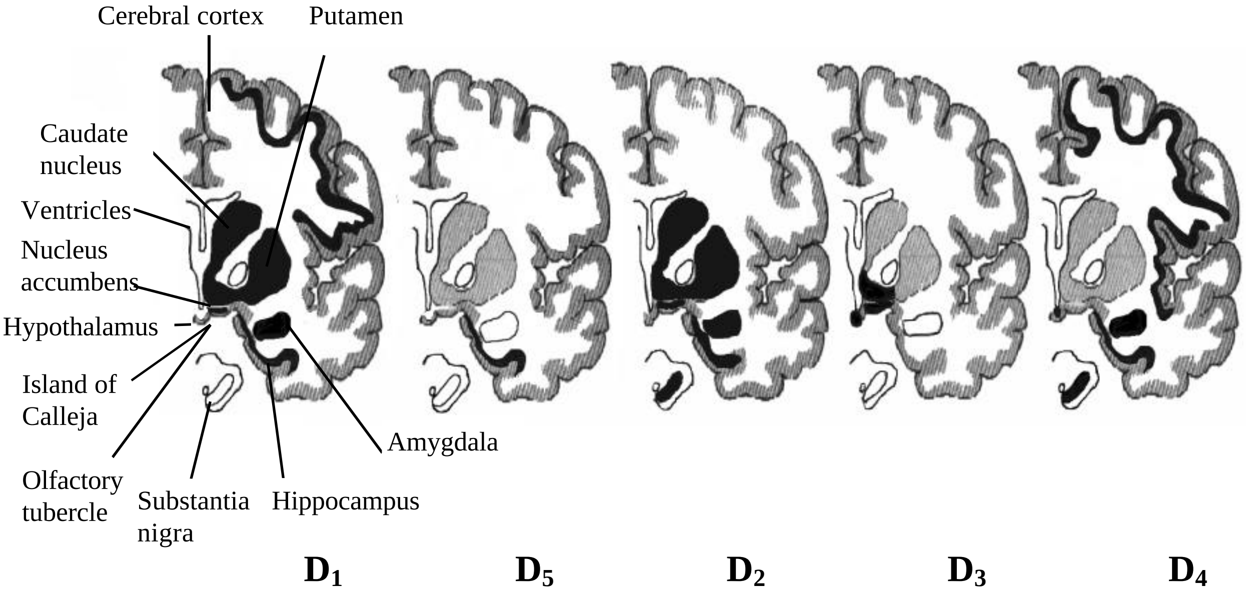


Figure 1.2 The distribution of dopamine receptors in the human brain as determined by the concentrations of mRNA for the respective receptor subtypes in the different brain areas.

G Protein-Coupled Receptors

The dopamine receptors belong to a class of proteins normally referred to as the G protein-coupled receptor (GPCR) superfamily. To date, no X-ray crystallographic structure of a GPCR is resolved, but along with molecular cloning and receptor binding studies the amino acid sequence of all five human DA receptors have been elucidated (Table 1.2). In 1993, Schertler *et al.*[63] provided evidence that the bovine rhodopsins, G protein-coupled receptors active as the photoreceptors in rod cells, were arranged in seven α-helices. In 1990, Henderson *et al.*[64] presented a high quality 3D model of bacteriorhodopsin, also a G protein-coupled receptor, based on cryo-microscopy experiments. Further refinements of this model have recently been published by Grigorieff *et al.*[65], Unger *et al.*[66] and Kimura *et al.*[67] The latter group collected structural data from bacteriorhodopsin crystals at 3.0 Å resolution with 90 % completeness using electron cryo-microscopy. Although the function of bacteriorhodopsin is different from rhodopsin both proteins bind retinal in a similar way,[68] and have similar topology with seven transmembrane helices.

The receptor protein may be folded through the cellular membrane forming seven hydrophobic trans-membrane α-helices connected, alternately, via intra- and extra-cellular loops. The amino terminal (*N*-terminal) and the carboxylic terminal (*C*-terminal) of the receptor protein reside at the extra- and the intracellular sides of the cell-membrane, respectively.

The intrinsic activity of a DA agonist is mediated by a signal transduction across the cellular membrane. That is, the drug-receptor interaction most probably induces a conformational change in the receptor protein which in turn, activates a G protein coupled to the third intracellular loop.[54,69] Accordingly, the activated G protein stimulates (or inhibits) adenylyl cyclase (see Table 1.2) to produce cAMP (a so-called second messenger) from AMP (Figure 1.3), which influences various processes in the cytosol.

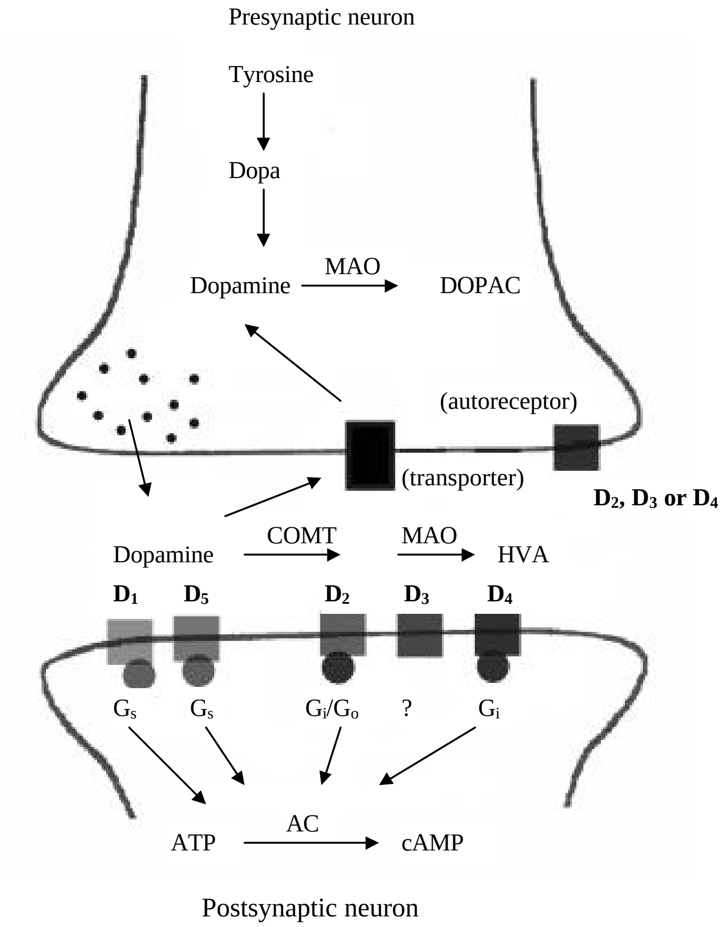
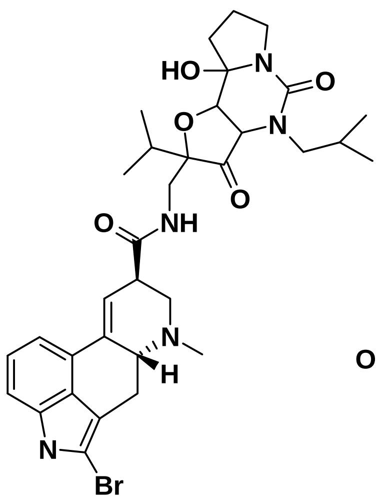
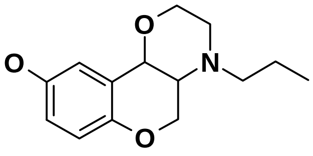


Figure 1.3 Schematic representation of a pre and post-synaptic dopaminergic cell.

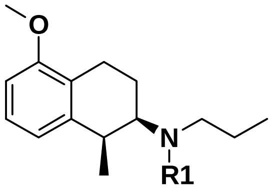
The third intracellular loop exhibit the largest sequence dissimilarities among the different DA receptors. The DA D_1 and DA D_5 receptors have relative short third intracellular loops, are coupled to G_s proteins and have long *C*-terminal tails. These two receptors, the D_1-like receptors, stimulate the activity of adenylyl cyclase and the pharmacological functions from known ligands are more or less, identical.[54,61] The D_2-like receptors, *i.e.*, D_2, D_3 and D_4 receptors on the other hand, all have long third intracellular loops with short *C*-terminal tails, might couple to G_i proteins (or G_0 proteins) and inhibit adenylyl cyclase. Interesting, two forms of the DA D_2 receptor have been found, differing in 29 amino acids in the third intracellular loop,[70-72] and seem to have identical pharmacology but their presence in various cerebral tissues differ.[70,73] Hence, a difference in functionality is likely still to be found. The long and the short forms of the D_2 receptor (*i.e.*, D_{2L} and D_{2S}) consist of 443 and 414 amino acid residues (see Table 1.2), respectively. The genes of the dopamine receptor superfamily can be divided into two different categories: 1) intronless genes that codes for the D_1-like receptors and 2) genes with their coding sequences in discontinuous DNA segments (exons) separated by sequences (introns) that do not form a part of the mature mRNA. The latter category of genes are found in the D_2-like family of dopamine receptors, which explains the occurrence of a long and a short form of the DA D_2 receptor. In the biosynthesis of mRNA a mechanism called alternative splicing, in which a given exon in the pre-mRNA is either present or absent in the final mRNA, result in two different proteins coded by the same gene.[61]



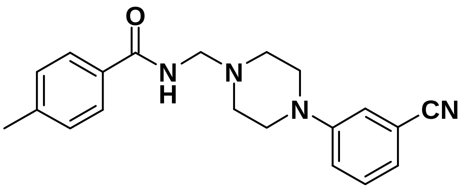
12 bromocriptine



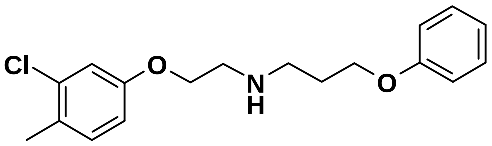
13 PD128907



14 AJ-76 R1 = H
15 UH-232 R1 = n-Pr



16 PD1 D4 agonist



17 PD2 D4 antagonist

Table 1.2 Specifics of the human DA receptors.

	D_1	**D_5**	**D_{2L}/D_{2S}**	**D_3**	**D_4**
# a.a.	446[a]	477[b]	443[c]/414	400[d]	467[e]
mRNA location	neostriatum	hypothalamus, hippocampus	neostriatum	isl. of Calleja, n. accumbens	frontal cortex, hippocampus
adenylyl cyclase	stimulates	stimulates	inhibits	?	inhibits
agonists	SKF38393 (**36**) SKF82958 (**37**)	SKF38393 (**36**)	bromocriptine (**12**)	PD128907(**13**)[f] 7-OH-DPAT (**25**)	PD1(**16**)[g]
antagonists	SCH23390 (**35**)	SCH23390 (**35**)	haloperidol (**4**)	AJ76(**14**)[h] UH232(**15**)[h]	clozapine (**6**) PD2(**17**)[i]

[a] Ref.57,59,60,74; [b] Ref.75-77; [c] Ref.72,78-86; [d] Ref.87-89; [e] Ref.90-93; [f] Ref. 94; [g] Ref.95; [h] Ref. 96; [i] Ref.97

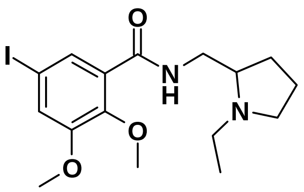
1.3 Computer-Assisted Molecular Design

The last decade new tools have become available for drug design including computational chemistry, high-throughput screening[98,99] and combinatorial chemistry.[100-102] Still, no matter how advanced our technology have developed or how fast new compounds can be synthesizes and tested, a medicinal chemist simply has two major questions to answer: Do I understand the structure activity-relationship for this series of compounds and which compound should I synthesize next? In order to provide answers to these questions two general work procedures are followed. First, one may try to build a 3D model, *e.g.*, homology modeling of the target protein and, accordingly, dock ligands into the active site of the protein. Simply, a potent ligand fits into the receptor while an inactive ligand does not. The second approach is more basic where ligands, initially, are superimposed on mutual tentative interaction points (*e.g.*, lone pair of electrons and midpoints of aromatic rings) with the receptor. Consequently, attempts to explain the potency of the ligands by comparison of their structures and their relative 3D orientations can be done. This approach is generally called an "active analogue approach"[103] or structure-activity relationship (SAR).

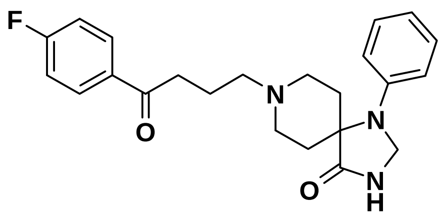
: Homology Modeling

Until recently, we had no knowledge about the amino acid sequence and less knowledge about the secondary structure of G protein-coupled receptors (GPCRs). Therefore, computational chemists have utilized the structure elucidated for bacteriorhodopsin (see above), although the sequence homology is poor, to create 3D models of GPCRs.[92,104-107] Recently, the dopamine D_2 receptor was constructed[68] based on the coordinates from bacteriorhodopsin[108] itself.

The first problem in GPCR modeling is to determine which amino acid residues that reside in the transmembrane domains, or stated differently, the alignment of the seven helices. Computer programs[107,109,110] that can perform this kind of alignment, are available. The dopamine receptors presented in Table 1.3 were downloaded from EMBL in Heidelberg[111] and aligned using the 'whatif' program.[112] The alignments were by no means perfect and refinements were applied manually as presented in Table 1.3.



18 (*S*)-epidepride



19 spiperone

Dopamine agonists are believed to interact with the third and the fifth transmembrane domain, while antagonists additionally interact with the seventh domain as reviewed by Savarese *et al.*[113] and Teeter *et al.*[68] The endogenous neurotransmitter, dopamine (**8**), most likely interacts with three amino acid residues[68,107] located at helices three and five. The protonated nitrogen forms a saltbridge with Asp$_{114}$ on helix three while, simultaneously, hydrogen bonds are formed between the *m*- and *p*-OH and Ser$_{194}$ and Ser$_{197}$ on helix five, respectively.

Table 1.3 Transmembrane domain amino acid sequences from the five dopamine receptor proteins. Alignment was originally taken from EMDL but was refined, manually, to obtain maximum sequence overlap between the receptors.

TM1																									
D_2		A	T	L	L	T	L	L	I	A	V	I	V	F	G	N	V	L	V	C	M	A	V	S	
D_3		A	L	S	Y	C	A	L	I	L	A	I	V	F	G	N	G	L	V	C	M	A	V	I	
D_4		L	V	G	G	V	L	L	I	G	A	V	L	A	G	N	S	L	V	C	V	S	V	A	
D_1		A	C	F	L	S	L	L	I	L	S	T	L	L	G	N	T	L	V	C	A	A			
D_5		A	C	L	L	T	L	L	I	I	W	T	L	L	G	N	V	L	V	C	A	A			

TM2										Asp$_{80}$															
D_2		L	I	V	S	L	A	V	A	D	L	L	V	A	T	L	V	M	P	W	V	V	Y	L	E
D_3		L	V	V	S	L	A	V	A	D	L	L	V	A	T	L	V	M	P	W	V	V	Y	L	E
D_4	S	V	I	V	S	L	A	A	A	D	L	L	L	A	L	L	V	L	P	L	F	V	Y		
D_1	F	F	V	I	S	L	A	V	S	D	L	L	V	A	V	L	V	M	P	W	K	A	V	A	E
D_5	V	F	I	V	S	L	A	V	S	D	L	F	V	A	L	L	V	M	P	W	K	A	V	A	E

TM3							Asp$_{114}$																		
D_2		I	F	V	T	L	D	V	M	M	C	T	A	S	I	L	N	L	C	A	I	S	I		
D_3		V	F	V	T	L	D	V	M	M	C	T	A	S	I	L	N	L	C	A	I	S	I		
D_4		A	L	M	A	M	D	V	M	L	C	T	A	S	I	F	N	L	C	A	I	S	V		
D_1	N	I	W	V	A	F	D	I	M	C	S	T	A	S	I	L	N	L	C	V	I	S	V		
D_5		V	W	V	A	F	D	I	M	C	S	T	A	S	I	L	N	L	C	V	I	S	V		

TM4																									
D_2		V	T	V	M	I	S	I	V	W	V	L	S	F	T	I	S	C	P	L	L	F	G	L	
D_3		V	A	L	M	I	T	A	V	W	V	L	A	F	A	V	S	C	P	L	L	F	G	F	
D_4		Q	L	L	L	I	G	A	T	W	L	L	S	A	A	V	A	A	P	V	L	C	G	L	N
D_1	A	A	F	I	L	I	S	V	A	W	T	L	S	V	L	I	S	F	I	P	V	Q	L	S	W
D_5				V	M	V	G	L	A	W	T	L	S	I	L	I	S	F	I	P	V	Q	L	N	W

TM5									Ser$_{194}$			Ser$_{197}$	Phe$_{198}$												
D_2		P	A	F	V	V	Y	S	S	I	V	S	F	Y	V	P	F	I	V	T	L	L	V	Y	I
D_3		P	D	F	V	I	Y	S	S	V	V	S	F	Y	L	P	F	G	V	T	V	L	V	Y	A
D_4				Y	V	V	Y	S	S	V	C	S	F	F	L	P	C	P	L	M	L	L	L	Y	W
D_1			T	Y	A	I	S	S	S	V	I	S	F	Y	I	P	V	A	I	M	I	V	T	Y	T
D_5			T	Y	A	I	S	S	S	L	I	S	F	Y	I	P	V	A	I	M	I	V	T	Y	T

TM6																			Phe$_{390}$						
D_2		M	L	V	A	I	V	L	G	V	F	I	I	C	W	L	P	F	F	I	T	H	I	L	N
D_3			M	V	A	I	V	L	G	A	F	I	V	C	W	L	P	F	F	L	T	H	V	L	
D_4			V	L	P	V	V	V	G	A	F	L	L	C	W	T	P	F	F	V	V	H	I		
D_1			T	L	S	V	I	M	G	V	F	V	C	C	W	L	P	F	F	I	L	N	C	I	L
D_5			T	L	S	V	I	M	G	V	F	V	C	C	W	L	P	F	F	I	L	N	C	M	V

TM7												Tyr$_{416}$													
D_2		V	L	Y	S	A	F	T	W	L	G	Y	V	N	S	A	V	N	P	I	I	Y	T	T	F
D_3						A	T	T	W	L	G	Y	V	N	S	A	L	N	P	V	I	Y	T	T	F
D_4	P	R	L	V	S	A	V	T	W	L	G	Y	V	N	S	A	L	N	P	V	I	Y	T		
D_1				F	D	V	F	V	W	F	G	W	A	N	S	S	L	N	P	I	I	Y	A	F	N
D_5				F	D	V	F	V	W	F	G	W	A	N	S	S	L	N	P	V	I	Y	A		

A Ala; C Cys; D Asp; E Glu; F Phe; G Gly; H His; I Ile; K Lys; L Leu; M Met; N Asn; P Pro; Q Gln; R Arg; S Ser; T Thr; V Val; W Trp; Y Tyr;

TM3: Asp$_{114}$, TM5: Ser$_{194}$ Ser$_{197}$	(white rectangles)[68,107,116]
TM6: Phe$_{390}$, TM5: Phe$_{198}$, TM7: Tyr$_{416}$	(light grey rectangles)[107]
TM2: Asp$_{80}$	(dark grey rectangle)[61,68,73,107]

From Table 1.3 it is also clear that these amino acid residues are preserved in all five receptor subtypes, supporting this hypothesis. (In order not to confuse, the numbering of the amino acid residues starts with number one from the extra-cellular end, and is valid for the DA D_{2L} receptor if not stated otherwise.) Trump-Kallmeyer *et al.*[107] also suggested that the OH-group from Tyr$_{416}$ on helix seven helps to stabilize the transmitter-receptor complex by interacting with the charged dopamine nitrogen. Another interesting feature is the narrow aromatic cleft, defined by Phe$_{390}$ on helix six and Phe$_{198}$ on helix five, that may interact with the flat aromatic part of catecholamine related ligands.[107] Finally, site-directed mutagenesis study have shown[73,114,115] the importance of Asp$_{80}$ in the regulation of D$_2$ affinity for drugs, coupling to adenylate cyclase and sensitivity to Na$^+$ and pH. Sodium decreases the binding of D$_2$ agonists (*e.g.*, dopamine (**8**))[114,115] and increases binding for some substituted benzamides (*e.g.*, epidepride (**18**) and sulpiride (**40**))[114,115] but does not affect binding of other D$_2$ antagonists (*e.g.*, spiperone (**19**)).[68,115] This amino acid residue, Asp$_{80}$, is preserved in all five dopamine receptors (see Table 1.3) and Teeter *et al.*[68] called it the 'sodium site'.

: Structure-Activity Relationships (SAR)

Although homology modeling is a direct consequence of recent analytical refinements (*e.g.*, cloning, crystallization and X-ray techniques) and increased computer efficiency the technique is not, as yet, able to explain the structure-activity relationship (SAR) for most ligands. Computer models of receptors cannot, for instance, account for the obvious flexibility of the receptor protein, nor predict the conformational change of the receptor caused when an agonist binds to the active site. There are, simply, to many uncertain parameters that we cannot simulate. Therefore, traditional methods where the activity and inactivity of ligands are explained by superimposition of mutual and possible interaction points with the receptor, are as valid today as for ten years ago. This approach is, generally, referred to as "the active analogue approach",[103] upon which many recent computational methods rely (*e.g.*, CoMFA[117]). During the last two decades several attempts to construct models that can explain the SAR of dopamine receptor ligands have been presented. In the following, a review of the most successful agonist and antagonist models will be given, and the SAR of dopaminergic compounds will be discussed.

: The McDermed Dopamine Receptor Concept

One of the first dopamine receptor models was presented by McDermed *et al.*[118] in 1979, and was two dimensional (Figure 1.4). The model was based on two tentative interaction points with the receptor: one for the basic nitrogen atom and one for the hydroxyl group *meta* to the ethylamine chain. Additionally, a steric boundary defined the receptor excluded volume and explained the low affinity for ligands possessing steric bulk (*i.e.*, a substituent or a part of the molecule) protruding into this region (see below). McDermed and coworkers rationalized their model on the fact that dopamine receptor agonists like **20** ((6aR)-apomorphine) and **22** ((2R)-5,6-di-OH-dipropylaminotetralin) have the dopamine moiety in its α-rotameric conformation, while dopamine receptor agonists like **21** ((6aS)-isoapomorphine) and **23** ((2R)-6,7-di-OH-dipropylaminotetralin) have the dopamine moiety in the β-rotameric conformation, and have to be flipped and rotated in order to fit properly into the presumed active site, as illustrated in Figure 1.4. The same model could rationalize the inactivity of **21** since one of the aromatic rings protrudes into the steric boundary and render binding to the tentative active site impossible (Figure 1.4).

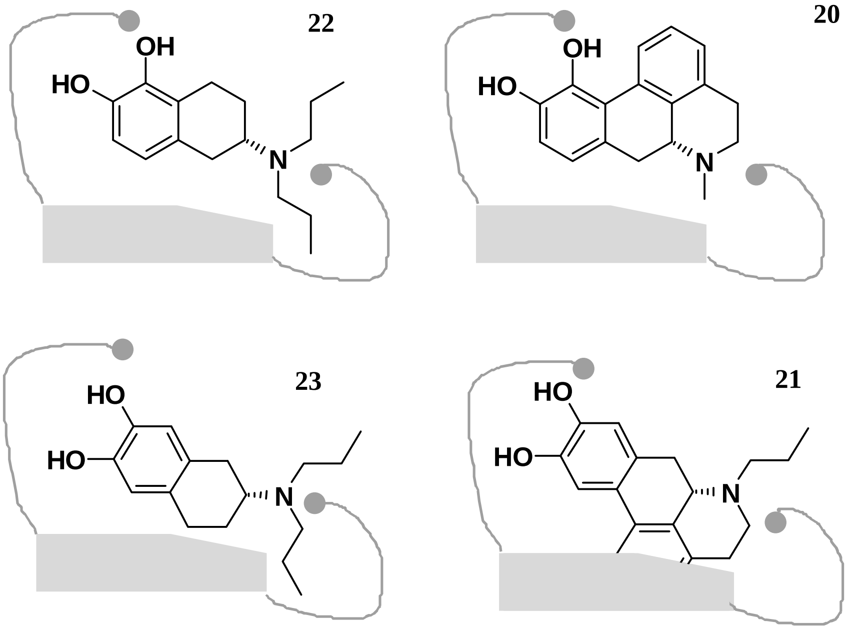
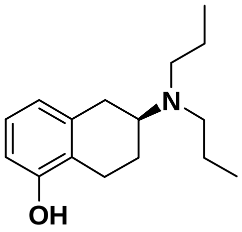
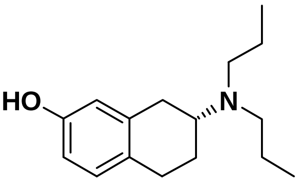
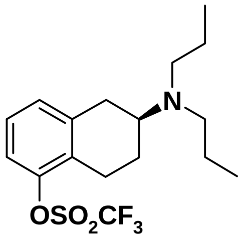
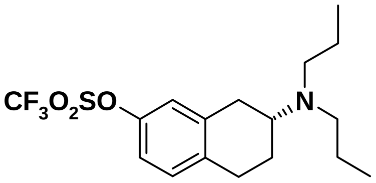


Figure 1.4 The McDermed receptor concept defined by two tentative interaction points with the receptor and a steric boundary preventing binding with "bulky" ligands.









24 (2S)-5-OH-DPAT **25** (2R)-7-OH-DPAT **26** (2S)-5-OTf-DPAT **27** (2R)-7-OTf-DPAT

The mono-hydroxy DPATs[*], **24** and **25**, exhibit the same stereo-isomeri as **22** and **23**, with the S-enantiomer of the 5-OH-DPAT (the α-rotameric conformation) and the *R*-enantiomer of the 7-OH-DPAT (the β-rotameric conformation) being the more potent isomers.[119] Both conformers display affinity to the DA D_{2L} and the D_3 receptors.[120] However, the α-conformer (**24**) display less preference for the DA D_3 receptor as compared to the β-conformer (**25**), with a D_{2L}/D_3 ratio of 26 and 60, respectively. No preference for the α- or β-conformer was observed for the DA $D_{4.2}$ receptor.[120] The triflate analogues, **26** and **27**, displayed similar dopamine agonist profiles as their hydroxy counterparts, **24** and **25**, both in *in-vitro* binding assays and in *in-vivo* biochemical and behavioral assays in rat.[121] The *in vitro* affinity towards the DA receptors were lower for the triflated compounds, as compared to the hydroxy compounds. Interestingly, **26** was found to have mixed

---

[*] DPAT is short for di-(n-propyl)-aminotetralin.

DA/5-HT_{1A} properties after oral administration not observed after subcutaneous administration. This suggest that active metabolite(s) may be formed. (Similar findings were found for 8-OTf-DPAT, a potent 5-HT_{1A} receptor agonist where the dominating metabolite, the mono-propyl analogue, turned out to be more potent *in vivo* than 8-OTf-DPAT itself.[121])

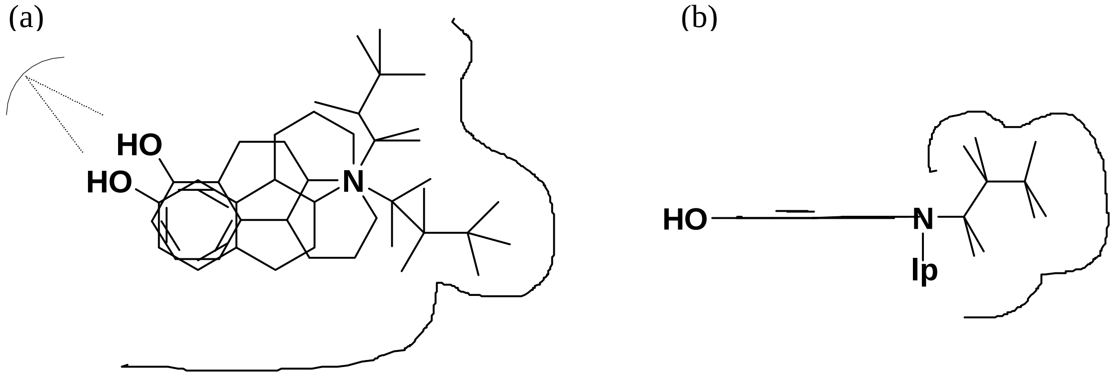
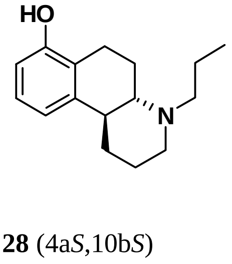


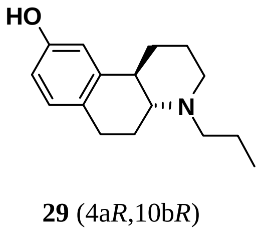
Figure 1.5 The Extended McDermed model as presented by Liljefors *et al*. (a) Molecules **28** and **29** superimposed with the receptor excluded volume marked. (b) The same superimposition as in (a) but now flipped 90 degrees forward. For clarity, only hydrogens on the propyl groups are shown.

: The Extended McDermed Receptor Model

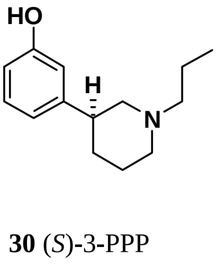
Liljefors[122,123] and Wikström[124] extended the McDermed dopamine receptor concept to a receptor model able to explain the activity and inactivity of a larger number of structurally related compounds. Initially, two modes of receptor interaction and two *N*-alkyl directions were defined, by superimposition of the nitrogen atoms and the midpoints of the aromatic rings of compounds **28** and **29** in their calculated lowest energy conformations[122] (Figure 1.5). The aromatic rings were constrained to be coplanar.



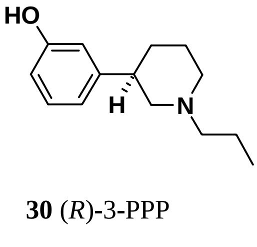
28 (4a*S*,10b*S*)



29 (4a*R*,10b*R*)



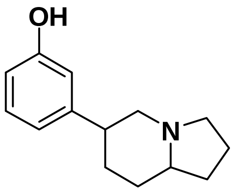
30 (*S*)-3-PPP



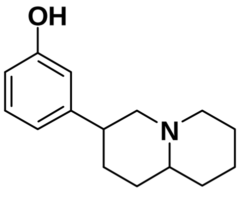
30 (*R*)-3-PPP

The sizes and orientations of the *N*-alkyl substituents are, according to the authors,[122,123] of crucial importance in order to understand the dopaminergic activity and presynaptic selectivity. The model in Figure 1.5(a) defines one “upward” and one “downward” direction for the *N*-alkyl substituents. The “downward“ direction is a narrow cleft complementary to maximum a n-propyl group, while the “upward” direction is sterically less restricted.[122] The model was rationalized since compounds **20** and **29**, both having a *N*-n-propyl substituent, are potent pre- and postsynaptic agonists and display high enantioselectivity. The *N*-n-butyl analogues of the same compounds were found inactive at both pre- and post-synaptic receptors,[122,124] most likely due to the fact their *N*-n-butyl substituents are too large to fit into the "downward" propyl-cleft. The potency of the *N*-n-butyl analogue[125] of **28**, could be explained since its *N*-n-butyl protrudes "upwards" in the less sterically restricted direction when aligned properly into the "active site" (Figure 1.5(a)). In fact, compounds with as large substituents as phenylethyl or thiopenethyl groups directed "upwards" may still be active.[122,126,127]

Liljefors *et al.*[122,123] also investigated the biological active conformations of the enantiomers of 3-PPP, *i.e.*, compounds (*S*)-**30** and (*R*)-**30**. Pharmacologically, the (*R*)-**30** enantiomer displays classical pre- and postsynaptic receptor agonist properties, while the (*S*)-**30** enantiomer is a presynaptic agonist with postsynaptic antagonistic properties (Table 1.4). In an attempt to explain the opposed profiles of the enantiomers (*R*)-**30** was fitted in the model with the *N*-n-propyl directed in the "downward" propyl-cleft and (*S*)-**30** with the *N*-n-propyl directed "upwards". The superimposition of (*R*)-**30** on (4a*R*,10b*R*)-**29** exerts an excellent fit explaining the pre- and postsynaptic properties of (*R*)-**29**. The (*R*)-**30** does not fit in its "global energy minimum conformation" but the *N*-n-propyl directed in the lipophilic "propyl-cleft" helps to stabilize the compound in the agonist conformation. Liljefors and coworkers[122] found and defined two different conformations for compound (*S*)-**30** to take: one agonist and one antagonist conformation. The requirement to activate the postsynaptic receptors seems to be a demand for lipophilicity around the nitrogen. The (*S*)-**30** has no n-propyl directed into the "propyl-cleft". Obviously, the methylene group alfa to the nitrogen in the piperidine ring that is directed into the "propyl-cleft" is not lipophilic enough to maintain the (*S*)-**30** in a postsynaptic activating conformation. Thus, (*S*)-**30** does not activate postsynaptic receptors. In compound (4a*R*,10b*S*)-**28**, however, the methylene group is maintained in the "propyl-cleft", since the OHB[*f*]Q skeleton is rigid, and the postsynaptic receptors can be activated. The (*S*)-**30** can, as well, assume a low energy conformation that fits into the model in Figure 1.5 and explain its presynaptic agonist properties.



31



32

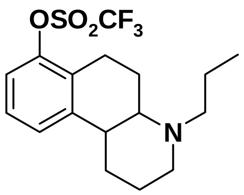
The model in Figure 1.5(a) cannot explain the inactivity of compounds **31** and **32** since the van der Waal volume of **30** fits perfectly while the volume from **32** is too large. The orientations of the propyl groups in Figure 1.5 which can be oriented either in an anti or a gauche conformation with respect to the nitrogen lone pair of electrons provides an explanation. Liljefors *et al.*[123] concluded that if one propyl assumes an anti conformation the other one must be gauche and *vice versa*. If the "downward" oriented propyl group is oriented in an anti conformation (Figure 1.5(b)) the steric boundary in front of the nitrogen atom becomes more narrow as compared to in Figure 1.5(a) and the propyl cleft is located above the plane (Figure 1.5(b)). In this improved model, Figure 1.5(b), neither of the inactive compounds **31** and **32** fit while the active compounds in Table 1.4 do.

Table 1.4 Intrinsic activity of the ligands discussed

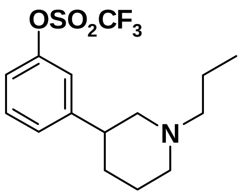
	presynaptic agonism ED$_{50}$,[a] nmol/kg		postsynaptic agonism motor activity[b]	
compound	limbic	striatum	dose μmol/kg sc	acc.counts/30 min[g]
(*R*)-**20**[c,d]	190	220	2.3	361 ± 42
(*S*)-**24**[d]	3.7	3.7	0.31	155 ± 27
(*R*)-**25**[d]	9.5	11	0.31	46 ± 18
(*S*)-**26**[e]	830	1100	12.5	785 ± 123
(*R*)-**27**[e]	I	I	50	2153 ± 650
(4a*S*,10b*S*)-**28**[d]	14	14	1.30	62 ± 11
(4a*R*,10b*R*)-**29**[c]	4	5	1.06	155 ± 32
(*S*)-**30**[c]	800	1700	213	12 ± 2
(*R*)-**30**[c]	1000	1300	13	78 ± 14
33[f]	-	-	50	211±63
34[f]	-	-	100	290±85

[a] Measured indirectly as inhibition of DA synthesis rate (see ref 124); [b] Motor activity measured in motility meters on raserpinized rats (see ref 124); [c] Data taken from ref 122; [d] Data taken from ref 124; [e] Data taken from ref 128; [f] Data taken from ref 129; [g] Values expressed as percentage of saline controls; mean ± SEM.

It was demonstrated for the triflated aminotetralines (*e.g.*, 8-OTf-DPAT and **24**) that the triflate group induced biochemical changes as compared to their hydroxy analogues (see above). This was confirmed also for the OHB[*f*]Qs in experiments performed by Sonesson *et al.*[121,129] They found that compound (±)-**33** was inactive as an agonist, even at high doses (50 μmol/kg), although the hydroxyl analogue ((±)-**28**) is a potent agonist. Instead, presynaptic DA receptor antagonistic properties was demonstrated for (±)-**33**, by the increase of DOPA levels in nonpretreated habituated rats.[129] Additionally, (±)-**33** also decreased significantly the locomotor activity to 56 ± 4 %. Obviously, (±)-**33** is a compound with postsynaptic agonistic and presynaptic antagonistic properties. The more flexible analogue, compound (±)-**34**, did not portray the same affinity for postsynaptic receptors as (±)-**33**.



33 (±)-trans-7-OTf-OHB[f]Q



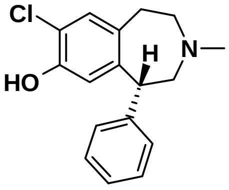
34 (±)-OTf-3-PPP

The triflate group has, obviously, great impact on the phenyl ring due to its electron withdrawal ability[130] distorting the conjugated aromatic system. However, information concerning the physicochemical properties of the (aryl-)triflate group is very sparse, and to date, no X-ray crystallographic structure of a (aryl-)triflate group has been resolved. In Chapter 3, the triflate group will be discussed further, also in comparison with other sulfonyl esters.

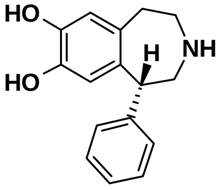
Dopamine D_1 Agonist and Antagonist Models

The McDermed receptor concept is an example of an early working model of the dopamine receptor proven useful in the design of new potent ligands. Today, however, with five different dopamine receptor subtypes known and sophisticated molecular modeling tools available we aim for models that enable us to explain and understand the structure-activity relationships within as well as in-between different receptor subtypes.

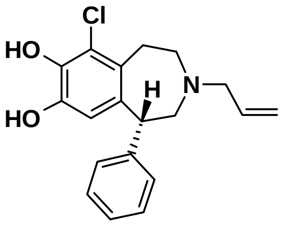
The SAR of the DA D_1 receptor subtype has extensively been scrutinized in the literature during the last decade.[131-133] A challenge has been, and still is, to fully explain the ligand-receptor interactions of the potent benzazepines. To date, no theory exist that explains why compound **35** (SCH23390) is a selective D_1 antagonist while the structurally similar compound, **36** (SKF38393), is a potent agonist at DA D_1 receptors. Compound **35** has a selectivity for the DA D_1 over DA D_2 receptors with a factor 2093,[131] and fully inhibits dopamine stimulated adenylyl cyclase[131] ($K_i = 0.47 \pm 0.06$ nM). In contrast, compounds **36**,**37** and **38** are reported as potent and selective agonist for the D_1 receptors, all able to activate adenylyl cyclase to the same extent as dopamine (**8**).[133]



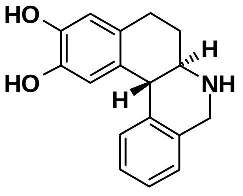
35 (*R*)-SCH23390



36 (*R*)-SKF38393



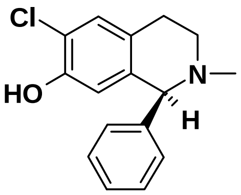
37 (*R*)-SKF82958



38 (6a*R*,12b*S*)-dihydrexidine

Pettersson *et al.*[134,135] have through extensive conformational analysis and electrostatic potential calculations proposed biologically active conformations for **35** and **36**. Both compounds were proposed to have their seven rings in chair conformations with the *N*-methyl (or *N-H*) and the 1-phenyl rings in (pseudo-)equatorial positions. Additionally, the plane of the 1-phenyl rings do not deviate more than 30 degrees from being orthogonal to the plane of the catechol aromatic ring in the main skeleton, since the energy penalty for such rotations would be too large.[134] In the case of compound **38**[133] and other rigid compounds[134] deviations larger than 30 degrees are possible (*i.e.*, an accessory phenyl ring closer to coplanar with the catechol phenyl ring), since no energy penalties are involved. As further support of this hypothesis, probing of the electrostatic surroundings of **35** by means of the GRID program[136] with two different probes (*e.g.*, the cationic NH_3 probe and the anionic carboxy oxygen probe), indicates that the hydroxyl group and the accessory phenyl ring may interact with the same receptor site via electrostatic interactions.[135]

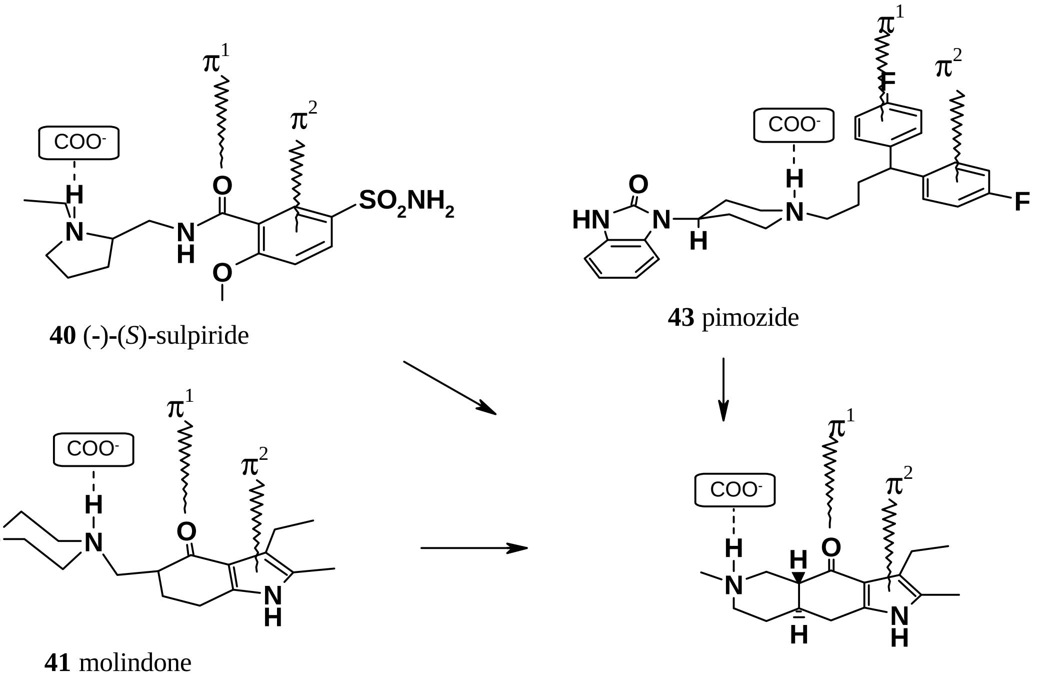
Mottola *et al.* [133] suggested the following agonist pharmacophore derived from the analogues of benzazepine and **38**: 1) the two catechol hydroxyl groups; 2) the nitrogen atom (ca. 7 Å from the *m*-hydroxyl) and 3) the accessory phenyl ring (ca. 5 Å from the catechol ring). Despite the calculations performed by Pettersson *et al.*[134] Mottola *et al.*[133] defined the accessory phenyl ring, in their pharmacophore, to be close to coplanar with the catechol ring. Finally, Mottola *et al.* conclude that alkylation on the nitrogen diminishes the affinity for DA D_1 receptors, defining a steric receptor boundary in that direction.



39

Charifson and coworkers[131,132] proposed a pharmacophore for DA D_1 antagonists, very similar to the agonist pharmacophore of Mottola *et al.* but with the second hydroxyl group replaced by a chlorine atom, derived from analogues of **35** and the tetrahydroisoquinoline **39**. Also in the tetrahydroisoquinoline series elongation of the nitrogen alkyl (*i.e.*, longer than methyl) group was not favorable for DA D_1 receptor binding, although the interatomic distance between the chlorine and the nitrogen is decreased as compared to the benzazepines.[132] Additionally, the stereochemistry for the tetrahydroisoquinolines is reversed as compared to the benzazepines; both (*R*)-SCH23390 (**35**) and (*S*)-**39** are potent and selective DA D_1 antagonists.[132]

: Dopamine D_2 Antagonist Pharmacophores



40 (-)-(*S*)-sulpiride

43 pimozide

41 molindone

42 (-)-(4a*R*,8a*R*)-piquindone

Figure 1.6 The rational for Olson et al. to synthesize (–)-(4a*R*,8a*R*)-piquindone (**42**): Assemble the tentative pharmacophoric elements, *i.e.*, the benzamide carbonyl group, the basic nitrogen of **40** and **41** and the benzamide aromatic moiety, in one rigid skeleton.

Among dopamine antagonist, benzamides are generally characterized by a high selectivity for the DA D_2 receptor subtype with low affinity for the DA D_1 receptor and other non-dopamine receptors. Additionally, the pharmacological profile of benzamides in general is unique, with low propensity to induce neurological side effects (*e.g.*, extrapyramidal syndromes and tardive dyskinesia)[137] and effective in the treatment of negative symptoms in schizophrenia. Thus, it seems a lot to gain by learning about the SAR of benzamides. In 1981 Olson *et al.*[138] proposed a pharmacophore based on potent, but flexible, dopaminergic compound such as sulpiride (**40**), haloperidol (**4**), pimozide (**43**) and molindone (**41**). The pharmacophore comprised four different elements as pictured in Figure 1.6 (adopted from Olson *et al.*[138]): 1) the benzamide phenyl ring (π-π interaction); 2) the amide carbonyl oxygen atom (π-π interaction) or, alternatively, an aromatic moiety (π-π interaction); 3) the basic nitrogen atom ($^{+}NH^{\cdots}COO^{-}$ interaction) and 4) a lipophilic tail corresponding to the heterocyclic moiety of pimozide (Figure 1.6, **43**). Interestingly, the authors introduce the carbonyl oxygen as an isosteromer to a phenyl ring in the ligand-receptor interaction, supported by a crystal structure where a similar interaction has been observed.[138]

In order to validate their model Olsen *et al.*[138] attempted to include the pharmacophoric elements in one single structure, and the resulting (–)-(4a*R*,8a*R*)-piquindone (**42**) turned out to have similar properties as haloperidol (**4**) but with significant decreased cataleptogenic liability. For a computational chemist a rigid and potent compound (Table 1.5), like **42,** is ideal to utilize as a starting point for molecular modeling and SAR studies. Accordingly, Rognan *et al.*[137] used (–)-(4a*R*,8a*R*)-piquindone (**42**) for elucidation of the SAR from a number of optically active DA D_2 antagonists. From the crystallographic structure of (–)-**42** three initial pharmacophoric elements were defined (Figure 1.7): an aromatic ring Ar_1 (pyrrole); a dipole coplanar to Ar_1 (amide carbonyl); a basic nitrogen atom with a lone pair of electrons directed orthogonal to the Ar_1 plane towards a dummy atom Du representing the receptor site (*e.g.*, an aspartic acid). The aromatic pharmacophoric element Ar_2 was defined by superimposition of low energy conformations of domperidone (**44**) with the crystallographic structure of (–)-**42** on the chlorophenyl ring (from domperidone), the pyrrole ring (from piquindone) and the Du in the direction of the nitrogen lone pair of electrons, respectively. In one high quality fit the second phenyl ring of **44** is protruding into and defines the Ar_2 pharmacophore element (Figure 1.7).

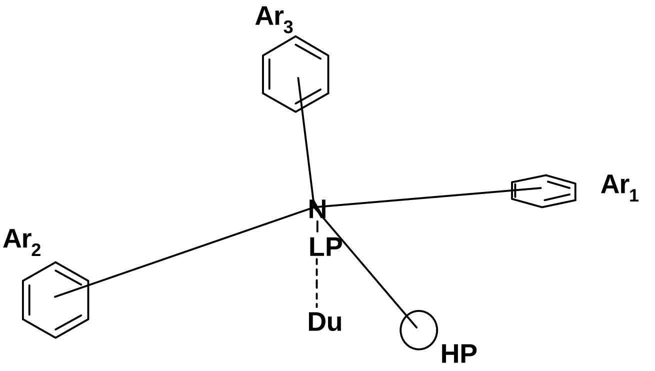
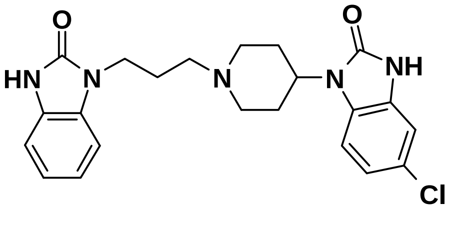
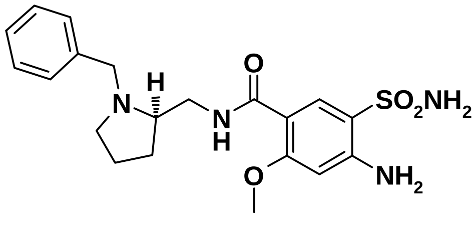


Figure 1.7 The dopamine antagonist pharmacophore as defined by Rognan *et al.* Ar_{1-3} are aromatic moieties, the Du is a dummy atom in the direction of the lone pair of electrons and HP is a hydrophobic pocket.



44 domperidone



45 (*R*)-DO 749

Table 1.5 Affinities of the discussed antagonists for the dopamine D_2 receptor subtype

Compound	**K_i (nM)[a]**
1 chlorpromazine	4.1
4 haloperidol	1.8
19 spiperone	0.04
6 clozapine	220
40 (*S*)-sulpiride	7.6
42 (4a*R*,8a*R*)-piquindone	4.1
43 pimozide	2.5
44 domperidone	0.7
45 (*R*)-DO749	2.4

[a] Data from reference 137 obtained by inhibition of [^{125}I]iodosulpiride to rat striatal membranes.

Another interesting feature of this series of compounds is the stereo chemistry. In benzamides with short substituents (*e.g.*, ethyl, propyl and allyl) attached to the pyrrolidine nitrogen atom the activity resides solely, in the (*S*)-enantiomer[137,139] (see Table 1.5, (*S*)-sulpiride). If, however, the substituent is replaced by a benzyl group (**45**) the stereo chemistry is switched; the activity now resides in the (*R*)-enantiomer[137,139] (see Table 1.5, (*R*)-D0749). In order to account for the stereochemistry in the pharmacophoric model (Figure 1.7) a third aromatic element Ar_3 was defined for large pyrrolidine substituents and a small hydrophobic pocket HP for small pyrrolidine substituents for benzamides with (*R*) and (*S*) configuration, respectively.

Later, in Chapters 4, 5 and 6, two different series of benzamides will be discussed as they form the bases for a couple of publications[140,141] involving 3D QSAR and multivariate statistical analysis. Furthermore, structure-activity relationships of benzamides will be dealt with also in these chapters.

1.4 Quantitative Structure-Activity Relationships (QSAR)

So far, two different approaches to design new drugs have been discussed: homology modeling and the active analogue approach.[103] Despite sophisticated computational tools,[142,143] with all known dopamine receptors cloned and site mutagenesis techniques available, the active analogue approach is still the most widely used approach to design new drugs. A speculative reason for this may be the fact that in receptor modeling too many parameters need to be estimated: conformational flexibility in the receptor protein, solvent dependency (dielectric constant), pH at the active site (protonated ligands or not), water improved receptor binding[144], induced fit[145,146] and many more. Hence, most drugs are developed with ideas based on structure-activity relationships and simple molecular modeling studies. In 1964 Hansch and Fujita[147] introduced a way to predict biological activity from theoretically derived molecular descriptors, often referred to as Hansch-analysis. The ideas behind the Hansch analysis are as valid today as in 1964 since theoretically (not necessarily) generated descriptors are assumed to be independent of the ligands conformations. More important, if the predictive ability of the model is high enough, valuable time will not be spent synthesizing inactive compounds.

Physicochemical Molecular Descriptors

The applicability of physicochemical descriptors is manifold. First, prediction of the biological activity (*e.g.*, receptor affinity) as introduced by Hansch *et al.*[147] (see below). Second, they enable us to search large molecular databases (*e.g.*, Chemical Abstracts structural database) by simply define a physicochemical profile for the target molecule.[148-150] Third, recently developed techniques like combinatorial chemistry[100-102] and high-throughput screening,[98,99] call for methods in order to design the synthesis of the most diverse compounds and for the definition of new specific targets, respectively.

Examples of commonly used physicochemical descriptors are listed in Table 1.6. which may be divided into steric, electrostatic and hydrophobic (lipophilic) types of descriptors, although many of them are a combination of all three types. Some of these descriptors are listed in the literature, while some may be obtained through analytical experiments[151,152] or through computational calculations.[142] Steric descriptors like FW, L, B and VdWV (Table 1.6) are simply different attempts to quantify the molecular structure; electrostatic descriptors like σ_m, HOMO or LUMO portray the electronic features of a compound (*e.g.*, the influence of a specific substituent on an aromatic system) and the hydrophobic descriptors π, logP and logD accounts for a molecules ability to, for instance penetrate the blood-brain barrier.[153,154] Taken together, the array of descriptors collected for a specific molecule compares to the fingerprint from a human being, hence, it should be unique.

Hansch Analysis

In the early 1960s, Hansch and co-workers[147] investigated the possibility of expressing a relationship between structural and physicochemical properties and biological activity, quantitatively. Typically, properties as logP, σ, or E_s representing 1-octanol/water partition coefficient, the well-known Hammett constant and Tafts steric descriptors (Table 1.6), respectively, were used as descriptors. In the Hansch analysis, the biological activity log(1/*C*) were correlated with the descriptors using Multiple Linear Regression (MLR, see Chapter 2), also called Ordinary Least Squares (OLS).[155,156] The so-called Hansch equation (Equation 1.1) comprise the relationship established by Hansch *et al.*, where *a*, *b*, *c*, *d* and *e* are constants obtained through regression analysis.

$$\log\left(\frac{1}{C}\right) = -a\left(\log \mathrm{P}\right)^2 + b\log \mathrm{P} + c\,s + d\mathrm{E}_s + e \qquad (1.1)$$

As was emphasized by Van de Waterbeemd,[157] Hansch analysis is a method aiming at describing the relationship between only a few variables and the biological activity and should not be considered too much as a predictive model. By employing MLR for the regression analysis a couple of crucial items need to be considered: 1) keep the ratio of compounds to variables greater than approximately five and 2) multicollinearity may cause spurious solutions. Today, methods to circumvent these problems are available, where the original variables are replaced by underlying orthogonal latent variables (*i.e.*, PCR and PLS in Chapter 2). The predictive ability of a model may then be validated with crossvalidation[155] or by predictions of an external test set.

: Molecular Diversity and Experimental Design

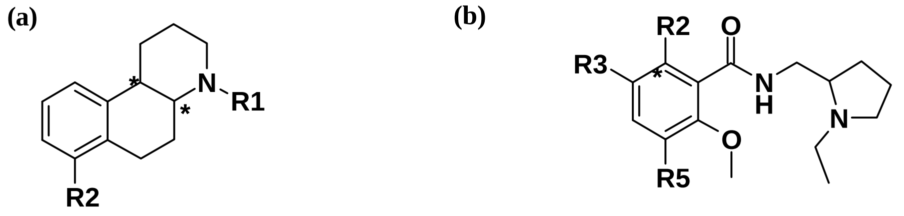


Figure 1.8 The basic skeleton of *trans*-OHB[*f*]Qs (a) and 6-methoxybenzamides (b) used by Nilsson *et al.* and Norinder *et al.*, respectively.

In Chapter 3, physicochemical descriptors are employed to guide the selection of which compounds to synthesize.[158] From a library of a 88 tentative OHB[*f*]Qs (Figure 1.8(a)), only the 15 most diverse compounds were selected to be synthesized by means of a factorial design in the Principal Properties (PPs). The PPs are score-vectors obtained from a Principal Component Analysis (PCA, see Chapter 2) of the physicochemical descriptors, and comprise the variation that significantly discriminate between the compounds. Norinder *et al.*[159] used physicochemical descriptors in order to increase the understanding of the structure-activity relationships of a series of benzamides (Figure 1.8(b)). They used a fractional factorial design in the first three PPs, for the selection of 16 representative compounds, out of 70, for the training set. In the following regression analysis the original descriptors (not the PPs, thus) from the training set were correlated with the biological activity (pIC_{50} for the DA D_2 receptor), using PLS. The remaining compounds were utilized as a test set in order to estimate the predictability of the PLS model. The profound difference in the approaches used in Chapter 3 and by Norinder *et al.*, should be obvious. Nilsson *et al.* generated descriptors for whole molecules, *e.g.*, logP was used rather than the contribution from single substituents (π). The objective with that investigation was to select the most diverse compounds to synthesize, not necessarily to create a predictive model. In contrast, Norinder *et al.* already had a large data set with compounds tested for biological activity and their aim was to elucidate also the influence of single substituent positions on the biological activity. Therefore, the training set was selected such that the diversity in each position R2, R3 and R5, in Figure 1.8(b), was maximized. The conclusions drawn from this investigation will be discussed further in Chapter 6, where this data set was analyzed with 3D QSAR (see next section) using multilinear PLS[160] as regression method.

: Molecular Fields

The obvious extension of the Hansch analysis is modeling where molecular flexibility is considered. The concept of Comparative Molecular Field Analysis (CoMFA)[161] as presented by Cramer *et al.*[117] in 1988 is such a method, commercially available as a module in the molecular modeling package SYBYL.[142]

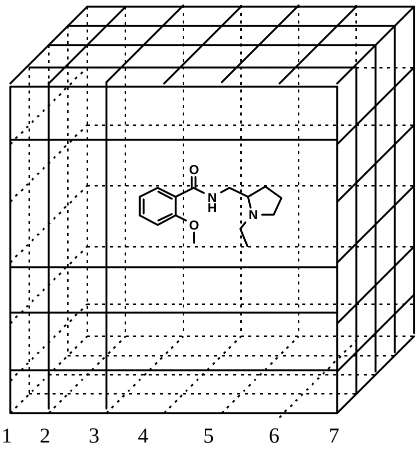


Figure 1.9 The three dimensional grid used in CoMFA to generate molecular field descriptors. For clarity, grid points within the grid are omitted

Table 1.6 The most commonly used molecular descriptors available in the literature or through computational calculations.

Descriptor	**Abbr.**	**type**[a]	**ref.**	**Descriptor**	**short**	**type**[a]	**ref.**
Formula Weight	FW	ste		Ionisation constant	pK_b		162
Hammett constant	σ_m	ele	163	Swain-Lupton field	F		163
Hammett constant	σ_p	ele	163	Swain-Lupton res.	R		163
Tafts polar constant	σ^*	ele	157	VdWaals Volume	VdWV	ste	142
Tafts steric parameter	E_s	ste	157	VdWaals Area	VdWA	ste	142
Hansch aromatic fragment	π	lip	162,163	Connolly Surface Vol.	CoVo	ste	142
Lipophilicity	logP	lip	157,162	Connolly Surface Area	CoAr	ste	142
Lipophilicity (pH=7.4)	logD	lip	157,162	Electronic Energy	ELEC	ele	142
Connectivity index (Randic)	$^1\chi$	ste	157	Core-Core interaction	CoCo	ste	142
Connectivity index	$^2\chi$	ste	157	Heat of Formation	HoFo	ele	142
Molar Refractivity	MR	ele	164	Ionization potential	HOMO	ele	142
Verloop Sterimol	L	ste	157	Electron affinity	LUMO	ele	142
Verloop Sterimol	B	ste	157	Dipole moment	Dipo	ele	142
Eudismic Index	EI		165	Point charges	chaX	ele	142
Ionization constant	pK_a		162				

[a] ele electronic; ste steric; lip lipophilicity

Molecular fields basically are three dimensional representations of the steric, electrostatic and hydrophobic surroundings of a molecule. A molecular field is generated by enclosing the molecule in a three dimensional grid (Figure 1.9) and assigning nonbonded interactions between a probe atom and the molecule in each grid point. Obviously, the difference between two different types of fields is the algorithm with which the nonbonded interactions are calculated. In SYBYL/CoMFA the steric field interaction energies (E_{ste}) are Lennard-Jones potentials, also referred to as steric 6-12 potentials,[166] which are sensitive to changes in the distance between the probe and the atoms (r_i) as can be seen in Equation 1.2. N is the number of atoms in the molecule; A and B are constants characteristic for the probe atom type and the type of the ith atom in the molecule, respectively.

$$\mathrm{E}_{\mathrm{ste}} = \sum_{i=1}^{N}\left[\frac{A}{\mathrm{r}_i^{12}} - \frac{B_i}{\mathrm{r}_i^{6}}\right] \tag{1.2}$$

The electrostatic field interaction energies are less influenced by the distances between the probe and the atoms but instead the charge of the probe and the point charges of the atoms are important. In addition, E_{ele} is very sensitive to spatial dielectric behavior of the environment[167] and a distance-dependent dielectric term has been proposed.[168] The magnitude of the electrostatic potential (E_{ele}) between two ions with charges Q and q separated by a distance r is given by Coulomb's law[166]:

$$\mathrm{E}_{\mathrm{ele}} = \sum_{i=1}^{N}\frac{\mathrm{Qq}_i}{\mathrm{K}z}\left[\frac{1}{\mathrm{r}} + \frac{(z - e)/(z + e)}{\sqrt{\mathrm{r}^2 + 4\mathrm{s}_{\mathrm{Q}}\mathrm{s}_{\mathrm{q}}}}\right] \tag{1.3}$$

where N is the number of atoms in the molecule; Q is the charge of the probe atom; q_i is the point charge on the ith atom and K is a constant term. In this formula a homogenous protein phase and a homogenous solution phase with dielectrics ζ and ε, respectively, are assumed to be present.[167] The depth of each protein atom (s_p) in the protein phase is assessed by counting the number of neighboring protein atoms whose nuclei lie within 4 Å. For the probe atom the depth (s_Q) is calculated similarly. Consequently, Equation 1.3 leads to an effective dielectric of ζ when the pairwise groups of atoms are so deep in the protein and so close together that the solvent effects can be neglected. However, when one or both of the atoms approach the surface of the protein the effective dielectric becomes $(\zeta + \varepsilon)/2$ since the term $4s_Q s_q$ is set to zero.

Prior to calculation of the electrostatic field point charges of the atoms need to be calculated. In the original CoMFA article[117] charges were calculated by the method of Gasteiger and Marsili[169] as was implemented in SYBYL. Today, however, several options are available although most often point charges are estimated by semi-empirical AM1 single point calculations.[170]

There are also variations of the above described steric field. Kroemer *et al.*[171] replaced the steric 6-12 potential interactions with atom-based indicator variables. That is, they assigned the values 30 or 0 kcal/mol to a grid point if an atom was present in the adjacent small cube or not. Similar, Floersheim *et at.*[172] assigned values of either 1 or 0 to a grid point, depending on whether the grid point was within, or outside, the van der Waals radius of any atom in the target molecule.

In the GRID program[136,167] a different and intuitively more appealing (authors comment) approach is used. Each grid point is assigned with the sum of three different non-bonded interactions, *i.e.*, E_{ste}, E_{ele} and E_{hb}, as in Equation 1.4. The two former terms are calculated as in Equations 1.2 and 1.3, while the hydrogen bonding contribution[167] is calculated as in Equation 1.5. C and D are tabulated values for specific atoms; d is the distance between the atoms; m is usually four but the whole E_{hb} term is set to zero when $\theta \leq 90°$. If the probe group donates the hydrogen bond it is assumed that the probe can orient itself in order to form the most effective hydrogen bond, and the cos θ is set to unity.

$$E_{tot} = \sum E_{ste} + \sum E_{ele} + \sum E_{hb} \tag{1.4}$$

$$E_{hb} = \left[C/d^6 - D/d^4 \right] \cos^m \theta \tag{1.5}$$

Consequently, and in contrast to other methods (*e.g.*, SYBYL/CoMFA), fields generated in the GRID program portrays the specifics of certain probes. Thus, it is the users task to select a probe that best reflects what needs to be investigated. For instance, a water molecule, a carbon atom or a Ca^{+2} ion could be chosen to display hydrogen bonding, steric and electrostatic characteristics of the ligands, respectively. Goodford[173] has demonstrated how GRID probes explicitly can reflect individual properties of specific chemical groups attached to the target molecule.

Fields generated in SYBYL or GRID are the most commonly used since they are commercially available. However, other molecular descriptors are available, *e.g.*, Molecular Shapes,[174-176] Molecular Lipophilicity Potentials,[177,178] Molecular Similarity Indices[179] (*i.e.*, CoMSIA) and Molecular Similarity Matrices.[180-182]

Independent of which program used to generate molecular fields, the following parameters need to be specified by the user: 1) the grid size; 2) the grid resolution, *i.e.*, grid points per Å; 3) the type of probes and 4) the probe charge.

Comparative Molecular Field Analysis (CoMFA)

As stated above, the logical extension of the Hansch analysis[147] is Comparative Molecular Field Analysis (CoMFA),[117] where the physicochemical parameters are replaced or combined with field descriptors. In analogy with the Hansch analysis, in CoMFA the molecular fields are correlated with the biological activity. CoMFA takes the three dimensional conformations of the molecules into consideration, hence, several crucial items need to be considered.

First, in order to find low energy conformations, or rather the global minimum energy conformation of each compound under investigation, conformational analyses[183-186] are conducted. According to the Boltzman distribution,[166] a low energy conformation is more abundant than a high energy conformation and, consequently, also more likely to be involved in the ligand-receptor interaction.

Second, all molecules must be aligned in the same coordinate system and several options to perform this are possible. If a pharmacophore is available, one might choose to superimpose all molecules on mutual and likely interaction points with the receptor (see Chapter 4). The compounds could be docked into the active site of a receptor homology model (see Chapter 6) or, alternatively, the molecular fields could be superimposed in a least square manner.[142] Independent of which alignment approach that is employed, in CoMFA the differences between the aligned molecular fields are correlated with the biological activity. Therefore, an alignment procedure where the global overlap between structurally related compounds[140,187] (Chapter 4) are maximized, is likely to perform just as good as a more elaborated and rational alignment[140] (Chapter 6). This is the case when the ligands are flexible and the rationale is pure statistical: If the alignment is not performed with maximized overlap between the molecules an increased level of insignificant variation, *i.e.*, noise is inevitable. Noise may, or may not, affect the predictability detrimentally.

Third, molecular fields are generated first when the molecules are properly aligned.

In Figure 1.10, a typical CoMFA data set consisting of I molecules characterized with a single molecular field is shown. In order to perform the PLS analysis each grid, with the dimensions J, K and L, is unfolded to form a row with JKL number of columns. Traditionally, only one response variable is considered in CoMFA, *e.g.*, the affinity for a central dopamine receptor and, therefore, a bilinear PLS1 algorithm[188-190] is used for the regression analysis. The theory covering the basics of PLS analysis is discussed in Chapter 2, but specific details important for CoMFA are pointed out in the following.

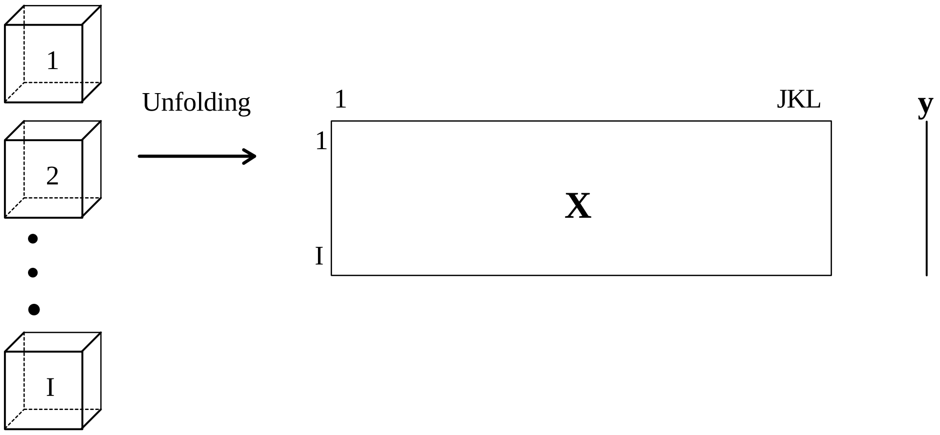


Figure 1.10 The unfolding of the molecular fields from the *N* molecules into a matrix **X** with *I* rows and *JKL* columns. In CoMFA, the biological activity is usually represented in one column (**y**). The number columns in **X** equals the number of grid points in each grid.

In PLS, the original variables are replaced by latent variables, *i.e.*, linear combinations of the original variables and, therefore, the number of objects should be larger than the number of latent variables included in the model (see Chapter 2). However, each additional component adds also insignificant variation to the model and should be added only if it improves the predictability. The predictability is usually estimated with crossvalidation as described in Chapter 2.

The results from a CoMFA model are often interpreted by studying contour plots of the PLS-coefficients $\mathbf{b}_{\mathrm{PLS}}$ (Equation 1.6, see also Chapter 2).

$$\hat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \ldots + b_P\mathbf{x}_P = \mathbf{X}\mathbf{b}_{\mathrm{PLS}} \qquad (1.6)$$

In SYBYL/CoMFA a steric and an electrostatic contour plot is generated, and each b_{PLS} coefficient is multiplied with the standard deviation for the corresponding variable. In effect, it is an enhancement of the b_{PLS} coefficients in grid points where the variation is large. This is performed in order to simplify the interpretations. The data set in Chapter 4 is analyzed using GRID/GOLPE which produces slightly different contour plots, since each field corresponds to a specific probe. For example, in the contour plot from a water probe, regions are revealed where hydrogen bonding is favorable for high affinity, considering steric, electrostatic and hydrogen bonding interactions, simultaneously. Molecular fields generated from the water (OH2), carbon (C3) and the calcium (CA+2) probes in the GRID-program, will be used for the description of the molecules included in the two data sets scrutinized in this thesis.

1.5 References

1. Andreasen, N.C. and Olsen, S. Negative v Positive Schizophrenia. Definition and Validation. *Arch. Gen. Psychiatry* **1982**, *39*, 789-794.
2. Carpenter, W.T.; Heinrichs, D.W.; Hanlon, T.E. A Comparative Trial of Pharmacological Strategies in Schizophrenia. *Am. J. Psychiatry* **1987**, *144*, 1466-1470.
3. Carpenter, W.T. The Deficit Syndrome. *Am. J. Psychiatry* **1994**, *151*, 327-329.
4. Siris, S.G. Diagnosis of Secondary Depression in Schizophrenia: Implications for DSM-IV. *Schizophr. Bull.* **1991**, *17*, 75-98.

5. Roy, A. Suicide in Chronic Schizophrenia. *Brit. J. Psychiat.* **1982**, *141*, 171-177.
6. Borison, R.L.; Diamond, B.I.; Pathiraja, A.; Meibach, R.C. Novel Antipsychotic Drugs. Meltzer, H.Y. Ed.; Raven: New York, **1992**; pp. 223
7. Chouinard, G.; Jones, B.; Remington, G.; Bloom, D.; Addington, D.; MacEwan, G.W.; Labelle, A.; Beauclair, L.; Arnott, W. A Canadian Multicenter Placebo-Controlled Study of Fixed Doses of Risperidone and Haloperidol in the Treatment of Chronic Schizophrenic Patients. *J. Clin. Psychopharmacol* **1993**, *13*, 25-40.
8. Kay, S.R.; Fisbein, A.; Opler, L.A. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. *Schizophr. Bull.* **1987**, *13,* 261-276.
9. Claghorn, J.; Honingfeld, G.; Abuzzahab, F.S.; Wang, R.; Steinbook, R.; Tuason, V.; Klerman, G. The Risks and Benefits of Clozapine versus Chlorpromazine. *J. Clin. Psychopharmacol* **1987**, *7,* 377-384.
10. Kane, J.; Honigfield, G.; Singer, J.; Meltzer, H.Y. Clozapine for the Treatment-Resistant Schizophrenic. *Arch. Gen. Psychiatry* **1988**, *45*, 789-796.
11. Meltzer, H.Y.; Bastani, B.; Kwon, K.Y.; Ramirez, L.F.; Burnett, S.; Sharpe, J. A Prospective Study of Clozapine in Treatment-Resistant Schizophrenic Patients. I. Preliminary report. *Psychopharmacology* **1989**, *99,* S68-S72.
12. Juul Povisen, U.; Norig, U.; Fog, R.; Gerlach, J. Tolerability and Therapeutic Effect of Clozapine. *Acta Psychiatr. Scand.* **1985**, *71,* 176-185.
13. Casey, D.E. Clozapine: Neuroleptic-Induced EPS and Tardive Dyskinesia. *Psychopharmacology* **1989**, *99,* S47-S53.
14. Peacock, L.; Solgaard, T.; Lublin, H.; Gerlach, J. Clozapine versus Typical Antipsychotics. Aretro- and Prospective Study of Extrapyramidal Side Effects. *Psychopharmacology* **1996**, *124*, 188-196.
15. Lieberman, J.A.; Johns, C.A.; Kane, J.M.; Rai, K.R.; Pisciotta, A.V.; Saltz, B.L.; Howard, A. Clozapine-Induced Agranulocitosis: Non-Cross-Reactivity With Other Psychotropic Drugs. *J. Clin. Psychiatry* **1988**, *49,* 271-277.
16. Stockton, M.E. and Rasmussen, K. Olanzapine, a Novel Atypical Antipsychotic, Reverses d-Amphetamine-induced Inhibition of Midbrain Dopamine cells. *Psychopharmacology* **1996**, *124,* 50-56.
17. Moore, N.A.; Calligaro, D.O.; Wong, D.T.; Bymaster, F.; Tye, M.C. The Pharmacology of Olanzapine and Other New Antipsychotic Agents. *Curr. Opin. Invest. Drugs* **1993**, *2*, 281-293.
18. Beasley, C.M.; Tollefson, G.; Tran, P.; Satterlee, W.; Sanger, T.; Hamilton, S. The Olanzapine HGAD Study Qroup (1995) Olanzapine versus Placebo and Haloperidol: Acute Phase Result of the North American Double blind Olanzapine trial. *Neuropsychopharmacology* **1996**, *14,* 105-118.
19. Bunney, B.S. Animal Models in Psychology and Neurology. Hanin and Usdin, Eds.; Pergamon Press: New York, **1977**; pp. 91-104.
20. Roth, B.L.; Craigo, S.C.; Choudhary, M.S.; Uluer, A.; Monsma, F.J.; Shein, Y.; Meltzer, H.Y. Binding of Typical and Atypical Antipsychotic Agents to 5-Hydroxytryptamine-6 and 5-Hydroxytryptamine-7 Receptors. *J. Pharmacol. Exp. Ther.* **1994**, *268,* 1403-1410.
21. Koehler, K.F.; Radesater, A.C.; Karlsson-Boethius, G.; Bryske, B.; Widman, M. Regional Distribution and *in vivo* Binding of the Typical Antipsychotic Drug remoxipride. *J. Neural Transm.* **1992**, *87*, 49-62.
22. Walker, J.M.; Bowen, W.D.; Walker, F.O.; Matsumoto, R.R.; De Costa, B.; Rice, K.C. Sigma Receptors: Biology and Function. *Pharmacol. Rev.* **1990**, *42,* 355-390.
23. Assie, M.-B.; Sleight, A.J.; Koek, W. Biphasic Displacement of [^{3}H]YM-09151-2 Binding in the Rat Brain by Thioridazine, Risperidone and Clozapine, but not by Other Antpsychotics. *Eur. J. Pharmacol.* **1993**, *237,* 183-189.
24. Canton, H.; Verriele, L.; Colpaert, F.C. Binding of Typical and Atypical Antipsychotics to $5HT_{1C}$ and $5HT_2$ Sites: Clozapine Potently Interacts with $5HT_{1C}$ Sites. *Eur. J. Pharmacol.* **1990**, *191,* 93-96.
25. Watling, K.J.; Beer, M.S.; Stanton, J.A.; Newberry, N.R. Interaction of the Atypical Neuroleptic Clozapine with 5-HT_3 Receptors in the Cerbral Cortex and Superior Cervical Ganglion of the Rat. *Eur. J. Pharmacol.* **1990**, *182,* 465-471.
26. Meltzer, H.Y.; Matsubar, S.; Lee, J.-C. Classification of Typical and Atypical Antipsychotic Drugs on the Basic of Dopamine D-1, D-2 and Serotonin2 pK_i Values. *J. Pharmacol. Exp. Ther.* **1989**, *251,* 238-246.
27. Bolden, C.; Cusack, B.; Richelson, E. Antagonism by Antimuscarinic and Neuroleptic Compounds at the Five Cloned Human Muscarinic Cholinergic Receptors Expressed in Chinese Hamster Ovary Cells. *J. Pharmacol. Exp. Ther.* **1992**, *260,* 576-580.
28. Sonesson, C. PhD Thesis. Arylpiperidine and Arylpyrrolidine Derivatives with Potential Antipsychotic Efficacy. Synthesis and Quantitative Structure-Activity Relationships Faculty of Pharmacy, Uppsala University, Sweden. **1995**
29. Carlsson, A. and Lindquist, M. Effect of Chlorpromazine or Haloperidol on Formation of 3-Methoxythyramine and Normethanephrine in Mouse Brain. *Acta Pharmacol. Toxicol.* **1963**, *20,* 140-144.
30. Randrup, A. and Munkvad, I. Special Antagonism of Amphetamine-Induced Abnormal Behavior: Inhibition of Stereotyped Activity with Increase of some Normal Activities. *Psycopharmacologia* **1965**, *7,* 416-422.

31. Snyder, L.A. *Science* **1974**, *184,* 1243-1253.
32. Seeman, P. *Fed. Proc.* **1974**, *33,* 246
33. Carlsson, A. Antipsychotic Drugs, Neurotransmitters, and Schizophrenia. *Am. J. Psychiatry* **1978**, *135,* 164-173.
34. Carlsson, A. Does Dopamine Have a Role in Schizophrenia? *Biol Psychiatry* **1978**, *13,* 3-21.
35. Van Rossum, J.M. The Significance of Dopamine-Receptor Blockade for the mechanism of Action of Neuroleptic Drugs. *Arch. Int. Pharmacodyn. Ther.* **1966**, *160,* 492-494.
36. Matthysse, S. Antipsychotic Drug Actions: a Clue to the Neuropathology of Schizophrenia? *Fed. Proc.* **1973**, *32*, 200-205.
37. Seeman, P. Antipsychotic Drugs: Direct Correlation Between Clinical Potency and Presynaptic Action on Dopamine Neurons. *Science* **1975**, *188,* 1217-1219.
38. Seeman, P. Dopamine Receptors and the Dopamine Hypothesis of Schizophrenia. *Synapse* **1987**, *1,* 133-152.
39. Peroutka, S.J. and Snyder, S.H. Relationship of Neuroleptic Drug Effects at Brain Dopamine, Serotonin, alfa-Adrenergic, and Histamine Receptors to Clinical Potency. *Am. J. Psychiatry* **1980**, *137,* 1518-1522.
40. Chiodo, L.A. and Bunney, B.S. Typical and Atypical neuroleptics: Differential Effects of Chronic Administration on the Activity of A9 and A10 Midbrain Dopaminergic Neurons. *J. Neurosci.* **1983**, *3,* 1607-1619.
41. Wooley *Proc. Natl. Acad. Sci. USA* **1954**, *40,* 228-231.
42. Roth, B.L.; Ciaranello, R.D.; Meltzer, H.Y. Binding of Typical and Atypical Antipsychotic Agents to Transiently Expressed 5-HT$_{1C}$ Receptors. *J. Pharmacol. Exp. Ther.* **1992**, *260,* 1361-1365.
43. Hicks, P.B. The Efffect of Serotonergic Agents on Haloperidol Induced Catalepsy. *Life Sci.* **1990**, *47,* 1609-1615.
44. Neal-Beliveau, B.S.; Joyce, J.N.; Lucki, I. Serotonergic Involvement in Haloperidol-Induced Catalepsy. *J. Pharmacol. Exp. Ther.* **1993**, *265,* 207-217.
45. Hornykiewicz, O. Dopamine in the Basala Ganglia. Its Role and Therapeutic Implications (Including the Clinical Use of L-DOPA). *Br. Med. Bul.* **1973**, *29,* 172-178.
46. Tandon, R. and Greden, J.F. Cholinergic Hyperactivity and Negative Schizophrenic Symptoms. *Arch. Gen. Psychiatry* **1989**, *46,* 745-753.
47. Tandon, R.; Greden, J.F.; Silk, K.R. Treatment of Negative Schizophrenic Symptoms with Trihexyphenidyl. *J. Clin. Psychopharmacol* **1988**, *8,* 212-215.
48. Tandon, R. and Greden, J.F. Trihexyphenidyl Treatment of Negative Schizophrenic Symptoms. *Acta Psychiatr. Scand.* **1987**, *76,* 732
49. Leysen, J.E.; Janssen, P.M.F.; Schotte, A.; Luyten, W.H.M.L.; Megens, A.A.H.P. Interaction of Antipsychotic Drugs with Neurotransmitter Receptor Sites *in vitro* and *in vivo* in Relation to Pharmacological and Clinical Effects: Role of 5-HT$_2$ Receptors. *Psychopharmacology* **1993**, *112,* S40-S54.
50. Rivest, R. and Marsden, C.A. Muscarinic Antagonists Attenuate the Increase in Accumbens and Striatum Dopamine Metabolism Produced by Clozapine but not by Haloperidol. *Br. J. Pharmacol.* **1991**, *104,* 234-238.
51. Stein, L. and Wise, C.D. Possible Etiology of Schizophrenia: Progressive Damage to the Noradrenergic Reward System by 5-Hydroxydopamine. *Science* **1971**, *171,* 1032-1036.
52. Van Kammen, D.P.; Peters, J.; Yao, J.; Van Kammen, W.B.; Neylen, T.; Shaw, D.; Linnoila, M. Norepinephrine in Acute Exacerbations of Chronic Schizophrenia. *Arch. Gen. Psychiatry* **1990**, *47,* 161-168.
53. Breier, A.; Buchanan, R.W.; Waltrip, R.W.; Liswak, S.; Holmes, C.; Goldstein, D.S. The Effect of Clozapine on Plasma Norepinephrine: Relationship to Clinical Efficacy. *Neuropsychopharmacology* **1994**, *10,* 1-7.
54. Csernansky, J.G. Antipsychotics. Ed.; Springer: Berlin, **1996**
55. Seeman P. Dopamine Receptors and Psychosis. Scientific American. **1995**, 28-37.
56. Civelli, O.; Bunzow, J.; Albert, P.; Van Tol, H.; Grandy, D. Molecular Biology of the Dopamine D$_2$ Receptor. *NIDA. Res. Monogr.* **1991**, *111*, 45-53.
57. Dearry, A.; Gingrich, J.A.; Falardeau, P.; Fremeau, R.T.; Bates, M.D.; Caron, M.G. Molecular Cloning and Expression of the Gene for a Human D$_1$ Dopamine Receptor. *Nature* **1990**, *347,* 72-76.
58. Monsma, F.J. Molecular Cloning and Expression of a D$_1$ Dopamine Receptor Linked to Adenylyl Cyclase Activation. *Proc. Natl. Acad. Sci. USA* **1990**, *87,* 6723-6727.
59. Sunahara, R.K.; Niznik, H.B.; Weiner, D.M.; Stormann, T.M.; Brann, M.R.; Kennedy, J.L.; Gelernter, J.E.; Rozmahel, R.; Yang, Y.; Israel, Y.; Seeman, P.; O'Dowd, B.F. Human Dopamine D$_1$ Receptor Encoded by an Intronless Gene on Chromosome 5. *Nature* **1990**, *347,* 80-83.
60. Zhou, Q.-Y.; Grandy, D.K.; Thambi, L.; Kushner, J.A.; Van Tol, H.; Cone, R.; Pribnow, D.; Salon, J.; Bunzow, J.R.; Civelli, O. Cloning and Expression of Human and Rat D$_1$ Dopamine Receptors. *Nature* **1990**, *347,* 76-80.
61. Schwartz, J.-C.; Giros, B.; Martres, M.-P.; Sokoloff, P. The Dopaminergic Receptor Family: Molecular Biology and Pharmacology. *The Neurosciences* **1992**, *4,* 99-108.

62. Bouthenet, M.L.; Souil, E.; Martres, M.P.; Sokoloff, P.; Giros, B.; Schwartz, J.C. Localization of Dopamine D_3 Receptor mRNA in the Rat Brain Using *in situ* Hybridization Histochemistry: Comparison with Dopamine D_2 Receptor mRNA. *Brain Res.* **1991**, *564,* 203-219.
63. Schertler, G.F.X.; Villa, C.; Henderson, R. Projection Structure of Rhodopsin. *Nature* **1993**, *362,* 770-772.
64. Henderson, R.; Baldwin, J.M.; Ceska, T.A.; Zemlin, F.; Beeleman, E.; Downing, K.H. Model for the Structure of Bacteriorhodopsin Based on High Resolution Electron Cryo-Microscopy. *J. Mol. Biol.* **1990**, *213,* 899-929.
65. Grigorieff, N.; Ceska, T.A.; Downing, K.H.; Baldwin, J.M.; Henderson, R. Electron-Crystallographic Refinement of the Structure of Bacteriorhodopsin. *J. Mol. Biol.* **1996**, *259,* 393-421.
66. Unger, V.M.; Hargrave, P.A.; Baldwin, J.M.; Schertler, G.F.X. Arrangement of Rhodopsin Transmembrane α-Helices. *Nature* **1997**, *389,* 203-206.
67. Kimura, Y.; Vassylyev, D.G.; Miyazawa, A.; Kidera, A.; Matsushima, M.; Mitsuoka, K.; Murata, K.; Hirai, T.; Fujiyoshi, Y. Surface of Bacteriorhodopsin Revealed by High-Resolution Electron Crystallography. *Nature* **1997**, *389,* 206-211.
68. Teeter, M.M.; Froimowitz, M.; Stec, B.; DuRand, C.J. Homology Modeling of the Dopamine D_2 Receptor and its Testing by Docking of Agonists and Tricyclic Antagonists. *J. Med. Chem.* **1994**, *37*, 2874-2888.
69. Rippman, F. and Böttcher, H. Molecular Modelling of G Protein-Coupled Receptors- A New Approach to Studying Structure Activity Relationships. *Kontakte* **1994**, *1,* 30-36.
70. Giros, B.; Sokoloff, P.; Martres, M.P.; Riou, J.F.; Emorine, L.J.; Schwartz, J.C. Alternative Splicing Directs the Expression of two D_2 Dopamine Receptor Isoforms. *Nature* **1989**, *342,* 923-926.
71. Monsma, F.J. Multiple D_2 Dopamine Receptors Produced by Alternative RNA Splicing. *Nature* **1989**, *342,* 926-929.
72. Dal Toso, R.; Sommer, B.; Ewert, M.; Herb, A.; Pritchett, D.B.; Bach, A.; Shivers, B.D.; Seeburg, P.H. The Dopamine D_2 Receptor: Two Molecular Forms Generated by Alternative Splicing. *EMBO J.* **1989**, *8,* 4025-4034.
73. Schwartz, J.-C.; Levesque, D.; Martres, M.-P.; Sokoloff, P. Dopamine D_3 Receptor: Basic and Clinical Aspects. *Clinical Neuropharmacology* **1993**, *16,* 295-314.
74. Ohara, K.; Ulpian, C.; Seeman, P.; Sunahara, R.K.; Van Tol, H.; Niznik, H.B. Schizophrenia: Dopamine D_1 Receptor Sequence is Normal, but has DNA Polymorphisms. *Neuropsychopharmacology* **1993**, *8,* 131-135.
75. Sunahara, R.K.; Guan, H.-C.; O'Dowd, B.F.; Seeman, P.; Laurier, L.G.; NG, G.; George, S.R.; Torchia, J.; Van Tol, H.; Niznik, H.B. Cloning of the Gene for a Human Dopamine D_5 Receptor with Higher Affinity for Dopamine than D_1. *Nature* **1991**, *350*, 614-619.
76. Grandy, D.K.; Zhang, Y.; Bouvier, C.; Zhou, Q.-Y.; Johnson, R.A.; Allen, L.; Buck, K.; Bunzow, J.R.; Salon, J.; Civelli, O. Multiple Human D_5 Dopamine Receptor Genes: a Functional Receptor and Two Pseudogenes. *Proc. Natl. Acad. Sci. USA* **1991**, *88,* 9175-9179.
77. Sobell, J.L.; Lind, T.J.; Sigurdson, D.C.; Zald, D.H.; Snitz, B.E.; Grove, W.M.; Heston, L.L.; Sommer, S.S. The D_5 Dopamine Receptor Gene in Schizophrenia: Identification of a Nonsense Change and Multiple Missense Changes but Lack of Association with Disease. *Hum. Mol. Genet.* **1995**, *4,* 507-514.
78. Selbie, L.A.; Hayes, G.; Shine, J. The Major Dopamine D_2 Receptor: Molecular Analysis of the Human D_{2A} Subtype. *DNA* **1989**, *8,* 683-689.
79. Robakis, N.K.; Mohamadi, M.; Fu, D.J.; Sambamurit, K.; Refolo, L.M. Human Retina D_2 Receptor cDNAs have Multiple Polyadenylation Sites and Differ from a Pituitary Clone at the 5' Non-Coding Region. *Nucleic Acids Res.* **1990**, *18,* 1299.
80. Grandy, D.K.; Marchionni, M.A.; Makam, H.; Stofko, R.E.; Alfano, M.; Frothingham, L.; Fisher, J.B.; Burke-Howie, K.J.; Bunzow, J.R.; Server, A.C.; Civelli, O. Cloning of the cDNA and Gene for a Human D_2 Dopamine Receptor. *Proc. Natl. Acad. Sci. USA* **1989**, *86,* 9762-9766.
81. Stormann, T.M.; Gdula, D.C.; Weiner, D.M.; Brann, M.R. Molecular Cloning and Expression of a Dopamine D_2 Receptor from Human Retina. *Mol. Pharmacol.* **1990**, *37,* 1-6.
82. Selbie, L.A.; Hayes, G.; Shine, J. DNA Homology Screening: Isolation and Characterization of the Human D_{2A} Dopamine Receptor Subtype. *Adv. Second Messenger Phosphoprotein Res.* **1990**, *24*, 9-14.
83. Araki, K.; Kuwano, R.; Morii, K.; Hayashi, S.; Minoshima, S.; Shimizu, N.; Katagiri, T.; Usui, H.; Kumanishi, T.; Takahashi, Y. Structure and Expression of Human and rat D_2 Dopamine Receptor Genes [published erratum appears in Neurochem Int 1992 Oct;21(3):465]. *Neurochem. Int.* **1992**, *21,* 91-98.
84. Dearry, A.; Falardeau, P.; Shores, C.; Caron, M.G. D_2 Dopamine Receptors in the Human Retina: Cloning of cDNA and Localization of mRNA. *Cell. Mol. Neurobiol.* **1991**, *11,* 437-453.
85. Seeman, P.; Ohara, K.; Ulpian, C.; Seeman, M.V.; Jellinger, K.; Tol, H.H.; Niznik, H.B. Schizophrenia: Normal Sequence in the Dopamine D_2 Receptor Region that Couples to G-proteins. DNA Polymorphisms in D_2. *Neuropsychopharmacology* **1993**, *8,* 137-142.
86. Itokawa, M.; Arinami, T.; Futamura, N.; Hamaguchi, H.; Toru, M. A Structural Polymorphism of Human Dopamine D_2 receptor, D_2(Ser311→Cys). *Biochem. Biophys. Res. Commun.* **1993**, *196,* 1369-1375.

87. Giros, B.; Martres, M.-P.; Sokoloff, P.; Schwartz, J.-C. Gene Cloning of Human Dopaminergic D_3 Receptor and Identification of its Chromosome. *C. R. Acad. Sci. , III, Sci, Vie* **1990**, *311,* 501-508.
88. Schmauss, C.; Haroutunian, V.; Davis, K.L.; Davidson, M. Selective Loss of Dopamine D_3-type Receptor mRNA Expression in Parietal and Motor Cortices of Patients with Chronic Schizophrenia. *Proc. Natl. Acad. Sci. USA* **1993**, *90,* 8942-8946.
89. Liu, K.; Bergson, C.; Levenson, R.; Schmauss, C. On the Origin of mRNA Encoding the Truncated Dopamine D_3-type Receptor D_3nf and Detection of D_3nf-like Immunoreactivity in Human Brain. *J. Biol. Chem.* **1994**, *269,* 29220-29226.
90. Van Tol, H.; Wu, C.M.; Guan, H.C.; Ohara, K.; Bunzow, J.R.; Civelli, O.; Kennedy, J.; Seeman, P.; Niznik, H.B.; Jovanovic, V. Multiple Dopamine D_4 Receptor Variants in the Human Population [see comments]. *Nature* **1992**, *358,* 149-152.
91. Van Tol, H.; Bunzow, J.R.; Guan, H.-C.; Sunahara, R.K.; Seeman, P.; Niznik, H.B.; Civelli, O. Cloning of the Gene for a Human Dopamine D_4 Receptor with High Affinity for the Antipsychotic Clozapine. *Nature* **1991**, *350,* 610-614.
92. Livingstone, D.J.; Strange, P.G.; Naylor, L.H. Molecular Modelling of D_2-like Dopamine Receptors. *Biochem. J.* **1992**, *287,* 277-282.
93. Seeman, P.; Ulpian, C.; Chouinard, G.; Van Tol, H.; Dwosh, H.; Lieberman, J.A.; Siminovitch, K.; Liu, I.S.C.; Waye, J.; Voruganti, P.; Hudson, C.; Serjant, G.R.; Masibay, A.S.; Seeman, M.V. Dopamine D4 Receptor Variant, D_4GLYCINE$_{194}$, in Africans, but Not in Caucasians: no Association with Schizophrenia. *Am. J. Med. Genet.* **1994**, *54,* 384-390.
94. Dijkstra, D.; Mulder, T.B.; Rollema, H.; Tepper, P.G.; Van der Weide, J.; Horn, A.S. Synthesis and Pharmacology of *trans*-4-n-Propyl-3,4,4a,10b-tetrahydro-2*H*,5*H*-1-benzopyrano[4,3-b]-1,4-oxazin-7- and -9-ols: the Significance of Nitrogen pK_a Values for Central Dopamine Receptor Activation. *J. Med. Chem.* **1988**, *31,* 2178-2182.
95. Glase, S.A.; Akunne, H.C.; Georgic, L.M.; Heffner, T.G.; MacKenzie, R.G.; Manley, P.J.; Pugsley, T.A.; Wise, L.D. Substituted [(4-Phenylpierazinyl)-methyl]benzamides: Selective Dopamine D_4 Agonists. *J. Med. Chem.* **1997**, *40,* 1771-1772.
96. Johansson, A.M.; Fredriksson, K.; Hacksell, U.; Grol, C.J.; Svensson, K.; Carlsson, A.; Sundell, S. Synthesis and Pharmacology of the Enantiomers of *cis*-7-hydroxy-3-methyl-2-(dipropylamino)tetralin. *J. Med. Chem.* **1990**, *33,* 2925-2929.
97. Unangst, P.C.; Capiris, T.; Connor, D.T.; Doubleday, R.; Heffner, T.G.; MacKenzie, R.G.; Miller, S.R.; Pugsley, T.A.; Wise, L.D. (Aryloxy)alkylamines as Selective Human Dopamine D_4 Receptor Antagonists: Potential Antipsychotic Agents. *J. Med. Chem.* **1997**, *40,* 4026-4029.
98. Chapman, D. The Measurement of Molecular Diversity: a Three-Dimensional Approach. *J. Comput. -Aided Mol. Design* **1996**, *10,* 501-512.
99. Broach, J.R. and Thorner, J. High-Throughput Screening for Drug Discovery. *Nature.* **1996**, *384,* 14-16.
100. Verdine, G.L. The Combinatorial Chemistry of Nature. *Nature.* **1996**, *384,* 11-13.
101. Murray, C.W.; Clark, D.E.; Auton, T.R.; Firth, M.A.; Li, J.; Sykes, R.A.; Waszkowycz, B.; Westhead, D.R.; Young, S.C. PRO_SELECT: Combining Structure-Based Drug Design and Combinatorial Chemistry for Rapid Lead Discovery. 1. Technology. *J. Comput. Aided. Mol. Des.* **1997**, *11,* 193-207.
102. Lam, K.S. Application of Combinatorial Library Methods in Cancer Research and Drug Discovery. *Anticancer. Drug. Des.* **1997**, *12,* 145-167.
103. Marshall, G.R.; Barry, C.D.; Bosshard, H.E.; Dammkoehler, R.A.; Dunn, D.A. The Conformational Parameter in Drug Design: the Active Analogue Approach. *ACS Symp. Ser.* **1979**, *112,* 205-226.
104. Dahl, S.G.; Edvardsen, O.; Sylte, I. Molecular Dynamics of Dopamine at the D_2 Receptor. *Proc. Natl. Acad. Sci. USA* **1991**, *88,* 8111-8115.
105. Findlay, J. and Eliopoulos, E. Three-dimensional Modeling of G protein-linked Receptors. *TIPS* **1990**, *11,* 492-499.
106. Hibert, M.; Trumpp-Kallmeyer, S.; Bruinvels, A.; Hoflack, J. Three-Dimensional Models of Neurotransmitter G-Binding Protein-Coupled Receptors. *Mol. Pharmacol.* **1991**, *40,* 8-15.
107. Trumpp-Kallmeyer, S.; Hoflack, J.; Bruinvels, A.; Hibert, M. Modeling of G-Protein-Coupled Receptors: Application to Dopamine, Adrenaline, Serotonin, Acetylcholine, and Mammalian Opsin Receptors. *J. Med. Chem.* **1992**, *35,* 3448-3462.
108. Shertler, G.F.X.; Villa, C.; Henderson, R. Projection Structure of Rhodopsin. *Nature* **1993**, *362,* 770-772.
109. Needleman, S.B. and Wunsch, C.D. A General Method to Applicable to the Search for Similarity in Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48,* 443-453.
110. Dayhoff, H.O.; Schwartz, R.M.; Orcott, B.C. Atlas of Protein Sequence and Structure. NBFR: Washington DC, **1978**;
111. EMBL in Heidelberg. http://swift.embl-heidelberg.de/7tm.
112. The Whatif Program. http://www.sander.embl-heidelberg.de/whatif.

113. Savarese, T.M. and Frazer, C.M. *In Vitro* Mutagenesis and the Search for Structure-Function Relationships Among G Protein-Coupled Receptors. *Biochem. J.* **1992**, *283*, 1-19.
114. Neve, K.A. Regulation of Dopamine D_2 Receptors by Sodium and pH. *Mol. Pharmacol.* **1991**, *39*, 570-578.
115. Neve, K.A.; Cox, B.A.; Henningsen, R.A.; Spannoyannis, A.; Neve, R.L. Pivotal Role for Aspartate-80 in the Regulation of Dopamine D_2 Receptor Affinity for Drugs and Inhibition of Adenyl Cylase. *Mol. Pharmacol.* **1991**, *39*, 733-739.
116. Strader, C.D.; Candelore, M.R.; Hill, W.S.; Sigal, I.S.; Dixon, R.A. Identification of Two Serine Residues Involved in Agonist Activation of the β-Adrenergic Receptor. *J. Biol Chem.* **1989**, *264*, 13572-13578.
117. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analyses (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
118. McDermed, J.D.; Freeman, H.S.; Ferris, R.M. Catecholamines: Basic and Clinical Frontiers. Usdin, E.; Kopin, I.J.; Barchas, J. Eds.; Pergamon Press: New York, **1979**; pp. 568-577.
119. Copinga, S. PhD Thesis. The 2-Aminotetralin System as a Structural Base for New Dopamine- and Melatonin-Receptor Agents, Dept. of Medicinal Chemistry, University of Groningen, The Netherlands. **1994**
120. Van Vliet, L.A.; Tepper, P.G.; Dijkstra, D.; Damsma, G.; Wikström, H.; Pugsley, T.A.; Akunne, H.C.; Heffner, T.G.; Glase, S.A.; Wise, L.D. Affinity for Dopamine D_2, D_3 and D_4 Receptors of 2-Aminotetralins. Relevance of D_2 Agonist Binding for Determination of Receptor Subtype Selectivity. *J. Med. Chem.* **1996**, *39*, 4233-4237.
121. Sonesson, C.; Boije, M.; Svensson, K.; Ekman, A.; Carlsson, A.; Romero, A.G.; Martin, I.J.; Duncan, J.N.; King, L.J.; Wikström, H. Orally Active Central Dopamine and Serotonin Receptor Ligands: 5-, 6-, 7-, and 8-[[Trifluoromethyl)sulfonyl]oxy]-2-(di-n-propylamino)tetralins and the Formation of Active Metabolites *in vivo*. *J. Med. Chem.* **1993**, *36*, 3409-3416.
122. Liljefors, T. and Wikström, H. A Molecular Mechanics Approach to the Understanding of Presynaptic Selectivity for Centrally Acting Dopamine Receptor Agonists of the Phenylpiperidine Series. *J. Med. Chem.* **1986**, *29*, 1896-1904.
123. Liljefors, T.; Bøgesø, K.P.; Hyttel, J.; Wikström, H.; Svensson, K.; Carlsson, A. Pre- and Postsynaptic Dopaminergic Activities of Indolizidine and Quinolizidine Derivatives of 3-(3-Hydroxyphenyl)-*N*-(n-propyl)piperidine (3-PPP). Further Developments of a Dopamine Receptor Model. *J. Med. Chem.* **1990**, *33*, 1015-1022.
124. Wikström, H.; Andersson, B.; Sanchez, D.; Lindberg, P.; Arvidsson, L.E.; Johansson, A.M.; Nilsson, J.L.; Svensson, K.; Hjorth, S.; Carlsson, A. Resolved Monophenolic 2-Aminotetralins and 1,2,3,4,4a,5,6,10b-octahydrobenzo[*f*]quinolines: Structural and Stereochemical Considerations for Centrally Acting Pre- and Postsynaptic Dopamine-Receptor Agonists. *J. Med. Chem.* **1985**, *28*, 215-225.
125. Wikström, H.; Sanchez, D.; Lindberg, P.; Arvidsson, L.; Hacksell, U.; Johansson, A.M.; Nilsson, J.L.; Hjorth, S.; Carlsson, A. Monophenolic Octahydrobenzo[*f*]quinolines: Central Dopamine- and Serotonin-Receptor Stimulating Activity. *J. Med. Chem.* **1982**, *25*, 925-931.
126. Froimowitz, M.; Deng, Y.; Jacob, J.N.; Li, N.; Cody, V. Drug Design and Discovery. Dopaminergic (4a*R*,10b*S*)-*cis*- and (4a*S*,10b*S*)-*trans*-Octahydrobenzoquinolines have Similiar Pharmacophores. Harward Academic Publishers GmbH.: **1995**; pp. 73-81.
127. Horn, A.S.; Tepper, P.G.; Kebabian, J.W.; Beart, P.M. N-O434, a Very Potent and Specific New D_2 Dopamine Receptor Agonist. *Eur. J. Pharmacol.* **1984**, *99*, 125-126.
128. Sonesson, C.; Barf, T.; Nilsson, J.; Dijkstra, D.; Carlsson, A.; Svensson, K.; Smith, M.W.; Martin, I.J.; Duncan, J.N.; King, L.J.; Wikström, H. Synthesis and Evaluation of Pharmacological and Pharmacokinetic Properties of Monopropyl Analogs of 5-, 7-, and 8-[[(Trifluoromethyl)sulfonyl]oxy]-2-aminotetralins: Central Dopamine and Serotonin Receptor Activity. *J. Med. Chem.* **1995**, *38*, 1319-1329.
129. Sonesson, C.; Lin, C.H.; Hansson, L.; Waters, N.; Svensson, K.; Carlsson, A.; Smith, M.W.; Wikström, H. Substituted (*S*)-Phenylpiperidines and Rigid Congeners as Preferential Dopamine Autoreceptor Antagonists: Synthesis and Structure-Activity Relationships. *J. Med. Chem.* **1994**, *37*, 2735-2753.
130. Sheppard, W.A. The Effect on Fluorine Substitution on the Electronic Properties of Alkoxy, Alkylthio and Alkylsulfonyl Groups. *J. Am. Chem. Soc.* **1962**, *85*, 1314-1318.
131. Charifson, P.S.; Wyrick, S.D.; Hoffman, A.J.; Simmons, R.M.; Bowen, J.P.; McDougald, D.L.; Mailman, R.B. Synthesis and Pharmacological Characterization of 1-phenyl-, 4-phenyl-, and 1-benzyl-1,2,3,4-tetrahydroisoquinolines as Dopamine Receptor Ligands. *J. Med. Chem.* **1988**, *31*, 1941-1946.
132. Charifson, P.S.; Bowen, J.P.; Wyrick, S.D.; Hoffman, A.J.; Cory, M.; McPhail, A.T.; Mailman, R.B. Conformational Analysis and Molecular Modeling of 1-phenyl-, 4-phenyl-, and 1-benzyl-1,2,3,4-tetrahydroisoquinolines as D_1 Dopamine Receptor Ligands. *J. Med. Chem.* **1989**, *32*, 2050-2058.
133. Mottola, D.M.; Laiter, S.; Watts, V.J.; Tropscha, A.; Wyrick, S.D.; Nichols, D.E.; Mailman, R.B. Conformational Analysis of D_1 Dopamine Receptor Agonists: Pharmacophore Assessment and Receptor Mapping. *J. Med. Chem.* **1996**, *39*, 285-296.

134. Pettersson, I.; Liljefors, T.; Bøgesø, K. Conformational Analysis and Structure-Activity Relationships of Selective Dopamine D_1 Receptor Agonists and Antagonists of the Benzazepine Series. *J. Med. Chem.* **1990**, *33,* 2197-2204.
135. Pettersson, I.; Gundertofte, K.; Palm, J.; Liljefors, T. A Study on the Contribution of the 1-phenyl Substituent to the Molecular Electrostatic Potentials of some Penzazepines in Relation to Selective Dopamine D_1 Receptor Activity. *J. Med. Chem.* **1992**, *35,* 502-507.
136. GRID, Goodford, P.J. Molecular Discovery Ltd, University of Oxford, England, SGI.
137. Rognan, D.; Sokoloff, P.; Mann, A.; Martres, M.P.; Schwartz, J.C.; Costentin, J.; Wermuth, C.G. Optically Active Benzamides as Predictive Tools for Mapping the Dopamine D_2 Receptor. *Eur. J. Pharmacol* **1990**, *189,* 59-70.
138. Olson, G.L.; Cheung, H.C.; Morgan, K.D.; Blount, J.F.; Todaro, L.; Berger, L.; Davidson, A.B.; Boff, E. A Dopamine Receptor Model and its Application in the Design of a New Class of Rigid Pyrrolo[2,3-g]isoquinoline Antipsychotics. *J. Med. Chem.* **1981**, *24,* 1026-1034.
139. Högberg, T. Novel Substituted Salicylamides and Benzamides as Selective D_2-Receptor Antagonists. *Drugs Fut.* **1991**, *16,* 333-357.
140. Nilsson, J.; Wikström, H.; Smilde, A.K.; Glase, S.; Pugsley, T.A.; Cruciani, G.; Pastor, M.; Clementi, S. A GRID/GOLPE 3D-QSAR Study on a Set of Benzamides and Naphthamides, with Affinity for the Dopamine D_3 Receptor Subtype. *J. Med. Chem.* **1997**, *40,* 833-840.
141. Nilsson, J.; De Jong, S.; Smilde, A.K. Multiway Calibration in 3D QSAR. *J. of Chemometrics* **1997**, *11,* 511-524.
142. SYBYL- Molecular Modeling Software, *6.3,* Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA,
143. MacroModel-Interactive Molecular Modeling System, *4.5,* Mohamadi, F.; Richards, N.G.J.; Guida, W.C.; Liskamp, R.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W.C. New York, Macromodel Program v.4.5
144. Computer-Assisted Lead Finding and Optimization. Current Tools for Medicinal Chemistry. van de Waterbeemd, H.; Testa, B.; Folkers, G. Eds.; VHCA and Wiley: Weinheim, **1997**
145. Gschwend, D.A.; Good, A.C.; Kuntz, I.D. Molecular Docking Towards Drug Discovery. *J. Mol. Recognit.* **1996**, *9,* 175-186.
146. Fersht, A.R.; Knill Jones, J.W.; Bedouelle, H.; Winter, G. Reconstruction by Site-Directed Mutagenesis of the Transition State for the Activation of Tyrosine by the Tyrosyl-tRNA Synthetase: a Mobile Loop Envelopes the Transition State in an Induced-Fit Mechanism. *Biochemistry.* **1988**, *27,* 1581-1587.
147. Hansch, C. and Fujita, T. rho-sigma-pi Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86,* 1616-1626.
148. Boyd, S.M.; Beverley, M.; Norskov, L.; Hubbard, R.E. Characterising the Geometric Diversity of Functional Groups in Chemical Databases. *J. Comput. -Aided Mol. Design* **1995**, *9,* 417-424.
149. Shemetulskis, N.E.; Dunbar Jr, J.B.; Dunbar, B.W.; Moreland, D.W.; Humblet, C. Enahancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput. -Aided Mol. Design* **1995**, *9,* 407-416.
150. Albano, C.; Lundstedt, T.; Carlsson, R. Screening of Suitable Solvents in Organic Synthesis. Strategies for Solvent Selection. *Acta Chem. Scand.* **1985**, 79-91.
151. Collin, S.; Tayar, N.A.; Van de Waterbeemd, H.; Moureau, F.; Vercauteren, D.P.; Durant, F.; Langlois, M.; Testa, B. QSAR of Nortropane-Substituted Benzamides: Use of Lipophilic (RP-HPLC) and Electronic (^{1}H-NMR) Parameters. *Eur. J. Med. Chem* **1989**, *24,* 163-169.
152. Tayar, N.e.; Kilpatrick, G.J.; Van de Waterbeemd, H.; Testa, B.; Jenner, P.; Marsden, C.D. Interaction of Neuroleptic drugs with Rat Striatal D_1 and D_2 Dopamine Receptors: a Quantitaive Structure-Activity Relationship Study. *Eur. J. Med. Chem* **1988**, *23,* 173-182.
153. Hansch, C. and Leo, A. Substituent Constants for Correlation Analysis in Chemistry and Biology. Wiley: New York, **1979**
154. Kessler, R.M.; Sib Ansari, M.; De Paulis, T.; Schmidt, D.E.; Clanton, J.A.; Smith, H.E.; Manning, R.G.; Gillespie, D.; Ebert, M.H. High Affinty Dopamine D_2 Receptor Radioligands. 1. Regional Rat Brain Distribution of Iodinated Benzamides. *J. Nuc. Med.* **1991**, *32,* 1593-1600.
155. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185,* 1-17.
156. Eriksson, L. and Johansson, E. Multivariate Design and Modeling in QSAR. *Chemom. and Intell. Lab. Syst.* **1996**, *34,* 1-19.
157. Chemometric Methods in Molecular Design. van de Waterbeemd, H. Ed.; VCH: Weinheim, **1995**
158. Nilsson, J.; Selditz, U.; Pugsley, T.A.; Sundell, S.; Lundmark, M.; Smilde, A.K.; Wikström, H. Design, Syntheses and QSAR of a Series *trans*-1,2,3,4,4a,5,6,10b-Octahydrobenzo[*f*]quinolines with Dopaminergic Affinity. *submitted* **1997**,
159. Norinder, U. and Högberg, T. A Quantitative Structure-Activity Relationship for some Dopamine D_2 Antagonists of Benzamide Type. *Acta Pharm. Nord.* **1992**, *4,* 73-78.

160. Bro, R. Multiway Calibration. Multilinear PLS. *J. of Chemometrics* **1996**, *10,* 47-61.
161. Cramer III, R.D. and Wold, S. inventors. Comparative Molecular Field Analyses (COMFA). 5025388. United States. Date Filed: **1988/08/26.**
162. Pallas 1.2, *1.2,* CompuDrug Chemistry Ltd. Program for calculation of logP, logD, and pKa, MS-Windows 3.1.
163. Hansch, C.; Leo, A.; Taft, R.W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91,* 165-195.
164. Ong, V.S. and Hites, R.A. Relationship Between Gas Chromatographic Retention Indexes and Computer-Calculated Physical Properties of Four Compound Classes. *Anal. Chem.* **1991**, *63,* 2829-2834.
165. Van de Waterbeemd, H.; el Tayar, N.; Testa, B.; Wikström, H.; Largent, B. Quantitative Structure-Activity Relationships and Eudismic Analyses of the Presynaptic Dopaminergic Activity and Dopamine D_2 and sigma Receptor Affinities of 3-(3-hydroxyphenyl)piperidines and octahydrobenzo[*f*]quinolines. *J. Med. Chem.* **1987**, *30,* 2175-2181.
166. Moore, W.J. Basic Physical Chemistry. Prentice-Hall: London, **1983**
167. Goodford, P.J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28,* 849-857.
168. Hopfinger, A.J. Conformational Properties of Macromolecules. 2, Chapter 2. Academic Press.: New York, **1973**
169. Gasteiger, J. and Marsili, M. *Tetrahedron* **1980**, *36,* 3219
170. Dewar, M.J.S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107,* 3902-3909.
171. Kroemer, R.T. and Hecht, P. Replacement of Steric 6-12 Potential-Derived Interaction Energies by Atom-based Indicator Variables in CoMFA Leads to Models of Higher Consistency. *J. Comput. -Aided Mol. Design* **1995**, *9,* 205-212.
172. Floersheim, P. Trends in QSAR and Molecular Modelling 92. Escom: Leiden, **1993**; pp. 227
173. Goodford, P.J. Multivariate Characterization of Molecules for QSAR Analyses. *J. of Chemometrics* **1996**, *10,* 107-117.
174. Green, S.M. and Marshall, G.R. 3D-QSAR: A Current Perspective. *TIPS* **1995**, *16,* 285
175. Dunn III, W.J.; Hopfinger, A.J.; Catana, C.; Duraiswami, C. Solution of Conformation and Alignment Tensors for the Binding of Trimethoprim and Its Analogs to Dihyrofolate Reductase: 3D-Quantitative Structure-Activity Relationship Study Using Molecular Shape analysis, 3-Way Partial Least-Squares Regression, and 3-Way Factor Analysis. *J. Med. Chem.* **1996**, *39,* 4825-4832.
176. Hopfinger, A.J. Theory and Application of Molecular Potential Energy Fields in Molecular Shape Analysis: A Quantitative Structure-Activity Relationship of 2,4-Diamino-5-Benzylpyrimidines as Dihydrofolate Reductase Inhibitiors. *J. Med. Chem.* **1983**, *26,* 990-996.
177. Gaillard, P.; Carrupt, P.A.; Testa, B.; Boudon, A. Molecular Lipophilicity Potential, a Tool in 3D QSAR: Method and Applications. *J. Comput. Aided. Mol. Des.* **1994**, *8,* 83-96.
178. Gaillard, P.; Carrupt, P.A.; Testa, B.; Schambel, P. Binding of Arylpiperazines, (Aryloxy)propanolamines, and Tetrahydropyridylindoles to the 5-HT_{1A} Receptor: Contribution of the Molecular Lipophilicity Potential to Three-Dimensional Quantitative Structure-Affinity Relationship Models. *J. Med. Chem.* **1996**, *39,* 126-134.
179. Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, *37,* 4130-4146.
180. Good, A.C.; So, S.S.; Richards, W.G. Structure-Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36,* 433-438.
181. Good, A.C.; Peterson, S.J.; Richards, W.G. QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36,* 2929-2937.
182. Good, A.C.; Hodgkin, E.E.; Richards, W.G. Similarity Screening of Molecular Data Sets. *J. Comput. Aided. Mol. Des.* **1992**, *6,* 513-520.
183. Howard, A.E. and Kollman, P.A. An Analysis of Current Methodologies for Conformational Searching of Complex Molecules. *J. Med. Chem.* **1988**, *31,* 1669-1675.
184. Saunders, M.; Houk, K.N.; Wu, Y.-D.; Still, W.C.; Lipton, M.; Chang, G.; Guida, W.C. Conformations of Cycloheptadecane. A Comparison of Methods for Conformational Searching. *J. Am. Chem. Soc.* **1990**, *112,* 1419-1427.
185. Lipton, M. and Still, W.C. The Multiple Minimum Problem in Molecular Modeling. Tree Searching Internal Coordinate Conformational Space. *J. Comp. Chem.* **1988**, *9,* 343-355.
186. Chang, G.; Guida, W.C.; Still, W.C. An Internal Coordinate Monte Carlo Method for Searching Conformational Space. *J. Am. Chem. Soc.* **1989**, *111,* 4379-4386.
187. Davis, A.M.; Gensmantel, N.P.; Johansson, E.; Marriott, D.P. The Use of the GRID Program in the 3-D QSAR Analysis of a Series of Calcium-Channel Agonists. *J. Med. Chem.* **1994**, *37,* 963-972.
188. Wold, H. Systems Under Indirect Observation. The Basic Design and Some Extensions. Jöreskog, K. and Wold, H. Eds. North-Holland.: Amsterdam, **1982**

189. Wold, S.; Ruhe, A.; Wold, H.; Dunn III, W.J. *J. Sci. Stat. Comput.* **1984**, *5*, 735-743.
190. Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons: New York, **1989**

Introduction to Chemometrics and Statistics

2

2.1 Introduction

In Quantitative Structure-Activity Relationships (QSAR), molecular descriptors (**X**) are correlated with one or more response variable (**y**). The objective with the analysis usually is to increase the understanding of the biological system under investigation, or to predict the response of objects not yet tested (*e.g.*, predict the potency of a compound not yet synthesized). The conclusions drawn from a regression analysis are dependent on the assumption of the regression model.[1] If it is assumed that the relationship is well represented by a model that is linear in the regressor variables, a suitable model may be

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \mathbf{e} \tag{2.1}$$

In Equation 2.1 the *b*s are unknown constants called regression coefficients and the objective of regression analysis is to estimate these constants. The number of available regression methods is large and in QSAR, during the last decade, Multiple Linear Regression[1,2] has more or less been replaced by Partial Least-Squares regression[2,3] and other related projection methods.

Throughout this thesis, scalars are written as italic characters; vectors as boldface lower case characters; matrices as boldface upper case characters and multiway matrices as underlined upper case characters. The lower case italic characters *i*, *j*, *k*, *l* and *m* will be used as running indices, where $i = 1,\ldots,I$; $j = 1,\ldots,J$; $k = 1,\ldots,K$; $l = 1,\ldots,L$ and $m = 1,\ldots,M$. It is assumed that all vectors are column vectors.

2.2 Data Pretreatment

It is well known in regression analysis that a proper pretreatment is crucial for the outcome of the result. QSAR data sets consist of variables that differ in range, variation and size. Consequently, prior to regression analysis auto-scaling is usually applied (Figure 2.1), *i.e.*, the *i*th column is mean-centered (with $\bar{\mathbf{x}}_i$) and scaled with $1/\mathrm{sd}(\mathbf{x}_i)$.[2]

In CoMFA the descriptors are divided in blocks, field-wise, which renders auto-scaling, as in Figure 2.1, without meaning. Instead, the total variation in whole fields are standardized or block-scaled. In SYBYL, block-scaling is called CoMFA_std scaling. However, if GRID descriptors are used, block-scaling must be carried out cautiously, since the type of interactions are identical in all fields. In SYBYL/CoMFA the steric and the electrostatic fields are calculated differently and, therefore, block-scaling makes sense.

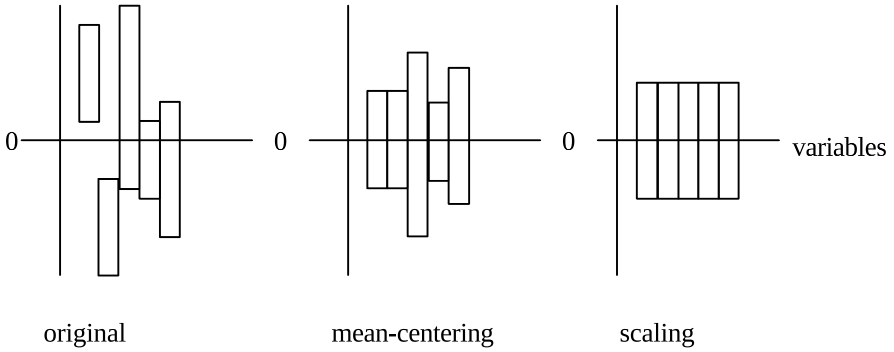


Figure 2.1 Auto-scaling as usually performed prior to QSAR analysis. First column mean-centering followed by column-wise scaling with the inverse of the corresponding standard deviation.

In the GOLPE-program[4] additional pretreatment options are possible: The high and low cut-off values for the interactions may be altered; interactions with absolute values lower than a specified value may be set to zero and grid points with standard deviation lower than a specified value may be omitted.

2.3 Multiple Linear Regression (MLR)

In order to establish a relationship between **X** and **y** in Figure 2.2, Multiple Linear Regression (MLR)[1] has until recently been the obvious method of choice. In MLR, it is assumed that **X** is of full rank and the x_{ij} are measured with negligible error. The algebraic MLR model is defined in Equation 2.1 and in matrix notation

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{2.2}$$

where $\mathbf{X} = [\mathbf{x}_0|\mathbf{x}_1|\ldots\mathbf{x}_J]$, $\mathbf{b}^T = [b_0, b_1, \ldots, b_J]$ and **e** is an error vector. Note that the first column in **X**, *i.e.*, $\mathbf{x}_0$ consists of only constants which, after mean-centering, becomes zero and consequently $\mathbf{x}_0$ is omitted. When **X** is of full rank the least squares solution is:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.3}$$

where $\hat{\mathbf{b}}$ is the estimator for the regression coefficients in **b**. An obvious disadvantage using MLR as regression method in QSAR is: when $I \leq J$ (Figure 2.2) **X** is not of full rank and $(\mathbf{X}^T\mathbf{X})^{-1}$ in Equation 2.3, is not defined and **b** can not be estimated. In the following section the problem with multicollinearity,[1] *i.e.*, the case when **X** not is of full rank, will be discussed.

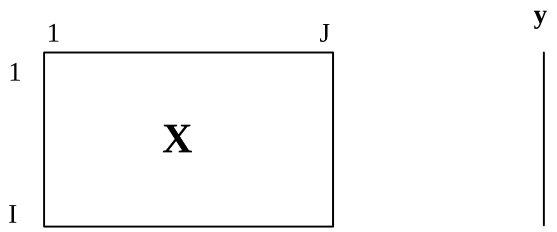


Figure 2.2 A typical QSAR data set: $\mathbf{X}$ is of the dimensions $I \times J$ where $J > I$ with a single response variable $\mathbf{y}$ $(I \times 1)$.

2.4 Multicollinearity

In the previous section, the potential danger of multicollinearity in combination with MLR was mentioned. Multicollinearity is present when the columns of $\mathbf{X}$ are approximately or exactly linearly dependent. In the case of exact linear dependency, $(\mathbf{X}^T\mathbf{X})^{-1}$ is not defined, and the estimation of the regression coefficients $\hat{\mathbf{b}}$ can not be expressed as in Equation 2.3 anymore.

If the linear dependency is approximate, assuming $\mathbf{X}$ is properly auto-scaled, at least one of the diagonal elements in the inverse covariance matrix, $(\mathbf{X}^T\mathbf{X})^{-1}$, will be large. Additionally, some of the diagonal elements of cov($\hat{\mathbf{b}}$), well-known to be $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ (where σ^2 is $(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})/(I-J)$ $(I > J)$),[1,3] may be large, indicating that some *b*s in $\hat{\mathbf{b}}$ are estimated with low precision. Consequently, multicollinearity may influence the interpretation of the model and affect external predictions detrimentally. Therefore, it is important to be able to detect whether $\mathbf{X}$ is collinear or not, prior to regression analysis.

The inverse covariance matrix, $(\mathbf{X}^T\mathbf{X})^{-1}$, provides a first indication of ill-conditioning (multicollinearity) among the variables in $\mathbf{X}$. Another commonly used indication of multicollinearity is the variance inflation factor (VIF)[1,5]:

$$\mathrm{VIF}_i = 1\big/\left(1-\mathrm{R}_i^2\right) \qquad (2.4)$$

where R_i^2 is the squared multiple correlation coefficient when $\mathbf{x}_i$ (the *i*th variable in $\mathbf{X}$) is regressed on the remaining variables. When the columns of $\mathbf{X}$ are close to linear dependence (*i.e.*, when the determinant of $\mathbf{X}^T\mathbf{X}$ is close to zero), R_i^2 will be close to unity and VIF_i will be large. In the ideal case, when $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, *i.e.*, when the variables in $\mathbf{X}$ are orthogonal, the VIF for the *i*th variable is unity. Thus, the VIF measures the increase (inflation) of the variance, for each variable, compared to the ideal case. A flag of warning is raised when VIF is greater than five, as suggested by Smilde.[5,6]

The condition index or number (ϕ) is defined as:

$$f = \frac{l_{\mathrm{max}}^{0.5}}{l_{\mathrm{min}}^{0.5}} \qquad (2.5)$$

where λ_{max} and λ_{min} represent the largest and the smallest eigenvalue, respectively, of $\mathbf{X}^T\mathbf{X}$ (scaled and centered $\mathbf{X}$). When $\mathbf{X}$ is ill-conditioned, at least one eigenvalue will be close to zero and, consequently, ϕ becomes large. As a rule of thumb, when ϕ exceeds 100, the effect of multicollinearity may be significant.

The influence of multicollinearity in QSAR is well known, and disqualified MLR as regression method years ago. In a chemical system, controlled by variables that are easily manipulated, an experimental design[7] may be a solution to avoid multicollinearity. In QSAR, however, the objects are generally molecules which make an experimental design complicated. Instead, methods to replace the original descriptors with underlying latent variables,[2,3,8,9] *e.g.*, PCR and PLS, have been developed.

2.5 Principal Component Regression (PCR)

In Principal Component Analysis (PCA), the original descriptors (Figure 2.3) are replaced by Principal Components (PCs) which are linear combinations of the columns in **X**. The extraction of PCs from **X** or the decomposition of **X** is algebraically expressed as

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \ldots + \mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \mathbf{E} \qquad (2.6)$$

Thus, the purpose of PCA simply is to decompose $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\ldots|\mathbf{x}_J]$ into A component score vectors ($\mathbf{T} = [\mathbf{t}_1|\mathbf{t}_2|\ldots|\mathbf{t}_A]$) and loading vectors ($\mathbf{P} = [\mathbf{p}_1|\mathbf{p}_2|\ldots|\mathbf{p}_A]$) where $A < J$. Prior to PCA, the variables in **X** usually are column mean-centered and often scaled to similar variation levels, *i.e.*, auto-scaled.

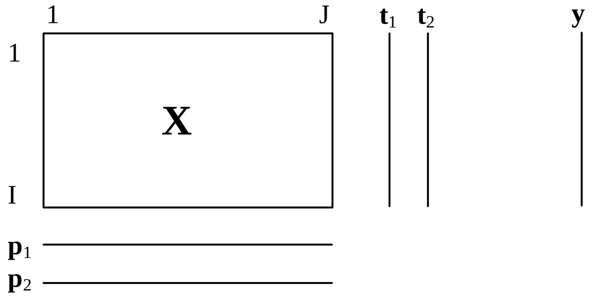


Figure 2.3 A graphical representation of the first two Principal Components. In PCR the component scores $\mathbf{T} = [\mathbf{t}_1|\mathbf{t}_2]$ are the regressors and the following holds true: $\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_2 = 0$ and $\mathbf{p}_1^{\mathrm{T}}\mathbf{p}_2 = 0$.

Each consecutive PC is chosen orthogonal to all the previous PCs ($\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_j = 0$) and accounts for a decreasing percentage of the variation in **X**. In addition, also orthogonal loading vectors ($\mathbf{p}_i^{\mathrm{T}}\mathbf{p}_j = 0$) are obtained which are scaled to be of length one ($\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}_A$). It can be shown that the loading vectors $\mathbf{p}_a$ ($a = 1,\ldots,A$) are eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ with $\boldsymbol{\lambda} = [\lambda_1, \lambda_2,\ldots, \lambda_A]$ as eigenvalues. Algebraically, this is recognized as $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{p}_a = \mathbf{p}_a\lambda_a$. The eigenvalues show how much of the variance the respective components account for. Similarly, it can be shown that the scores $\mathbf{t}_a$ ($a = 1,\ldots,A$) represent the eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, scaled to length $\lambda_a^{0.5}$.

The first few PCs may be considered as a proper representation of **X** since the variation not accounted for is assumed to represent only insignificant variation or noise. The obtained loading-vectors are important for the understanding of, *e.g.*, which **X** variables are important (large loadings), which **X** variables carry the same information and for the interpretation of the component scores (see Chapter 3). The score-vectors contain information about the similarities and dissimilarities between the objects (*e.g.*, compounds) included in the model.

In order to compute the PCA, the NIPALS algorithm[3,10] is often used where the component scores and loadings are calculated one component at a time. The second component is calculated from the residuals after the first component ($\mathbf{X}_1 = \mathbf{X}$):

$$\mathbf{X}_2 = \mathbf{X}_1 - \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} \tag{2.7}$$

Since the component scores (**T**) are orthogonal and account for most of the variation in **X** they are suitable as regressors for **y** using MLR. Accordingly, the following Principal Component Regression (PCR) model may be introduced

$$\mathbf{y} = q_1\mathbf{t}_1 + q_2\mathbf{t}_2 + \ldots + q_A\mathbf{t}_A + \mathbf{e} \tag{2.8}$$

where the qs are the regression coefficients describing the relationship between the response variable (**y**) and the A component scores (**T**). Analogous to MLR, the least squares solution for the estimation of **q** is

$$\hat{\mathbf{q}} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{2.9}$$

where $\hat{\mathbf{q}}$ is the estimator for the regression coefficients in **q**. If $\hat{\mathbf{q}}$, in Equation 2.9, could be expressed in regression coefficients **b**, used in Equation 2.1, the interpretation of the PCR model would be simplified. The regression coefficients **b** can be estimated, as was suggested by Martens and Næs,[3] from:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{T}\hat{\mathbf{q}} \tag{2.10}$$

By replacing **T** with **XP (P** is the loading matrix containing the A loading vectors) it is clear that one possible solution of $\hat{\mathbf{b}}$ is

$$\hat{\mathbf{b}} = \mathbf{P}\hat{\mathbf{q}} \tag{2.11}$$

but, due to the near singularity of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ Equation 2.11 does not provide an unique solution. In QSAR, Equation 2.11 can be used for external predictions but may also be utilized for interpretation purposes.

2.6 Partial Least Squares Regression (PLS)

The Principal Components describe the latent structure of **X** which, accordingly, can be used as regressors for **y** in PCR. In PLS, however, **y** is included in the decomposition procedure and a loading vector, *i.e.*, the weight vector $\mathbf{w}_a$ ($a = 1,\ldots,A$), that maximizes $\mathbf{w}_a^{\mathrm{T}}\mathbf{X}_{a-1}^{\mathrm{T}}\mathbf{y}$ under the constraint that $\mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a = 1$, is searched for. ($\mathbf{X}_{a-1}$ contains the residuals from the previous component as soon will be clear.)

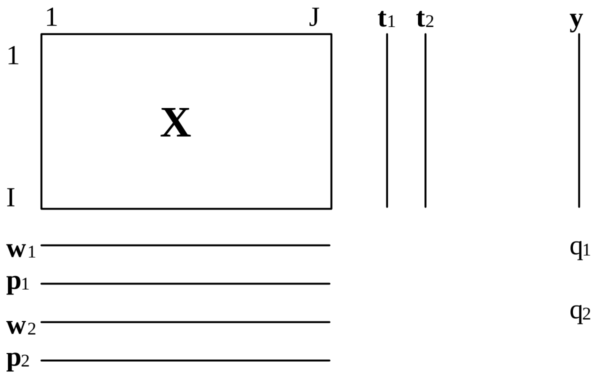


Figure 2.4 A graphical representation of the first two PLS components. In Wolds PLS algorithm, the following holds true: $\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_2 = 0$, $\mathbf{w}_1^{\mathrm{T}}\mathbf{w}_2 = 0$ and $\mathbf{p}_1^{\mathrm{T}}\mathbf{p}_2 \neq 0$.

This is the general definition of PLS (Partial Least Squares Regression),[2,3] where *partial* indicates that the least squares solution applies when **w** is determined as defined above, and not for general **w**.[11] In the following only PLS1 will be considered, *i.e.*, PLS with only one response variable, although PLS can handle multiple response variables simultaneously, *i.e.*, PLS2.

The original PLS algorithm, as presented by Wold *et al.*,[8,9] starts by estimating the weight vector $\mathbf{w}_a$ for the *a*th component, as the vector that maximizes the expression $\mathbf{w}_a^{\mathrm{T}}\mathbf{X}_{a-1}^{\mathrm{T}}\mathbf{y}$:

$$\mathbf{w}_a = \mathbf{X}_{a-1}^{\mathrm{T}}\mathbf{y} \Big/ \left\| \mathbf{X}_{a-1}^{\mathrm{T}}\mathbf{y} \right\| \tag{2.12}$$

Accordingly, the score vector **t** is determined as:

$$\mathbf{t}_a = \mathbf{X}_{a-1}\mathbf{w}_a \big/ \mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a \tag{2.13}$$

and since $\mathbf{w}_a^{\mathrm{T}}\mathbf{w}_a = 1$ Equation 2.13 simplifies to $\mathbf{t}_a = \mathbf{X}_{a-1}\mathbf{w}_a$. The loading vector $\mathbf{p}_a$, necessary for the calculation of new model residuals, is obtained by regression of **X** on $\mathbf{t}_a$:

$$\mathbf{p}_a = \mathbf{X}_{a-1}^{\mathrm{T}}\mathbf{t}_a \big/ \mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a \tag{2.14}$$

In order to make estimations of **y** from $\mathbf{t}_a$ possible, the regression coefficient q_{a} for the *a*th component is needed, which is determined by regression of **y** on $\mathbf{t}_a$

$$q_a = \mathbf{y}^{\mathrm{T}}\mathbf{t}_a \big/ \mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a \tag{2.15}$$

Finally, new residuals $\mathbf{X}_a$ are calculated by subtracting the effect of the previous component:

$$\mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}} \tag{2.16}$$

Analogous to PCR, the regression coefficients $\mathbf{b}_{\mathrm{PLS}}$[3] are useful for the interpretation of the PLS model and for predictions of external objects ($\mathbf{X}_{\mathrm{new}}$) as $\hat{\mathbf{y}} = \mathbf{X}_{\mathrm{new}}\hat{\mathbf{b}}_{\mathrm{PLS}}$. The $\mathbf{b}_{\mathrm{PLS}}$ coefficients are calculated after *A* components as

$$\mathbf{b}_{\mathrm{PLS}} = \mathbf{W}\left(\mathbf{P}^{\mathrm{T}}\mathbf{W}\right)^{-1}\mathbf{q} \tag{2.17}$$

where **W** is $(\mathbf{w}_1|\mathbf{w}_2|\ldots|\mathbf{w}_A)$, $\mathbf{P} = (\mathbf{p}_1|\mathbf{p}_2|\ldots|\mathbf{p}_A)$ and $\mathbf{q}^{\mathrm{T}} = (q_1,\ldots,q_{\mathrm{A}})$.

This algorithm is also called the orthogonalized PLS algorithm, since the estimated score and weight vectors are orthogonal, *i.e.*, $\mathbf{t}_i^{\mathrm{T}}\mathbf{t}_j = 0$ and $\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = 0$ where $i \neq j$ (Figure 2.4). If the number of components *A* extracted equals the number of columns in **X** ($J \leq I$; see Figure 2.4) the PLS solution is the MLR solution.

The second algorithm presented by Martens *et al.*[3] differs from Wold's algorithm since the components scores and weights are not orthogonal. Consequently, the internal regression coefficients **q** can not be calculated component-wise, but instead all components must be calculated simultaneously:

$$\mathbf{q} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{2.18}$$

In Martens non-orthogonalized algorithm, the residuals are calculated by subtracting $\mathbf{t}_a\mathbf{w}_a^{\mathrm{T}}$ from the previous component, as opposed to Equation 2.16 in Wold's algorithm. However, the predictions are identical in both algorithms, although in Martens' algorithm Equation 2.17 simplifies to

$$\mathbf{b}_{\mathrm{PLS}} = \mathbf{W}\mathbf{q} \tag{2.19}$$

2.7 PARAFAC and Tucker Decomposition

In PCA, a two-dimensional matrix **X** is decomposed into score (**T**) and loading (**P**) vectors (Figure 2.3) corresponding to the objects and the descriptors, respectively. Typically in Hansch analysis, *I* objects, *e.g.*, molecules, are described by *J* descriptors, *e.g.*, physicochemical parameters. Sometimes it is more convenient to view the raw data from a series of experiments in the form of a cube, *e.g.*, in analytical chemistry the retention time for I molecules are investigated by RP-HPLC on *J* different columns and *K* different mobile phase mixtures[12] or in pharmacology the intrinsic efficacy of *I* drugs are investigated at *J* different doses monitored for two hours every 15 minutes ($K = 9$). Normally, in order to analyze data sets that are arranged in three modes ($\underline{\mathbf{X}}$; $I \times J \times K$), with PCA, MLR or PLS, they are unfolded to form a two dimensional ($\mathbf{X}$; $I \times JK$) matrix, as in Figure 2.5.

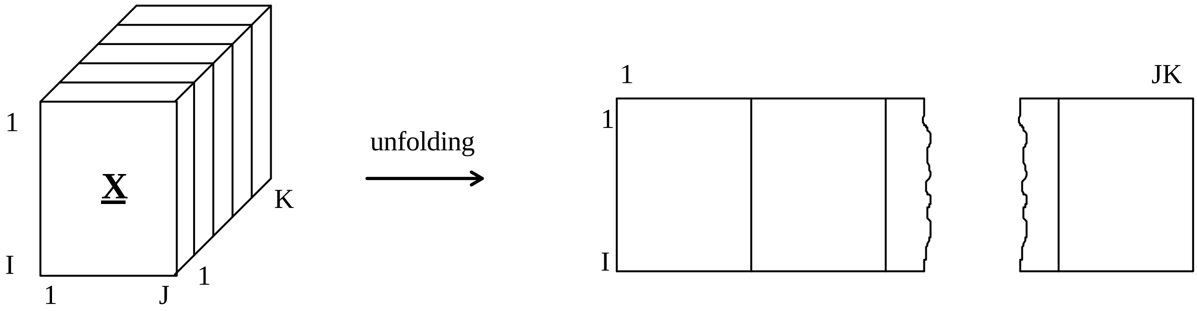


Figure 2.5 The unfolding of a three-way matrix, $\underline{\mathbf{X}}$ ($I \times J \times K$), into a two-way matrix $\mathbf{X}$ ($I \times JK$).

Today, alternatives to the PCA decomposition in Figure 2.2 are available, where the three-way or multiway structure of the data are maintained. In Figure 2.6 the one component PARAFAC[13,14] decomposition of $\underline{\mathbf{X}}$ ($I \times J \times K$) into three loading vectors $\mathbf{t}$ ($I \times 1$), $\mathbf{w}^J$ ($J \times 1$) and $\mathbf{w}^K$ ($K \times 1$) is presented graphically. Note that in multiway analysis, it is not always obvious which directions in $\underline{\mathbf{X}}$ correspond to objects and variables. Therefore, the general term 'mode' will be used in the following.

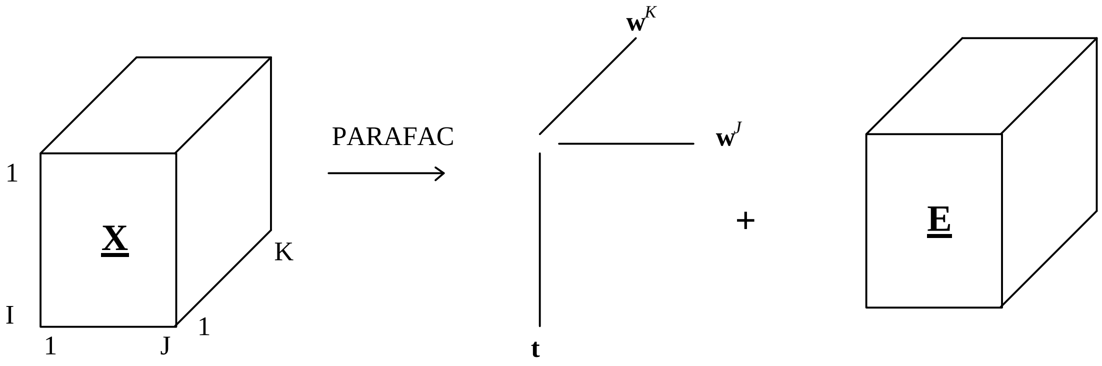


Figure 2.6 The one component PARAFAC decomposition of **X** ($I \times J \times K$) into three loading vectors **t** ($I \times 1$), $\mathbf{w}^J$ ($J \times 1$) and $\mathbf{w}^K$ ($K \times 1$). **E** ($I \times J \times K$) represents the residuals.

The PARAFAC model in Figure 2.6 is expressed formally in Equation 2.20 where t_{ia}, w_{ja} and w_{ka} are typical elements in the loading matrices **T** ($I \times A$), $\mathbf{W}^J$ ($J \times A$) and $\mathbf{W}^K$ ($K \times A$), respectively. A is the number of components extracted. Analogous to PCA, the loading matrices are chosen in such a way that the sum of squared residuals (**E**) is minimized. The PARAFAC solution is unique in the sense that every rotation of the loading matrices destroys the minimum sum of squares of the optimal solution.

$$x_{ijk} = \sum_{a=1}^{A} t_{ia} w_{ja} w_{ka} + e_{ijk} \tag{2.20}$$

The PARAFAC model actually is a special case of the Tucker[15,16] model differing only in the core-matrix **Z** in Figure 2.7. In a Tucker model, interactions between loading vectors from latent variables of the different modes are allowed. This is not the case for PARAFAC models. Hence, PARAFAC can be seen as a restricted version of a Tucker model. Consequently, a PARAFAC model is a Tucker model where only the superdiagonal of the **Z** matrix is non-zero.

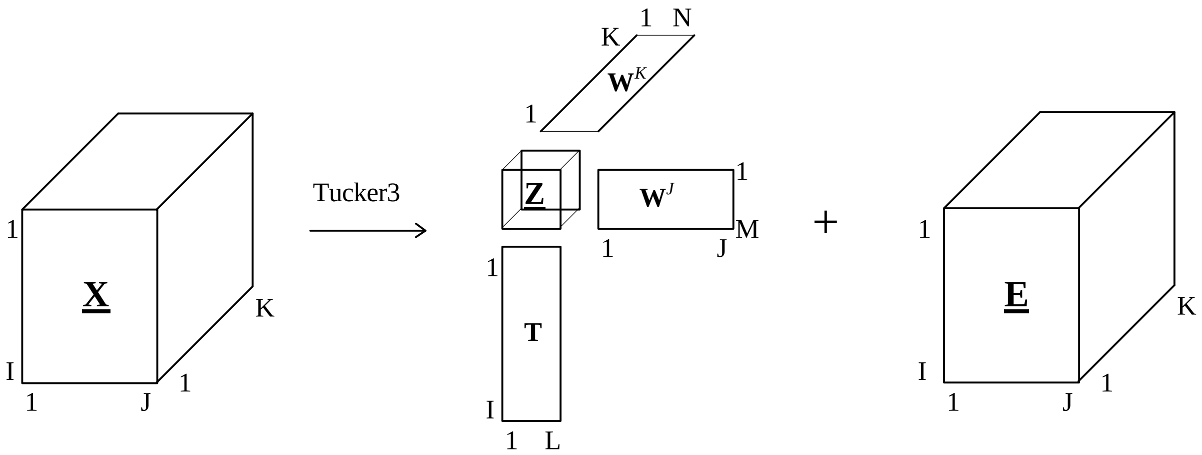


Figure 2.7 The Tucker3 decomposition of **X** ($I \times J \times K$) into three loading matrices and one core matrix **Z** ($L \times M \times N$) containing the interactions between the components in different modes. The **E** ($I \times J \times K$) matrix represents the residuals.

The model in Figure 2.7 is called a Tucker3 model, since $\underline{\mathbf{X}}$ is compressed in all three modes. Extensions to a higher number of modes are possible, as will be discussed in Chapter 8. Formally, the Tucker3 model is written as:

$$x_{ijk} = \sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N} t_{il} w_{jm} w_{kn} z_{lmn} + e_{ijk} \tag{2.21}$$

where t_{il}, w_{jm} and w_{kn} are typical elements in the loading matrices $\mathbf{T}$ ($I \times L$), $\mathbf{W}^J$ ($J \times M$) and $\mathbf{W}^K$ ($K \times N$), respectively. The z_{lmn} is a typical element in the core matrix $\underline{\mathbf{Z}}$ ($L \times M \times N$) and e_{ijk} is a typical element in the residual matrix $\underline{\mathbf{E}}$. The matrices $\mathbf{T}$, $\mathbf{W}^J$, $\mathbf{W}^K$ and $\underline{\mathbf{Z}}$ are chosen in such a way that the sum of squared residuals are minimized. In contrast to the PARAFAC solution the Tucker solution is not unique. Hence, rotations of the loading matrices and rotations of the core matrix can give another solution, with the same sum of squared residuals, as the one found.

As mentioned earlier, interactions between components from different modes are accounted for by the elements of the core matrix. The number of components may be chosen differently in each mode, *i.e.*, L, M and N are not necessarily equal.

Pretreatment of multiway data is not that straightforward, as in two-way data. However, it is common to keep one mode intact and then mean-center the unfolded multiway matrix column-wise. Scaling is more delicate and may be performed block-wise[17,18] by equalizing the sum of squares of whole blocks or by scaling of sub-regions in a (hyper-)cube[19] rather than column-wise scaling (*e.g.*, auto-scaling).

2.8 Multilinear PLS Regression (N-PLS)

The multilinear PLS[19] algorithm (N-PLS) is an extension of Martens' two-way PLS[3] (Section 2.6) in combination with the PARAFAC decomposition[13,14] (Section 2.7) to data of higher orders. In traditional QSAR, a PLS model between a two-way descriptor block (**X**) and a dependent (**y**) variable, is build. During the analysis the **X** matrix is decomposed into scores (**t**) and weights (**w**) as can be seen in Figure 2.4. In Chapter 5, N-PLS will be used for the analysis of a 3D QSAR data set where the descriptor matrix $\underline{\mathbf{X}}$ is multiway. In analogy with the PLS method for the case when $\underline{\mathbf{X}}$ is of order three, N-PLS decomposes the matrix into a score vector (**t**) and two weight vectors ($\mathbf{w}^J$ and $\mathbf{w}^K$) as pictured in Figure 2.8 for the first component.

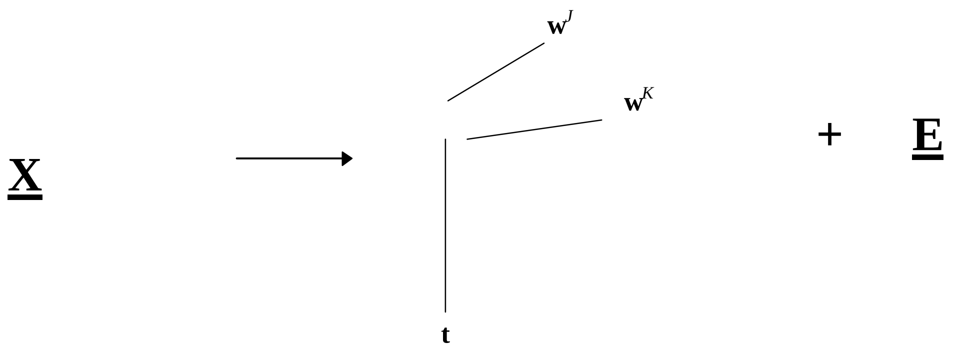


Figure 2.8 The one component decomposition of a multiway matrix, $\underline{\mathbf{X}}$ ($I \times J \times K$), into a score vector ($\mathbf{t}$) and two loading vectors ($\mathbf{w}^J$ and $\mathbf{w}^K$). $\underline{\mathbf{E}}$ ($I \times J \times K$) represents the residuals not considered in the model.

In more detail, if $\underline{\mathbf{X}}$ is a three-way matrix ($I \times J \times K$; Figure 2.8) and $\mathbf{y}$ is univariate ($I \times 1$) with typical elements x_{ijk} and y_i, respectively, $\underline{\mathbf{X}}$ is decomposed into one score vector $\mathbf{t}$ ($I \times 1$) and two weight vectors $\mathbf{w}^J$ ($J \times 1$) and $\mathbf{w}^K$ ($K \times 1$), *i.e.*, one vector per mode. It is assumed that $\underline{\mathbf{X}}$ and $\mathbf{y}$ are column mean-centered in all cases. The model of $\underline{\mathbf{X}}$ is given by:

$$x_{ijk} = t_i w_j^J w_k^K + e_{ijk} \tag{2.22}$$

The general idea is to find $\mathbf{t}$ such that the covariance between $\mathbf{t}$ and $\mathbf{y}$ is maximized under the constraint that $\mathbf{t}$ is the best least squares solution to Equation 2.22 ($\|\mathbf{w}^J\| = \|\mathbf{w}^K\| = 1$). Since

$$t_i = \sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk} w_j^J w_k^K \tag{2.23}$$

is the least squares solution to Equation 2.22, given the $\mathbf{w}$’s, the problem can be rewritten as:

$$\max_{\mathrm{w^J w^K}}\left(\mathrm{cov}(\mathbf{t},\mathbf{y})\right) \tag{2.24}$$

The covariance between $\mathbf{t}$ and $\mathbf{y}$ can be written as a summation over i, cov($\mathbf{t}$,$\mathbf{y}$) = $\sum_{i=1}^{I} t_i y_i$. Note that no correction for the degrees of freedom (I–1) is necessary, without loss of generality, since this number is constant for a given component. Now, the problem to solve is:

$$\max_{\mathrm{w^J w^K}}\left(\sum_{i=1}^{I} t_i y_i\right) \tag{2.25}$$

and by including t_i from Equation 2.23 the final problem can be rewritten as:

$$\max_{\mathrm{w^J w^K}}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_i x_{ijk} w_j^J w_k^K\right] \tag{2.26}$$

Note that the least squares property is valid for $\mathbf{w}^J$ and $\mathbf{w}^K$ that satisfy Equation 2.23, but not for general $\mathbf{w}^J$ and $\mathbf{w}^K$, hence Equation 2.22 is a *partial* least squares model.[11] The summation over i, in Equation 2.26, can be performed directly since $\underline{\mathbf{X}}$ and $\mathbf{y}$ already are known. This summation will yield a matrix $\mathbf{Z}$ of size ($J \times K$) with typical element $z_{jk} = \sum_{i=1}^{I} y_i x_{ijk}$. When $\mathbf{Z}$ is defined, Equation 2.26 simplifies to:

$$\max_{\mathrm{w^J w^K}}\left[\sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk} w_j^J w_k^K\right] \tag{2.27}$$

and written in matrix notation it is clear that $\mathbf{w}^J$ and $\mathbf{w}^K$ can be determined as the first singular vectors from a singular value decomposition (SVD) of **Z**:

$$\max_{\mathrm{w^J w^K}}\left[\left(\mathbf{w}^J\right)^{\mathrm{T}}\mathbf{Z}\mathbf{w}^K\right] \Rightarrow \left(\mathbf{w}^J,\mathbf{w}^K\right) = \mathrm{SVD}(\mathbf{Z}) \tag{2.28}$$

The parameters from the next component are calculated from the residuals after the first component, *i.e.*, $\mathbf{X}_2 = \mathbf{X}_1 - \mathbf{t}_1\mathbf{w}_1^{\mathrm{T}}$ (compare Martens non-orthogonalized PLS algorithm)[3] where $\mathbf{w}_1$ is $(\mathbf{w}_1^K \otimes \mathbf{w}_1^J)$ and $\mathbf{X}_1$ is the unfolded $\underline{\mathbf{X}}_1$. Accordingly, $\underline{\mathbf{X}}_2$ replaces $\underline{\mathbf{X}}_1$ in Equation 2.22 and the component scores and weights from the second component can be determined.

Since the scores from different components are not orthogonal the regression coefficients $\mathbf{b}_\mathbf{A}$, in Equation 2.29, have to be calculated taking all the component score vectors into account:

$$\mathbf{b}_{\mathrm{A}} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{2.29}$$

The score-matrix **T** has the dimension $I \times A$ where the ath column represents the ath score-vector.

In the case of 3D QSAR data, five different modes can often be defined: the object mode, the grid x direction, the grid y direction, the grid z direction and finally the probe mode (see Figure 2.9). Thus, a 3D QSAR data set in five modes, with one dependent variable requires a penta-linear PLS1 algorithm.

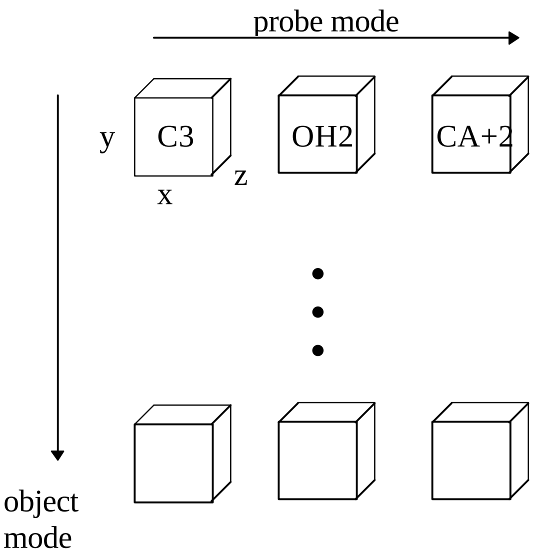


Figure 2.9 The complete data set defining five modes, *i.e.*, the object mode, x, y, z and the probe mode comprising 30, 31,15, 18, and 3 dimensions, respectively.

Analogous to the three-way problem, the five-way solution is obtained by finding the weight vectors $\mathbf{w}^J$, $\mathbf{w}^K$, $\mathbf{w}^L$ and $\mathbf{w}^M$. Since $\underline{\mathbf{X}}$ is of higher order than three, the solution can not be accomplished by a SVD, but similarly the weight vectors are now obtained by a one-component PARAFAC[12-14,16] (see previous section) decomposition of $\underline{\mathbf{Z}}$ with typical element $z_{jklm} = \sum_{i=1}^{I} y_i x_{ijklm}$:

$$\max_{\mathrm{w}^J\mathrm{w}^K\mathrm{w}^L\mathrm{w}^M}\left(\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M} z_{jklm} w_j^J w_k^K w_l^L w_m^M\right) \tag{2.30}$$

The multilinear PLS algorithm discussed above has been thoroughly scrutinized by Bro[19] and Smilde.[11]

2.9 Multiway Principal Covariate Regression (PCovR)

The theory of Principal Covariate Regression (PCovR)[11,20] is the last regression method discussed and will be applied to real 3D QSAR data in Chapter 7. PCovR can be regarded as a combination of a PCA on **X** and a simultaneous regression on **y**. In PCR, the components (**T**) are extracted component-wise with the objective to reconstruct **X** optimally and subsequently regress **y** on **T** using MLR (see Equation 2.9). In PLS,[8,9] a weight vector **w** and a score vector **t** are searched for such that the covariance between **t** and **y** is maximized under the constraint that **t** is the best least squares solution to the model $\hat{x}_{ij} = t_i w_j$. The PCovR method also provides a least squares solution, but in contrast to the previous methods, PCovR determines all components simultaneously. In PCovR, the data are fitted to the following model:

$$\mathbf{T} = \mathbf{XW} \tag{2.31}$$

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E}_{\mathrm{X}} \tag{2.32}$$

$$\mathbf{y} = \mathbf{Tb} + \mathbf{e}_{\mathrm{y}} \tag{2.33}$$

$$\min\left[a\left\|\mathbf{X} - \mathbf{XWP}^{\mathrm{T}}\right\|^2 + (1-a)\left\|\mathbf{y} - \mathbf{XWb}\right\|^2\right] \tag{2.34}$$

where **T** ($I \times A$) contains the A score vectors (principal covariates $\mathbf{t}_i$), **W** ($J \times A$) contains the component weights, **P** ($J \times A$) and **b** ($A \times 1$) are the regression parameters relating **X** ($I \times J$) and **y** ($I \times 1$), respectively, with the scores in **T** (see Figure 2.10) and $\mathbf{E}_{\mathrm{X}}$ and $\mathbf{e}_{\mathrm{y}}$ comprise the part of **X** and **y**, respectively, not accounted for by the model. The PCovR algorithm can be directed to reconstruct **X,** or fit **y,** by assigning α a value between one and zero, respectively. Obviously, when α = 1, **X** is reconstructed without fitting **y**, corresponding to a PCA of **X**. In the case of α = 0, the model resembles very much a MLR model since the emphasis totally is focused on fitting **y**. In the case where the number of components A is equal to the rank of **X**, the solution is equivalent to the full rank MLR solution (Section 2.2), independent of the value assigned to α.

In the applications used later in this thesis **y** is univariate, but the extension to several independent variables is possible and is described in the original PCovR article, by De Jong and Kiers.[20]

In PCovR, the data is fitted to the model in an iterative process using:

$$\alpha \mathrm{R}_{\mathrm{X}}^2 + (1-\alpha)\mathrm{R}_{\mathrm{y}}^2 \tag{2.35}$$

as the expression to maximize, where $\mathrm{R}_{\mathrm{X}}^2$ represents the variance in **X** accounted for by **T,** and $\mathrm{R}_{\mathrm{y}}^2$ represents the variance in **y** explained by **T**. It is convenient, however, to rewrite Equation 2.35 as the least squares loss function in Equation 2.34 and, subsequently, minimize the function with respect to **W**, **P** and **b** in the PCovR algorithm with constraint put on **P** (see below). The algorithm is converged when the relative change of Equation 2.34 from a subsequent iteration is less than a defined threshold value, the criterion of convergence, typically chosen very small (10^{-5}).

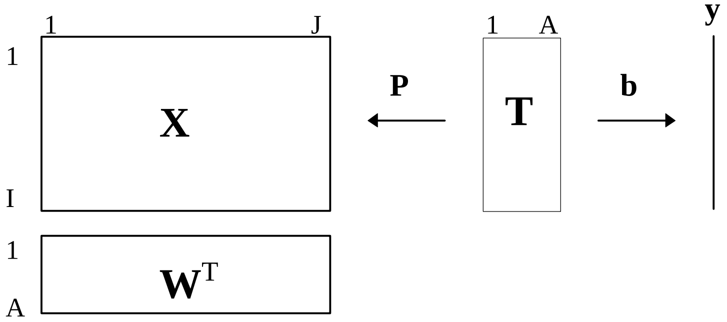


Figure 2.10 A graphical representation of PCovR, a simultaneous two-block regression method. Here, **X** ($I \times J$) is a two-way matrix and **y** ($I \times 1$) is univariate.

Smilde[11] generalized the PCovR algorithm (Equations 2.31–2.34) to be valid also for multiway data, where **X** ($I \times JK$) is the rearranged three-way matrix $\underline{\mathbf{X}}$ ($I \times J \times K$). Smilde postulated a number of different models by imposing different structures (constraints) on **P**. The simultaneous two-block Tucker3 model (PCovR/Tucker3) is obtained when

$$\mathbf{P}^{\mathrm{T}} = \mathbf{G}\left(\mathbf{W}_K^{\mathrm{T}} \otimes \mathbf{W}_J^{\mathrm{T}}\right) \tag{2.36}$$

where $\mathbf{W}_J$ ($J \times M$) and $\mathbf{W}_K$ ($K \times N$) are as defined in Figure 2.7, **G** ($L \times MN$) is the concatenated three-way core-matrix $\underline{\mathbf{G}}$ ($L \times M \times N$) where L, M and N are the number of dimensions used in the three modes. Similarly, the simultaneous two-block PARAFAC model (PCovR/PARAFAC) is obtained when

$$\mathbf{P}^{\mathrm{T}} = \left(\mathbf{W}_K^{\mathrm{T}} \otimes \mathbf{W}_J^{\mathrm{T}}\right) \tag{2.37}$$

where $\mathbf{W}_J$ ($J \times A$) and $\mathbf{W}_K$ ($K \times A$) are as defined in Figure 2.6, and A is the number of components used in all three modes.

In the PARAFAC model, the core matrix $\underline{\mathbf{G}}$ actually is a matrix where all off-superdiagonal elements are zero, since interactions between components from different modes are not allowed. In the Tucker model, however, this type of interactions are accounted for by the elements in $\underline{\mathbf{G}}$. Hence, the PARAFAC model is a special case of the Tucker model (*e.g.*, more constrained). Additionally, each rotation of the core matrix or of the loading matrices, from a Tucker model, may give solutions with the same sum of squared residuals as the model found. Hence, the Tucker model does not have a unique solution. The general PCovR model is presented graphically in Figure 2.10.

Since the PCovR algorithm is solved by an alternating least squares (ALS) algorithm, the initial starting parameters have to be selected *a priori*. Accordingly, the parameters, *e.g.*, loadings, scores, **P** and **b**, are updated alternately until convergence. It is a well known fact that in non-linear modeling an algorithm may occasionally converge into a local minimum, since the result depend on the starting parameters used. Unfortunately, there is no easy way to determine whether the minimum value found is a local or the global minimum function value. However, the chance of finding the global minimum value is increased if several calculations, with different starting parameters, are attempted.

In Chapter 7, the simultaneous two-block Tucker5[21] and five-way PARAFAC[22] methods have been utilized for the analysis of a 3D QSAR data set.

Predictions of new observations $\mathbf{X}_{\text{new}}$ are accomplished by first defining the regression model in terms of the original $\mathbf{X}$ variables:

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{\text{pred}} + \mathbf{e}_{\text{y}} \tag{2.38}$$

and, accordingly, define $\mathbf{b}_{\text{pred}}$ as

$$\mathbf{b}_{\text{pred}} = \mathbf{W}\mathbf{b} \tag{2.39}$$

where $\mathbf{b}$ are the regression coefficients (compare Equation 2.19) and $\mathbf{W} = \left(\mathbf{W}_K \otimes \mathbf{W}_J\right)$. Then, Equation 2.40 can be utilized to predict the responses $\hat{\mathbf{y}}_{\text{new}}$ of the objects in $\mathbf{X}_{\text{new}}$.

$$\hat{\mathbf{y}}_{\text{new}} = \mathbf{X}_{\text{new}}\mathbf{W}\mathbf{b} \tag{2.40}$$

Still to be solved, is a framework for the proper assignment of α. De Jong[20] and Smilde[11] suggested to use leave-one-out crossvalidation and find an α that maximizes the predictability. In Chapter 7, a different approach was used since crossvalidation was considered a too time consuming procedure.

2.10 Model Validation

The quality of a QSAR model is mostly determined by its ability to perform predictions of objects not included in the training set. However, it is often difficult to assemble enough compounds for sufficient large training and test sets. Therefore, the predictability is usually estimated with crossvalidation, *i.e.*, internal validation, where a subset of the training set is omitted during calibration and, subsequently, predicted with the obtained model. Obviously, the smaller the number of subsets used the harder the validation criterion becomes. This procedure is repeated until all compounds have been omitted once. Accordingly, the predictability is quantified as the crossvalidated Q^2 in Equation 2.41. PRESS (Equation 2.42) is the prediction error sum of squares and SSY (Equation 2.43) is the sum of squares of the response variable where $\hat{y}_{(i)}$ is the prediction of y_i using a model with the *i*th object omitted. A model with high predictability has a crossvalidated Q^2 close to unity, while a model with low or no predictability has a Q^2 close to or below zero.

$$\mathrm{Q}^2 = 1 - \mathrm{PRESS}/\mathrm{SSY} \tag{2.41}$$

$$\mathrm{PRESS} = \sum_{i=1}^{I}\left(y_i - \hat{y}_{(i)}\right)^2 \tag{2.42}$$

$$\mathrm{SSY} = \sum_{i=1}^{I}\left(y_i - \bar{y}\right)^2 \tag{2.43}$$

Thus, by definition, a model possesses predictability, only when the variation of the prediction errors is less than the variation of the response variable. The optimal number of components to include in the final model is generally[18] chosen equal to the number of components (lv) maximizing the function $Q^2 = f(\text{lv})$, or minimizing the function $\text{PRESS} = f(\text{lv})$. Actually, the crossvalidated Q^2 is an estimation of the predicted Q^2, *i.e.*, a models ability to predict the responses of an external test set.

The predicted Q^2 is expressed as in Equation 2.41 but now the SSY corresponds to the sum of squares of the responses from the test set and $\hat{y}_{(i)}$ is the prediction of the *i*th test compound. Clementi[23] advises not to use an external test set but rather include all available compounds in the training set. Statistically, real predictability can only be validated with an external test set, which is not included in the calibration process of the model. However, it is often hard to find enough compounds for both a training and a test set. Therefore, a compromise between an external test set and leave-one-out crossvalidation may be crossvalidation where more compounds are left out each time, and repeated several times,[24] like in bootstrapping.[25,26] However, if enough compounds are available, an external test set still is the best validation method and is, thus, recommended.

In PLS1 (PLS with one dependent variable), components scores **T** are extracted from the independent descriptor block (**X**) such that the covariance between **T** and the response variable (**y**) is maximized. By definition, the fraction of explained variance in **X** and **y** increases as the number of components in a model increases. The MLR method fits the data perfectly and due to the large number of parameters that are estimated the model becomes rigid and the predictability is often low. In PCR and PLS modeling the fit increases with each component and, if $J \leq I$, the solution converges towards the MLR solution, when all components are extracted. However, with crossvalidation the number of components necessary to account for the variation in **X** that significantly describe **y** can be estimated. Consequently, an overfitted model with too many components possesses a high degree of fit with low predictability (compare MLR). In contrast, an underfitted model with too few components does not account for sufficient variation in **X**. The degree of fit is expressed with the R^2 as defined in Equation 2.44, *i.e.*, the fraction of variation in **y** accounted for by **X**. SSY is the same as in Equation 2.43 and $\hat{y}_i$ is the model estimation of y_i.

$$R^2 = 1 - \sum_{i=1}^{I} \left(y_i - \hat{y}_i \right)^2 \Big/ \text{SSY} \tag{2.44}$$

Variations to both Q^2 and R^2 are suggested in the literature[24,27] but throughout this thesis crossvalidated Q^2, predicted Q^2 and R^2 are used for the presentation of crossvalidations, external predictions and model calibrations, respectively.

2.11 Variable Selection

A typical data set in 3D QSAR comprises thousands of columns, where a lot of them consist of insignificant variation, in particular grid points at large distances from the ligands. It has been shown[27] that, if irrelevant variables are maintained in the data set they give rise to worse predictions, since they express only random variation. Additionally, Clark *et al.*[28] reported that PLS might overlook ‘true’ correlations when too much redundant variables are present. Accordingly, one might suggest detecting this kind of variables and omit them from the analysis. However, there is a risk when a subset of variables is selected from a larger number; they might correlate with **y** purely by chance, which has been discussed in the literature by Topliss *et al.*[29] and Clark *et al.*[28] The latter authors found that the risk of chance correlation in CoMFA applications when the crossvalidated $Q^2 > 0.25$ to be, more or less, negligible. A frequently used method for the detection of chance correlation is permuting the elements in the **y**-vector, *i.e.*, a model calculated with the **y**-vector permuted must possess no predictive ability.[17] If a model is obtained not pure by chance then the average predictive Q^2, from several different models calculated with permuted **y**, must be significantly lower than the Q^2 from the model obtained with the original **y**-vector.

Several methods have been suggested for variable selection in 3D QSAR including GOLPE (Generating Optimal Linear PLS Estimations),[17] Cho region selection,[30] Norinder region selection,[31] and IVS-PLS (Interactive Variable Selection for PLS).[32,33]

In a previous version of GOLPE, redundant variables were preselected by means of a D-optimal selection in the PLS weights. The influence of the remaining variables on the predictability were estimated in a series of crossvalidation experiments where variables were included and excluded, alternately, following a Fractorial Factorial Design (FFD).[7] Only variables with significant positive influence on the predictability were selected. In the latest version of GOLPE, the D-optimal preselection procedure has been replaced by a smart region definition (SRD)[34] procedure. SRD aims at extracting groups of variables (regions) in 3D space carrying the same information. For a more elaborate description of GOLPE variable selection, see Chapter 4.

The region selection methods presented by Cho *et al.*[30] and Norinder[31] are very much related to GOLPE in the sense that they define and estimate the influence of regions in the grid on the predictability, as GOLPE does with single variables. In the Cho region selection method, the influence of whole regions on the predictability are estimated by performing crossvalidation experiments, with each region apart. Only variables within regions having a crossvalidated Q^2 above a defined limit are selected.

It is important to note, Norinder and Cho *et al.* define rectangular regions only on geometrical grounds while the regions defined with GOLPE/SRD are different in size and shape depending on the spatial location and the statistical significance of grid points in the close vicinity.

Norinder does the same with regions as GOLPE does with variables, hence, performing crossvalidation experiment where regions are left out, alternately, following a FFD protocol. As in GOLPE, variables within regions with significant positive influence on the predictability are selected.

Interestingly, crossvalidation experiments with both region selection methods were excellent with high crossvalidated Q^2s, but predictions of external test sets were not good. It indicate problems with overfitting. The approach presented by Cho *et al.*[30] has one disadvantage, it leaves the fundamental idea behind multivariate analysis and divides the grid into a number of separate sub-models. With Norinder's method the analysis is performed following a FFD protocol which, statistically, is more correct.

The last variable selection method IVS-PLS[32,33] does not really reject variables but instead reweights single elements in the PLS weight (**w**) vector, dimension-wise. After each PLS component, a lower cut-off level for the weights is searched for by performing crossvalidation experiments and, subsequently, increase the cut-off level. Weights below the cut-off level is set to zero prior to each experiment, and the cut-off level that minimizes the CV-value (PRESS/SSY) is adapted. An upper cut-off level is also determined in a similar manner, but now the crossvalidation experiments are started with a high cut-off level which, subsequently, is decreased. Weights exceeding the upper cut-off level are set to zero and, again, the upper cut-off level is determined as the level minimizing the CV value. One may object against setting high variable weights to zero, but Lindgren *et al.*[32] showed that large elements in **w** sometimes suppress smaller values which may affect the predictability negatively.

2.12 References

1. Myers, R.H. Classical and Modern Regression with Applications. PWS-KENT Publishing Company: **1997**
2. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185,* 1-17.
3. Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons: New York, **1989**
4. GOLPE, *3.0,* Clementi, S. Multivariate Infometric Analyses(MIA), Perugia, Italy, SGI.
5. Smilde, A.K. PhD Thesis. Multivariate Calibration of Reversed-Phase Chromatography Systems Research Group of Chemometrics, University of Groningen, The Netherlands. **1990**
6. Belsley, D.A.; Kuh, E.; Welsch, R.E. Regression Diagnostics. Wiley: New York, **1980**
7. Morgan, E. Chemometrics: Experimental Design. Chadwick, N. Ed.; John Wiley and Sons Ltd: Chichester, **1991**
8. Wold, H. Systems Under Indirect Observation. The Basic Design and Some Extensions. Jöreskog, K. and Wold, H. Eds. North-Holland.: Amsterdam, **1982**
9. Wold, S.; Ruhe, A.; Wold, H.; Dunn III, W.J. *J. Sci. Stat. Comput.* **1984**, *5,* 735-743.
10. Wold, H. Perspectives in Probability and Statistics. Soft Modeling by Latent Variable, the Non-Linear Iterative Partial Least Squares (NIPALS) Algorithm. Gani, J. Ed. Academic Press.: London, **1975**; pp. 117-142.
11. Smilde, A.K. Comments on Multilinear PLS. *J. of Chemometrics* **1997**, *11,* 367-377.
12. Smilde, A.K. and Doornbos, D.A. Three-way Methods for the Calibration of Chromatographic Systems: Comparing PARAFAC and Three-way PLS. *J. of Chemometrics* **1991**, *5,* 345-360.
13. Carroll, J. and Chang, J.J. Analysis of Individual Differences in Multidimensional Scaling with an N-way Generalization of the Eckart-Young Decomposition. *Psycometrika* **1970**, *35,* 283-319.
14. Harshman, R.A. Foundations of the PARAFAC Procedure: Models and Conditions for an Exploratory Multimodal Factor Analysis. *UCLA Working Papers in Phonetics* **1970**, *16,* 1-84.
15. Tucker, L. Problems of Measuring Change. Implications of Factor Analysis of Three-Way Matrices for Measurement of Change. Harris, C. Ed. University of Wisconsin Press.: Madison, **1963**; pp. 122-137.
16. Smilde, A.K. Three-way Analyses. Problems and Prospects. *Chemom. and Intell. Lab. Syst.* **1992**, *15,* 143-157.
17. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct. -Act. Relat.* **1993**, *12,* 9-20.
18. SYBYL- Molecular Modeling Software, *6.3,* Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA,
19. Bro, R. Multiway Calibration. Multilinear PLS. *J. of Chemometrics* **1996**, *10,* 47-61.
20. De Jong, S. and Kiers, H.A.L. Principal covariates regression Part I: Theory. *Chemom. and Intell. Lab. Syst.* **1992**, *14,* 155-164.
21. Kiers, H.A.L. Principal Covariates Regression with 5-Way Tucker Model for X, Matlab-Code, University of Groningen, **1997**.
22. Kiers, H.A.L. Principal Covariates Regression with 5-Way PARAFAC Model for X, Matlab-Code, University of Groningen, **1997**.
23. Clementi, S. *Personal Communication,* **1995**
24. Cruciani, G.; Baroni, M.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part 1: Standard Deviation of Prediction Errors (SDEP). *J. of Chemometrics* **1992**, *6,* 335-346.
25. Wold, S. Validation of QSAR's. *Quant. Struct. -Act. Relat.* **1991**, *10,* 191-193.
26. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* **1996**, *7,* 1-26.

27. Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model. *J. of Chemometrics* **1992**, *6,* 347-356.
28. Clark, M. and Cramer III, R.D. The Probability of Chance Correlation Using Partial Least Squares (PLSR). *Quant. Struct. -Act. Relat.* **1993**, *12,* 137-145.
29. Topliss, J.G. and Edwards, R.P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22,* 1238-1244.
30. Cho, S.J. and Tropscha, A. Crossvalidated R2-guided Region Selection for Comparative Molecular field Analyses: A Simple Method to Achieve Consistent Results. *J. Med. Chem* **1995**, *38,* 1060-1066.
31. Norinder, U. Single Domain Mode Variable Selection in 3D QSAR Applications. *J. of Chemometrics* **1996**, *10,* 95-105.
32. Lindgren, F.; Geladi, P.; Rännar, S.; Wold, S. Interactive Variable Selection (IVS) for PLS. Part 1: Theory and Algorithms. *J. of Chemometrics* **1994**, *8,* 349-363.
33. Lindgren, F.; Geladi, P.; Berglund, A.; Sjöström, M.; Wold, S. Interactive Variable Selection (IVS) for PLS. Part 2: Chemical Applications. *J. of Chemometrics* **1995**, *9,* 331-342.
34. Pastor, M.; Cruciani, G.; Clementi, S. Smart Region Definition: A New Way to Improve the Predictive Ability and Interpretability of Three-Dimensional Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1997**, *40,* 1455-1464.

Design, Syntheses and QSAR of a Series *trans*-1,2,3,4,4a,5,6,10b-Octahydrobenzo[*f*]quinolines with Dopaminergic Affinity

3

This chapter is based on the article: Nilsson, J.; Selditz, U; Pugsley, T.; Sundell, S.; Lundmark, M.; Smilde, A. K.; Wikström, H. Design, Syntheses and QSAR of a Series *trans*-1,2,3,4,4a,5,6,10b-Octahydrobenzo[*f*]quinolines with Dopaminergic Affinity. *submitted*

Summary

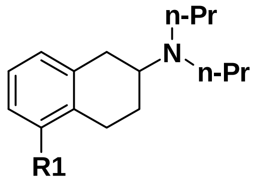
In order to investigate the influence of different substituents at the 4-N and 7-O positions of trans-1,2,3,4,4a,5,6,10b-octahydrobenzo[f]quinolines (trans-OHB[f]Qs) on the affinity for the dopamine D_{2L}*,* D_3 *and* $D_{4.2}$ *receptors, a large number of compounds were generated theoretically, and characterized with physicochemical descriptors. Accordingly, a subset of compounds was selected by means of a factorial design in the descriptor space, subsequently synthesized and screened for dopamine* D_{2L}*,* D_3 *and* $D_{4.2}$ *receptor affinities.*

In general, compounds with a hydroxy group at the seven position displayed significant high affinities for all three dopamine receptors while the compounds with a sulfon ester group were less potent. In addition, the sulfon ester group suppresses the affinity for the D_4 *receptor. The nitrogen substituent may be as large as a phenylethyl group without detrimentally affecting the affinity for the dopamine receptors. Finally, a compound with a 7-OH group and an N-propargyl group lacks affinity for the dopamine* D_4 *receptor. The somewhat rigid N-propargyl group and the low* pK_a *value (6.1) may be contributing factors to the low* D_4 *affinity.*

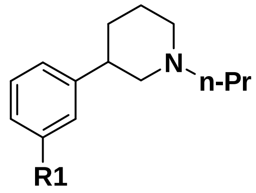
In analogy with the 2-aminotetralins, where the affinity for the dopamine receptors resides in the (2S)-enantiomers, the potency of the trans-OHB[f]Qs resides in the corresponding trans-(4aS,10bS) enantiomer.

3.1 Introduction

The OHB[*f*]Qs[1-7] (*trans*-**17** and *trans*-**1**) as rigid analogues of 2-aminotetralins[8] (**24**) and 3-PPP[5,6,9] (**22)** have, during the last two decades, achieved a lot of attention in the literature (see also Chapter 1). Attempts have been made to explain the structure-activity relationships[2,10] between this class of compounds[11,12] and other dopaminergic compounds like aporphines and ergolines.[8] For the OHB[*f*]Q system it was found that the potency resided in the *trans*-(4a*S*,10b*S*) enantiomer.[8] As was pointed out,[2,5] the *trans* isomer is rigid, assuming a flat molecular conformation, whereas the *cis* isomer is more flexible, with the piperidine-ring moiety protruding out of the plane, which may cause steric hindrance in the ligand-receptor interaction, and explain the higher affinity of the *trans* isomer.

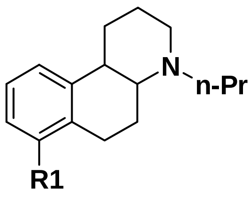


24 R1 = OH



22 R1 = OH

23 R1 = OSO_2CF_3



trans-**17** R1 = OH

trans-**1** R1 = OSO_2CF_3

Compound *trans*-**17** stimulated central presynaptic DA receptors at a low dose with no significant behavioral stimulation, reported by Wikström *et al.*[5] However, when administrated at higher doses, *trans*-**17** elicited typical and postsynaptical DA receptor stimulatory effects, *i.e.*, stereotypies, and increased locomotor activity.

Sonesson *et al.*[6] found *trans*-**1** to be inactive as an agonist even at high doses (50 μmol/kg). Instead, the striatal DOPAC levels in nonpretreated habituated rats were increased by 235 %, suggesting presynaptic DA receptor antagonistic properties. In the same assay, *trans*-**1** decreased significantly the locomotor activity to 56 %.

As previously reported by Wikström *et al.*,[3] when the *N*-*n*-propyl group was replaced by an *N*-*n*-butyl group or longer chains, these analogues of compounds **22** and *trans*-**17** became more active in the biochemical and the behavior models used by the authors.

In the present chapter a large number of OHB[*f*]Qs were generated, using compound **1** as the template, and subsequently characterized with physicochemical parameters. A few of the most diverse compounds were then selected and synthesized. After testing the compounds for *in vitro* affinity at the dopamine D_{2L}, D_3 and $D_{4.2}$ receptor subtypes, these data plus the physicochemical descriptors will provide information to the structure-activity relationships.

3.2 Computational Chemistry

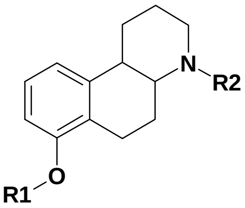
: Tentative Compounds

A large number of 4-*N* and 7-*O* substituted *trans*-OHB[*f*]Qs were designed by permuting all the combinations of the substituents listed in Table 3.1, at the R_1 and R_2 positions. On position R_1 and R_2, eight (ID R_1 = a, b, c, d, e, f, g, h) and eleven (ID R_2 = 1 to 11) different substituents were considered, respectively. Each compound can be identified by combining the IDs in Table 3.1. For example, compound h03 is a *trans*-OHB[*f*]Q with a triflate group on the 7 position (ID R_1 is h) and an ethyl group on the 4-*N* position (ID R_2 is 03).

: Physicochemical descriptors

Physicochemical descriptors were generated for all compounds listed in Table 3.1. The descriptors were obtained from different sources: Mopac AM1 single point calculations[13,14]; pKa and logP values were calculated using the Pallas 1.2 program[15] and, finally, descriptors were obtained from the literature.[16] The 19 descriptors in Table 3.2 were generated for all 88 compounds, and collected in the descriptor matrix **X** (88×19).

Table 3.1 The substituents permuted in order to generate the initial 88 compounds.



ID R_1	R_1	ID R_2	R_2
a	-H	01	$-CH_2CH_2CH_2CH_3$
b	$-CH_3$	02	$-CH_2CH_2CH_3$
c	$-CH_2CH_3$	03	$-CH_2CH_3$
d	$-SO_2-CH_3$	04	$-CH_3$
e	$-SO_2-C_6H_5$	05	-H
f	$-SO_2-C_6H_4-CH_3$	06	$-CH_2CH_2C_6H_5$
g	$-SO_2$-2-thiophene	07	$-CH_2CH_2$-2-thiophene
h	$-SO_2-CF_3$	08	$-CH_2CCH$
		09	$-CH_2CHCH_2$
		10	$-CH(CH_3)_2$
		11	$-CH_2C_6H_5$

: Principal Properties

In Quantitative Structure-Activity Relationships (QSAR), Principal Properties (PPs) are frequently used for the description of series of compounds (see Chapter 1).[17-19] A PP is a score vector (**t**) obtained from a Principal Component Analysis[20,21] of **X** ($I \times J$), containing the J physicochemical parameters (columns) characterizing the I compounds (rows). The PPs are linear combinations of the descriptors in **X** and all PPs are chosen orthogonal, *i.e.*, each PP does not correlate with any of the other PPs. Optimally, each PP represents clearly interpretable features in the molecules, like steric (*e.g.*, size) or electrostatic (*e.g.*, charge) properties. In PCA, each subsequently extracted PP accounts for less variation in **X** and, consequently, the variation accounted for by the first PP is more significant for the description of **X**, as compared with the following PPs.

Prior to PCA, in this investigation, each column was mean-centered and scaled to have unit standard deviation, often referred to as auto-scaling[20] (see Chapter 2).

Table 3.2 The physicochemical descriptors used for the characterization of the compounds in Table 3.1.

	descriptor	short	source
1	heat of formation	hofo	a
2	electronic energy	elec	a
3	Core-Core repulsion	coco	a
4	dipole-moment	dipo	a
5	ionization-potential	iopo	a
6	homo	homo	a
7	lumo	lumo	a
8	π R_2	piNs	b
9	π R_1	piOs	b
10	electrostatic potential (-1)	pom1	c
11	electrostatic potential (0)	pot0	c
12	electrostatic potential (1)	pop1	c
13	point charge on N	chaN	a
14	point charge on O	chaO	a
15	charge on phenyl C in C-O	chCO	a
16	molecular weight	mowe	a
17	van der Waals volume	vdWv	c
18	pKa_1	pKa1	d
19	logP	logP	d

[a] from Mopac AM1 single point calculations[13,14]; [b] tabulated in literature[16]; [c] calculated in SYBYL 6.1[14]; [d] calculated in Pallas 1.2[15]

Experimental Design

It is assumed that the descriptors in Table 3.2, *i.e.*, combinations of several descriptors, contain the specific information necessary to distinguish between the ligand-receptor interactions for the different dopamine receptor subtypes. An experimental design in the descriptor space will make it possible to select a few of the most diverse compounds. If the choice stands to select between two compounds, the compound more easily synthesized is selected. Eventually, a compound that portrays selectivity for a receptor subtype, will provide information about which descriptors that are responsible for the selectivity. In order to simplify the interpretation, the selection is performed following a factorial design[22] protocol.

An initial PCA of **X** (88 × 19), with two components accounting for 66 % of the variation, clearly divided the 88 compounds in two clusters (Figure 3.1(a)). One cluster contains all compounds with R_1 being a H atom, an OMe group and an OEt group corresponding to the compounds in Table 3.1 with ID R_1 being a, b and c, respectively. The remaining compounds, *i.e.*, 'the sulfon esters', form the second cluster. From Figure 3.1(a) it is also clear, the compounds with a triflate group at the 7 position (ID R_1 = h) form a subcluster within the sulfon ester group. Accordingly, compounds were selected from each cluster separately.

First, since compound **1** (*N*-*n*-propyl-7-OTf-OHB[*f*]Q; h02 in Table 3.1) displays affinity (*i.e.*, D_2 = 200 and D_3 = 21 nM) for dopamine receptors only one additional triflate compound was synthesized. Hence, compound h01 (**16**) was selected arbitrarily.

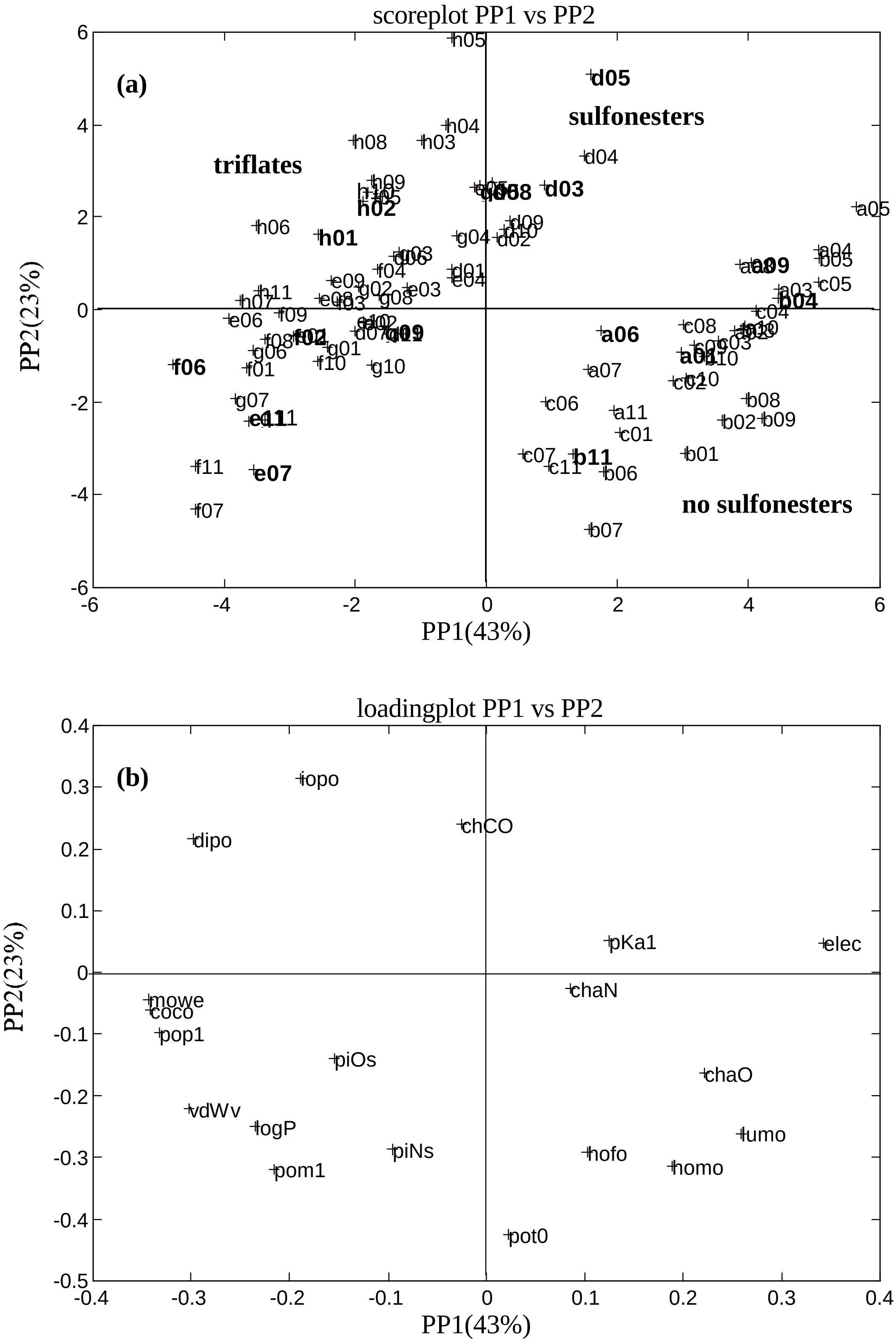


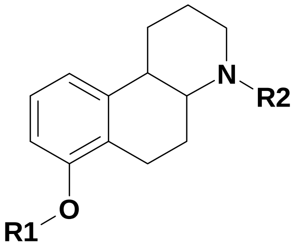
Figure 3.1 Score (a) and Loading (b) plots from the PCA of all the computer generated compounds (88 × 19). The selected compounds are high-lighted in (a) with bold face characters.

Second, from the sulfon ester cluster, including compounds with ID R_1 being d, e, f and g (thus, triflates not included), nine compounds were selected. The selection was performed by means of a factorial design[22] in the three first PPs, accounting for 77 % of the variation in the descriptor matrix (44×19).

Third, from the cluster without sulfon esters, including compounds with ID R_1 being a, b and c, five compounds were selected. Again, the selection was performed by means of a factorial design in the first two PPs, accounting for 64 % of the variation in the descriptor matrix (33×19).

All the selected compounds are high-lighted with bold face characters in Figure 3.1(a) and listed in Table 3.3. The decision to synthesize compound **20** (Table 3.3) was taken at a later stage in the investigation (see below).

Table 3.3 The *in vitro* receptor binding results from the synthesized compounds. All binding results are reported as K_i (nM) values.

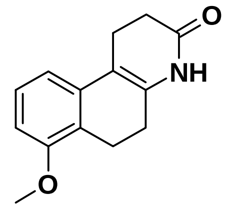


Compd	R1	R2	D_2	D_3	$D_{4.2}$
1	$-SO_2CF_3$	$-CH_2CH_2CH_3$	200	21	ND
1-(–)	$-SO_2CF_3$	$-CH_2CH_2CH_3$	56	19	>3300
1-(+)	$-SO_2CF_3$	$-CH_2CH_2CH_3$	240	120	>1000
2	-H	$-CH_2CHCH_2$	61	6.0	26
3[a]	-H	$-CH_2CH_2C_6H_5$	10	3.0	30
4	$-CH_3$	$-CH_3$	>5900	800	540
5	$-CH_3$	$-CH_2C_6H_5$	>5900	820	2000
6[b]	-H	$-CH_2CH_2CH_2CH_3$	29	4.0	50
7	$-SO_2CH_3$	$-CH_2CCH$	5100	1800	>3300
8	$-SO_2C_6H_5$	$-CH_2C_6H_5$	1600	720	3300
9	$-SO_2CH_3$	$-CH_2CH_3$	800	70	330
10	$-SO_2C_6H_5$	$-CH_2CH_2$-2-thiophene	350	140	>3300
11	$-SO_2CH_3$	-H	5700	380	>3300
12	$-SO_2C_6H_4CH_3$	$-CH_2CH_2C_6H_5$	>5900	>3000	>3300
13[c]	$-SO_2$-2-thiophene	-H	ND	ND	ND
14	$-SO_2C_6H_4CH_3$	$-CH_2CH_2CH_3$	220	37	>3300
15	$-SO_2$-2-thiophene	$-CH_2CHCH_2$	190	47	3300
16	$-SO_2CF_3$	$-CH_2CH_2CH_2CH_3$	180	29	>3300
20	-H	$-CH_2CCH$	15	13	730

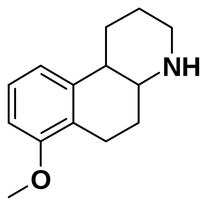
[a] Reference 23; [b] Reference 5; [c] Not synthesized

3.3 Chemistry

The syntheses of *cis*- and *trans*-OHB[*f*]Q have been discussed in a number of publications[2,4,5,8,12,24] during the last two decades. However, some initial synthetic points need to be mentioned also here.



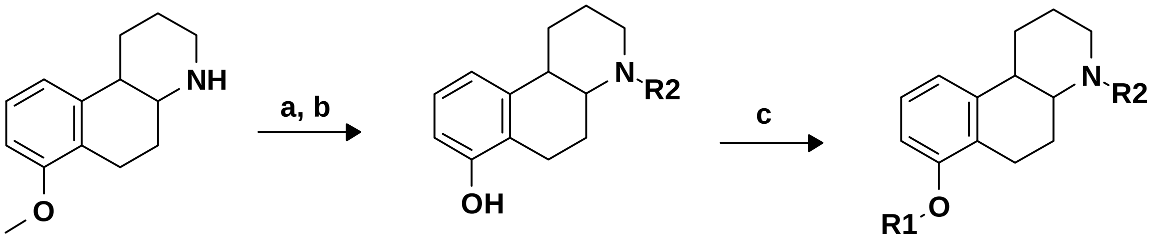
25



trans-**26**

The OHB[*f*]Q skeleton may be obtained via the intermediate compound **25** by reduction of the enamide in two steps to obtain **26**. Alternatively, **25** is alkylated at the 4-*N* position before the two reduction steps are carried out. Wikström *et al.*[5] chose for the latter route, hence, they benzylated **25** followed by reduction of the enamide with Pd/C under H_2 yielding a 1:1 mixture of *cis* and *trans* lactam. After reduction of the lactam with $LiAlH_4$ the resulting *cis* and *trans* isomers were separated by means of chromatography on a silica column. Cannon *et al.*[24] preferred the former route, since they managed to selectively reduce **25** with triethylsilane in trifluoroacetic acid, and obtain pure *trans* isomer as the product. The *trans*-**26** was obtained by a subsequent reduction with $LiAlH_4$.

The majority of the compounds in Table 3.3 were synthesized in three consecutive reaction steps (Scheme 3.1), starting from *trans*-**26**: a) alkylation of the 4-nitrogen, b) demethylation of the 7-OMe followed by c) sulfon ester formation at the 7-*O* position.



trans-**26**

Scheme 1 (a) alkylation; (b) demethylation; (c) sulfon ester formation of compound *trans*-**26**

The alkylation was performed with the alkylating reagent, in refluxing acetonitrile, using Cs_2CO_3 as the base. Since the intermediates were stable in acids, the demethylation was performed by refluxing the alkylated *trans*-**26** in 48 % HBr solution. Occasionally, the demethylation was performed using 1 M BBr_3 solution in dichloromethane cooled at -60 ºC. The sulfon esters were synthesized using either of two methods: in a two phase system with dichloromethane and 8 % NaOH with a proper catalyst or in dry dichloromethane, using triethylamine as the base.

For the preparation of *trans*-**5** used for the separation of the enantiomers (see below), the route described by Wikström *et al*.[5] was used.

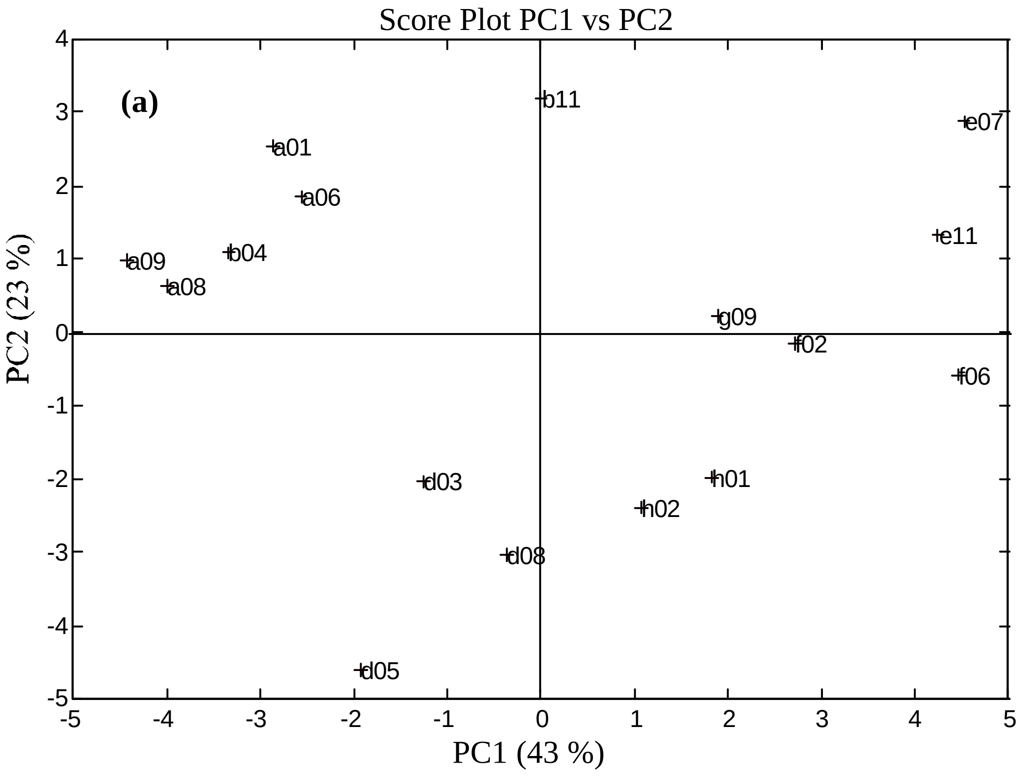
3.4 *In Vitro* Pharmacology

All the compounds synthesized were tested in three different *in vitro* receptor binding assays (Table 3.3), *i.e.*, dopamine D_{2L}, D_3 and $D_{4.2}$, using $[^3H]$-spiperone as the radioligand, performed as described in the Experimental Section.

3.5 Results and Discussion

Quantitative Structure-Activity Relationships (QSAR)

From a PCA of the matrix containing both the descriptors and the response variables, the first two PCs account for 66 % of the variation (Figure 3.2). The resemblance between the loading plots in Figure 3.1(b) and Figure 3.2(b), confirm that the selection of molecules were carried out properly. The response variables are highly correlated and cluster close to the center of the loading plot (Figure 3.2(b)), indicating a relative low association with the other descriptors. The response variable, pD_4, is placed a bit further to the left from the center of the plot, as compared with pD_2 and pD_3, indicating that pD_4 is more inversely correlated with the size related descriptors, *i.e.*, vdWv, pop1, coco and mowe. Consequently, the cluster of compounds in the upper left quadrant in Figure 3.2(a), do have affinity for the dopamine D_4 receptor, while the larger compounds to the right in the same plot, totally lack affinity for the dopamine D_4 receptor. The compounds with affinities for the dopamine D_4 receptor do not have a sulfon ester group at the 7 position.



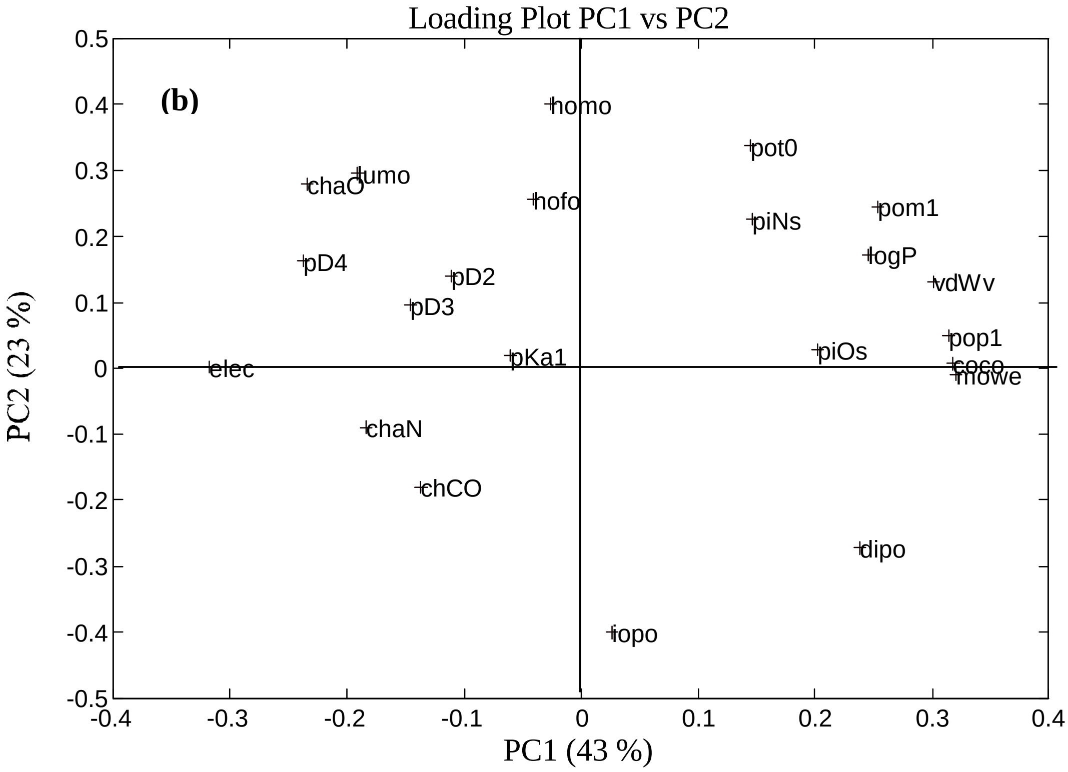


Figure 3.2 Score (a) and Loading (b) plots from a PCA of the descriptors from the synthesized and tested compounds in Table 3.3. In (b), pD2, pD3 and pD4 represent the –log(K_i(nM)) values of the receptor affinities from the respective receptors.

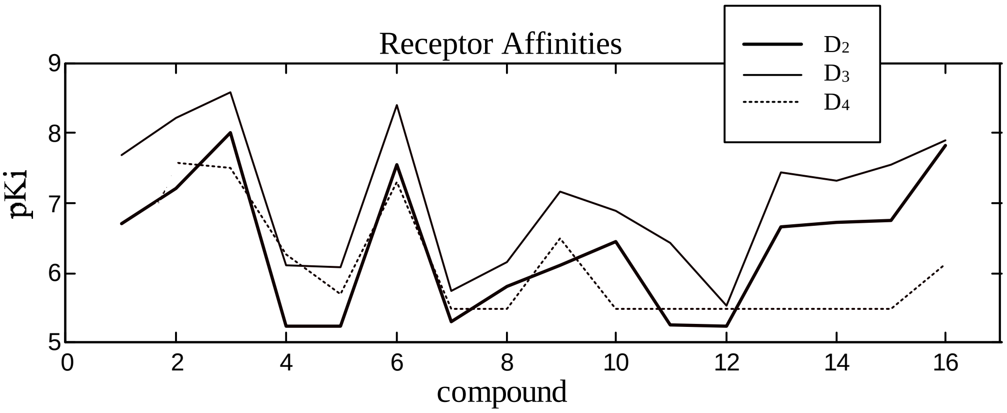


Figure 3.3 The affinities for the three dopamine receptors from the selected compounds (Table 3.3). Numbers 1–12 correspond to the compound numbers in Table 3.3; numbers 13–15 correspond to compounds **14**–**16** and number 16 corresponds to compound **20**.

The affinities for the three receptors are highly correlated and selectivity for any of the three receptors is not observed for any of the compounds (Figure 3.3). However, the compounds with significant high affinity for the dopamine receptors, *e.g.*, compounds **2**, **3** and **6**, all have a hydroxy group at the 7 position. Therefore, one additional hydroxy compound was synthesized in order to investigate the influence of different 4-*N* substituents on the receptor affinity. Accordingly, compound **20** (a08 in Table 3.1) was synthesized, and compared with all the other hydroxy compounds in Table 3.4. The pKa and logP were found to possess some association with the receptor affinity in less significant PCs than displayed in Figure 3.2, and were for reason of comparison appended to Table 3.4.

Table 3.4 Five OHB[*f*]Qs with a hydroxy group at the seven position and different substituents at the 4-*N* position. Binding affinities are K_i values as reported in Table 3.3.

Compd	*N*-substituent	D_2(nM)	D_3(nM)	D_4 (nM)	pKa[a]	logP[a]
6	$-CH_2CH_2CH_2CH_3$	29	4.0	49	9.8	4.1
3	$-CH_2CH_2C_6H_6$	10	3.0	30	8.4	4.7
2	$-CH_2CHCH_2$	61	6.0	26	9.7	3.2
20	$-CH_2CCH$	15	13	730	6.1	2.7

[a] *calculated with Pallas 1.2*[15]

Structure-Activity Relationships

From Table 3.4 it can be concluded that the *N* substituent may be as large as a phenylethyl group (*e.g.*, compound **3**) without detrimentally affecting the receptor affinities, confirming the conclusions drawn by Wikström *et al.*,[3] from *in vivo* biochemical experiments. This is also in agreement with the extended McDermed[25] model (see Chapter 1) as presented by Liljefors *et al.*[11,26]

The *N*-propargyl compound (**20**) has low affinity for the dopamine D_4 receptor (K_i = 730 nM), which is difficult to comprehend since the *N*-allyl compound (**2**) is much more potent (K_i = 26 nM). The most likely interaction point for the protonated nitrogen atom with the target receptor, is an aspartic acid residue on helix three (*i.e.*, Asp_{114}; Table 1.3),[27,28] which is preserved in all the D_2-

like dopamine receptors. Interaction with the aspartic acid residue requires a protonated nitrogen atom since the aspartic acid residue ($pK_a = 4.4$) is readily ionized at physiological pH. The pK_a for **20** and **2** were calculated to be 6.1 and 9.7, respectively, suggesting that **20** is less protonated than **2** at pH 7.4, *i.e.*, the pH used for the *in vitro* receptor binding experiments.

Dijkstra *et al.*[29] rationalized the low dopamine D_2 receptor affinity (competition with [^3H]-N0437) of compound PD128907 with the low measured pK_a value (6.1), indicating that only two percent of the compound is protonated at the nitrogen atom at pH 7.4. Today, PD128907 is one of the most selective D_3 agonist known. Thus, the significance of the pK_a value for the explanation of the difference between compounds **2** and **20** is not clear, since both compounds have affinity for the D_2 and D_3 receptors.

An alternative explanation to the lower D_4 affinity of **20** may be the *N*-propargyl group, which is less flexible as compared to the *N*-allyl, *N*-phenylethyl and *N*-butyl groups of compounds **2**, **3** and **6** (Table 3.4), respectively.

It is known from literature[30,31] that a triflate group increases the lipophilicity and has significant electrostatic influence. One effect induced by the triflate group is increased oral bioavailability,[31,32] as compared with hydroxy or methoxy substituents. An explanation for this, as suggested by Sonesson *et al.*,[31] may be that the electron-withdrawing effect of the triflate group results in a decrease of the aromatic hydroxylation in, *e.g.*, the liver (cytochrome P450). The three phenyl hydrogens were in general shifted downfield, for compounds with a sulfon ester group attached to the phenyl ring, as compared to compounds without a sulfon ester group (*e.g.*, OMe and OH groups). The three 'aromatic' hydrogens were always found in a range clearly above 7 ppm for the sulfon esters, while for the compounds without a sulfon ester group the range was ≤ 7 ppm. In that respect, no apparent difference between the triflate group and the structurally related mesylate group was observed

The descriptors included in this investigation did not provide any clues as to why the 7-triflates, *i.e.*, compounds **1** and **16**, have affinity for the D_3 receptor while the structurally related substituent, the mesylate group, was not present in any potent compounds. Actually, the affinity for the dopamine D_3 receptor of the *N*-propargyl-7-hydroxy compound, **20** (K_i = 13 nM), was reduced significantly to 1800 nM, after mesylation (compound **7**). To date, no explanation to why the triflate and the mesylate groups affect the *in vitro* and *in vivo* experiments differently, has been reported.

Chemistry

The *trans* geometry of **5** could be determined with NMR-spectrometry, since the difference in chemical shift between the *N*-benzyl methylene protons[3,4] was large (J = 215 Hz) and centered around δ 3.75, whereas the corresponding difference from the *cis*-**5** isomer could not be observed (singlet, δ 3.71). Additionally, the *trans* isomer was confirmed with single crystal X-ray analysis of *trans*-**19** (Figure 3.4), which crystallized in the triclinic P-1 space group with two molecules per unit cell (a = 8.233 Å; b = 9.423 Å; c = 11.480 Å; $\alpha = 103.55°$; $\beta = 98.62°$; $\gamma = 108.73°$).

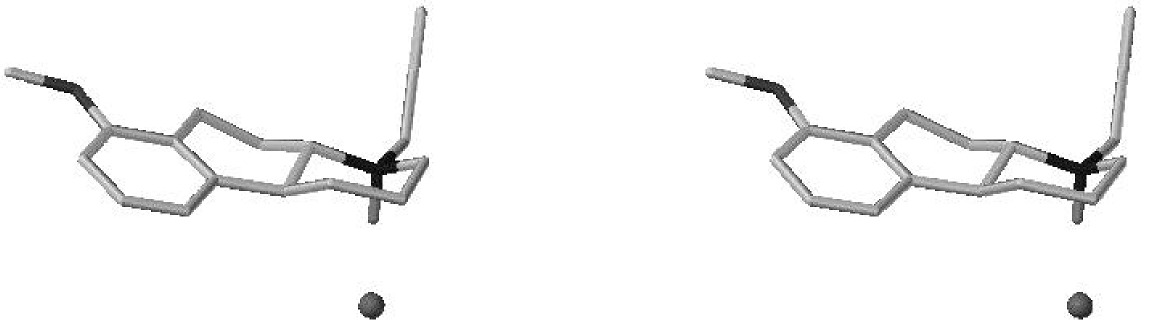


Figure 3.4 Stereo representation of the X-ray crystallographic structure from the hydrochloride salt of *trans*-**19**, which crystallized in the triclinic P-1 space group with two molecules per unit cell (a = 8.233 Å; b = 9.423 Å; c = 11.480 Å; α = 103.55°; β = 98.62°; γ = 108.73°).

Several attempts to separate the (+)- and the (–)-enantiomer of *trans*-**26** with fractional crystallization using (–)-dibenzoyl-L-tartaric acid were performed without success. In the following attempt, OMe-mandeloyl chloride, (*S*)-camphanic chloride and (+)-chlocyphos chloride were subsequently coupled to the nitrogen atom, but the separations of the diastereomers using thin layer chromatography were not sufficient. Eventually, the enantiomers of *trans*-**5** were successfully separated using semi-preparative HPLC.

3.6 Conclusions

Experimental design as a tool for rational drug design is effective provided that the descriptors used reflects the variations in the response variable. In the present investigation, the selected compounds were considered proper representatives of the large population of compounds from which they were selected. However, the response variables, *i.e.*, the affinities for the dopamine D_{2L}, D_3 and $D_{4.2}$ receptor subtypes, were found to correlate poorly with the descriptors, which complicated further QSAR analysis.

The compounds with high affinity for the three receptor subtypes, all had a hydroxy group attached at the seven position. A sulfon ester group at the seven position, however, suppressed the affinity for the $D_{4.2}$ receptor. It was also concluded that the *N* substituent may be as large as a phenylethyl group without detrimentally affect the ligand-receptor interaction. More difficult to rationalize is the low affinity for the $D_{4.2}$ receptor of compound **20** (*N*-propargyl-7-hydroxy-OHB[*f*]Q). One speculative explanation is that the somewhat rigid *N*-propargyl group may interfere in the ligand-$D_{4.2}$-receptor interaction. In addition, the significance of the low pK_a value (calculated to be 6.1) of compound **20** is hard to estimate, but may be a contributing factor to the low $D_{4.2}$ affinity.

3.7 Experimental Section

ː Computational Chemistry

All the compounds in Table 3.1 were built in SYBYL[14] by adding the proper substituents to a low energy conformation of *trans*-(4a*S*,10b*S*)-OHB[*f*]Q (*i.e.*, the enantiomer found active by Wikström *et.al.*[8]) Each molecule was energy minimized in SYBYL, using the Tripos molecular mechanics force field.[14,33] The 4-*N* was not protonated in any of the calculations. All settings were used default and all minimization iterations converged properly.

In order to generate some of the physicochemical descriptors in Table 3.2, Mopac AM1 single point calculations[13,14] with the keywords MULLIK, AM1, T=3600 and 1SCF activated, were performed.

ː Chemistry

General Remarks. NMR spectra were recorded at 200 or 300 MHz using a Varian Gemini 200 spectrometer. ^{1}H NMR chemical shifts are given in δ units (ppm) relative to the solvents and converted to the TMS scale using δ ($CDCl_3$) = 7.26 and δ (CD_3OD) = 3.30. ^{13}C NMR chemical shifts are given in δ units (ppm) relative to the solvents and converted to the TMS scale using δ ($CDCl_3$) = 76.91 and δ (CD_3OD) = 49.50. The splitting patterns are designated as follows: s (singlet), d (doublet), dd (double doublet), t (triplet), q (quartet), m (multiplet). Multiplets are given as the range from the first to the last peak, respectively. FT-IR spectra were obtained on a ATI-Mattson spectrometer. Elemental analyses were performed at Parke-Davis (Ann Arbor, MI) and were within 0.4 % of calculated percentages, if not stated otherwise. High resolution mass spectrometric analyses were performed at the Department of Chemistry at the University of Groningen. GC/MS mass spectra were recorded on a Unicam Automass 150 GC/MS system 1. Melting points were determined on an Electrothermal digital melting point apparatus and are uncorrected. Specific optical rotations were measured in methanol at RT on a Perkin Elmer 241 polariometer. For flash chromatography, silica gel 60 (0.040–0.063 mm, E. Merck, No. 9385) was used. All reagents used were commercially available and used without further purification.

Alkylation at the 4-*N* position. To a mixture of *trans*-**26** (1 eq.), dry acetonitrile and Cs_2CO_3 (approx. 3 eq.) an alkylating reagent (1.2 eq.) was added. The mixture was refluxed a couple of hours until the reaction was completed as indicated by GC or TLC. The reaction was quenched by adding water and EtOAc. The organic layer was separated and the water layer was extracted three times with EtOAc. The combined organic layers was washed once with brine, dried over Na_2SO_4, filtered and evaporated leaving an oil. The HCl salt was prepared and recrystallized.

Demethylation of the 7-methoxy group. The *N*-alkyl-7-methoxy-OHB[*f*]Q was refluxed under N_2 in 48 % HBr solution for two hours. The HBr was evaporated and the remaining *N*-alkyl-7-hydroxy-OHB[*f*]Q×HBr was recrystallized from an appropriate solvent.

***trans*-*N*-(*n*-Propyl)-7-[[(trifluoromethyl)sulfonyl]oxy]-OHB[*f*]Q (1).** Procedure as for (–)-**1** (below). mp 242–245 °C; ^{1}H NMR (CDCL$_3$) δ 7.33 (d, J=7.79, 1H), 7.23 (t, J=7.79, 1H), 7.10 (d, J=8.12, 1H), 3.09–2.98 (m, 2H), 2.89–2.10 (m, 8H), 1.89–1.76 (m, 2H), 1.70–1.43 (m, 3H), 1.37–1.16 (m, 1H), 0.91 (t, J=7.36, 3H); IR (KBr) 1144 (-SO$_2$O-), 1208 (C-F) cm^{-1}; Anal. ($C_{17}H_{22}NO_3SF_3$×HCl) C, N, H

(–)-*trans*-(4a*S*,10b*S*)-*N*-(*n*-Propyl)-7-[[(trifluoromethyl)sulfonyl]oxy]-OHB[*f*]Q ((–)-1). (–)-*trans*-(4a*S*,10b*S*)-*N*-Propyl-7-hydroxy-OHB[*f*]Q×HBr (64 mg, 0.20 mmol), 99 % *N*-phenyl-trifluoro-methane sulfonimide (111 mg, 0.29 mmol) and tetra-butyl-ammonium-hydrogen sulfate (a small spoonful) were suspended in CH_2Cl_2 (5 mL) and layered with 8 % NaOH (4 mL). The mixture was stirred vigorously for 19 hours and then quenched with water. The organic layer was separated and the water layer was extracted three times with CH_2Cl_2. The combined organic layers was washed once with 10 % $NaHCO_3$, dried over $MgSO_4$ and evaporated yielding a brownish oil (116 mg) that crystallized at RT. The product was purified with flash chromatography (gradient from pure CH_2Cl_2 to CH_2Cl_2/MeOH 30:1) and the HCl-salt was prepared. Recrystallization from aceton/ether yielded white crystals (46 mg, 57 %). mp 232–233 °C; $^{base}[\alpha]^{19}$ = -46.5° (MeOH c = 0.91); ^{1}H NMR (CDCL$_3$) δ 7.33 (d, J=6.93, 1H), 7.26 (t, J=7.77, 1H), 7.12 (d, J=8.97, 1H), 3.25–3.17 (m, 1H), 3.11–3.01 (m, 1H), 2.91–2.63 (m, 4H), 2.58–2.33 (m, 4H), 2.00–1.88 (m, 2H), 1.79–155 (m, 3H), 1.44–1.26 (m, 1H), 0.95 (t, J=7.34); ^{13}C NMR (CDCl$_3$) δ 148, 142, 129, 127.3, 125.7, 118.8, 115.5, 63.1, 54.8, 52.9, 41.5, 29.0, 25.0, 24.4, 23.2, 17.6, 11.7; FTIR (KBr) 1142 (-SO$_2$O-) cm^{-1}

(+)-*trans*-(4a*R*,10b*R*)-*N*-(*n*-Propyl)-[[(trifluoromethyl)sulfonyl]oxy]-OHB[*f*]Q ((+)-1). Procedure as for (–)-**1**. Yield (62 mg, 57 %). mp 231–232 °C; $^{base}[\alpha]^{20}$= +43.1° (MeOH c = 1.24); ^{1}H NMR (CDCl$_3$) δ 7.31–7.23 (m, 2H), 0.98 (t, J=7.31, 3H), 7.14 (d, J=7.12, 1H), 3.34 (d, J=9.84, 1H), 3.22–3.15 (m, 1H), 3.08–2.34 (m, 8H), 2.11–1.94 (m, 2H), 1.72–1.54 (m, 3H), 1.47–1.35 (m, 1H); ^{13}C NMR (CDCl$_3$) δ 147.8, 141.1, 128.7, 127.5, 125.7, 121.9, 119.1, 63.3, 54.4, 52.8, 40.7, 28.6, 24.4, 23.6, 23.1, 17.3, 11.5; FTIR (KBr) 1204 (C-F), 1142 (-SO$_2$O-) cm^{-1}

***trans*-*N*-Allyl-7-hydroxy-OHB[*f*]Q (2).** *Trans*-**18** (170 mg, 0.67 mmol) was dissolved in CH_2Cl_2 (4 mL) and added to a cooled (-60°C) 1M solution of BBr_3 (3.1 mL, 3.1 mmol) and dry CH_2Cl_2 (5 mL). The mixture was stirred at RT over day and boiled in MeOH (5 mL) for 15 minutes. The solvents were evaporated leaving a white solid which was triturated from ethanol. The white solid was filtered, washed with diethylether and dried (120 mg, 75 %). mp 257–260 °C; ^{1}H NMR (CD$_3$OD) δ 6.92–6.87 (m, 1H), 6.73 (d, J=7.82, 1H), 6.55 (t, J=7.45, 1H), 5.97–5.89 (m, 1H), 5.28–5.20 (m, 2H), 4.93 (s, 2H), 3.53 (d, J=13.44, 1H), 3.22–2.86 (m, 3H), 2.56–2.32 (m, 3H), 2.24–2.06 (m, 2H), 1.82–1.77 (m, 2H), 1.69–1.42 (m, 1H), 1.20–1.14 (m, 1H); FTIR (KBr) 3205 (-OH) cm^{-1}; MS (EIPI) 243; Anal: $C_{16}H_{21}NO$×HBr×½H_2O) C, H, N

***trans*-*N*-Phenylethyl-7-hydroxy-OHB[*f*]Q (3).** This compound has previously been characterized by Froimowitz.[23] mp 261–265 °C (lit.[23] 284–285); ^{1}H NMR (CD$_3$OD) δ 7.4–6.8 (m, 8H), 3.2–2.2 (m, 6H), 2.1–1.5 (m, 5H), 1.4–0.8 (m, 5H); FTIR (KBr) 3220 (-OH) cm^{-1}

***trans*-*N*-Methyl-7-methoxy-OHB[*f*]Q (4).** This compound was prepared from *trans*-**26** (160 mg, 0.74 mmol) following the general alkylation procedure above with methyl-iodide (70 µL, 1.1

mmol) leaving a colorless oil (230 mg, 135 %). mp 179–185 °C; ^{1}H NMR ($CDCl_3$) δ 7.23 (t, J=7.0, 1H), 6.91 (d, J=8.0, 1H), 6.74 (d, J=8.0, 1H), 4.00 (d, J=12.0, 1H), 3.74 (s, 3H), 3.63 (d, J=11, 1H), 3.51 (s, 2H), 3.17 (s, 3H), 3.11–2.99 (m, 1H), 2.78–2.65 (m, 1H), 2.63–2.36 (m, 2H), 2.22–1.92 (m, 2H), 1.79–1.46 (m, 2H); ^{13}C NMR ($CDCl_3$) δ 182.0, 161.8, 152.6, 148.5, 143.3, 133.2, 98.7, 90.1, 79.7, 68.9, 62.1, 53.0, 48.2, 47.6, 45.7; FTIR (KBr) 2936, 2835, 2589, 1585, 1464(s) cm^{-1}; MS (EIPI) 231; Anal: ($C_{15}H_{21}NO \times HCl$) C, N, H (0.65%)

***trans*-*N*-Benzyl-7-methoxy-OHB[*f*]Q (5).** *N*-Benzyl-7-methoxy-OHB[*f*]Q-3-on (23.78 g, 74.0 mmol) and $LiAlH_4$ (6.6 g, 0.95 mol) were mixed in THF (300 mL). The reaction was followed on GC which indicated an instant reaction. The reaction was quenched by consecutively adding H_2O (6 ml), NaOH (2 M, 6 mL) and H_2O (18 mL). The mixture was filtered, dried over $MgSO_4$, again filtered and the solvents were evaporated leaving a crude oil of *cis* and *trans* (22.40 g, 98 %). The *cis* and *trans* isomers were separated with gradient flash chromatography on a silica column starting with ether/petroleumether (ratio 5:1 with 0.1 % TEA) ending with pure ether yielding *cis* (3.36 g, 15 %), *trans* (4.87 g, 21 %) and a mixture of *cis* and *trans* (8.07 g, 36 %).

***trans*-5.** mp 232–235 °C; ^{1}H NMR ($CDCl_3$) δ 7.33–7.17 (m, 5H), 7.11 (t, J=7.69, 1H), 6.89 (d, J=8.06, 1H), 6.66 (d, J=8.06, 1H), 4.11 (d, J=13.18, 1 benzyl-H), 3.78 (s, 3H), 3.39 (d, J=13.19, 1 benzyl-H), 2.98–2.89 (m, 2H), 2.70–2.55 (m, 2H), 2.51–2.39 (m, 2H), 2.17–1.99 (m, 2H), 1.75–1.64 (m, 2H), 1.62–1.47 (m, 1H), 1.25–1.09 (m, 1H); ^{13}C NMR ($CDCl_3$) δ 156.7, 140.82, 138.9, 129.0, 128.0, 126.6, 126.1, 124.8, 117.6, 107.0, 63.6, 56.9, 55.1, 53.0, 42.3, 29.6, 26.8, 25.1, 23.0; FTIR (KBr) 3008, 2952, 1583, 1463 cm^{-1}; Anal ($C_{21}H_{25}NO \times HCl \times ½H_2O$) C, H, N

***cis*-5.** mp 216–221 °C; ^{1}H NMR ($CDCl_3$) δ 7.34 (d, J=7.33, 2H), 7.25 (t, J=7.33, 2H), 7.18 (d, J=8.06, 1H), 7.05 (t, J=7.69), 6.69 (d, J=8.06), 6.59 (d, J=8.05), 3.74 (s, 3H), 3.71 (s, 2H, -CH2-Ph), 3.06–2.96 (m, 1H), 2.95–2.91 (m, 1H), 2.50–2.47 (m, 1H), 2.43–2.29 (m, 1H), 2.00–1.84 (m, 3H), 1.79–1.47 (m, 4H), 1.28–1.09 (m, 1H); ^{13}C NMR ($CDCl_3$) δ 156.9, 143.0, 139.5, 128.6, 128.0, 126.6, 126.0, 124.6, 120.9, 106.7, 58.7, 57.0, 55.0, 46.1, 39.9, 29.9, 25.3, 22.8, 14.9; HR-MS Calcd (Obsd) for $C_{21}H_{25}NO$ 307.194 (307.192)

Enantiomeric separation of *trans*-*N*-Benzyl-7-methoxy-OHB[*f*]Q (5). The separation of the enantiomers was performed by means of semi-preparative HPLC on a Chiralcel OD (250 x 10 mm) column. A stock solution of the HCl salt of the racemic **5** and ethanol (100 mg/mL) was prepared. Each time 100 μL was injected on the column. The mobile phase used was ethanol mixed with diethylamine (0.1 %), in order to minimize the peak-tailing, at a flow rate of 1.5 mL per minute. The eluent was monitored with a UV-detector (270 nm). R_S: 2.02; α: 1.24. (Separation on a analytical Chiralcel OD column, flow rate: 0.5 mL/min; mobile phase: EtOH (gradient grade); $R_S = 1.65$; $\alpha = 1.62$)

(–)-*trans*-(4a*R*,10b*R*)-*N*-Benzyl-7-methoxy-OHB[*f*]Q ((–)-5). The (–)-enantiomer was the least retained one. 106 mg of the pure enantiomer was obtained.

(+)-*trans*-(4a*S*,10b*S*)-*N*-Benzyl-7-methoxy-OHB[*f*]Q ((+)-5). The (+)-enantiomer was the most retained one. 128 mg of the pure enantiomer was obtained.

***trans*-*N*-(*n*-Butyl)-7-hydroxy-OHB[*f*]Q (6).** Previously characterized by Wikström *et al.*[5] mp 270–275 °C (lit: 277–279 °C); HR-MS Calcd (Obsd) for $C_{17}H_{25}NO$ 259.194 (259.195)

***trans*-*N*-(1-Prop-2-ynyl)-7-[[(methane)sulfonyl]oxy]-OHB[*f*]Q (7).** *Trans*-**20** (60 mg, 0.25 mmol) was dissolved in dry CH_2Cl_2 (5 mL), a small amount of TEA (4 drops) was added followed by CH_3SO_2Cl (30 µL, 0.37 mmol). The mixture was stirred under N_2 at RT over night. The reaction was quenched by adding 10 % $NaHCO_3$ (3 mL). The organic layer was separated and the water layer was extracted three times with CH_2Cl_2. The combined organic layers was washed once with brine, dried over Na_2SO_4, filtered and evaporated leaving a crude oil (80 mg, 0.25 mmol). The oil was converted into the HCl salt. mp 225–227 °C; ^{1}H NMR ($CDCl_3$) δ 7.30–7.20 (m, 3H), 4.31 (d, J=17, 1H), 3.83 (d, J=17, 1H), 3.44 (m, 2H), 3.21–3.06 (m, 5H), 2.9–2.8 (m, 1H), 2.71–2.23 (m, 4H), 2.20–2.00 (m, 2H), 1.46–1.28 (m, 2H); ^{13}C NMR ($CDCl_3$) δ 146.7, 138.5, 128.4, 127.4, 124.6, 120.2, 79.7, 70.4, 63.0, 53.1, 42.4, 39.0, 38.2, 27.4, 22.5, 22.2; FTIR (KBr) 1173 (s, $-SO_2O-$) cm^{-1}; HR-MS Calcd (Obsd) for $C_{17}H_{21}NO_3S$ 319.124 (319.125)

***trans*-*N*-Benzyl-7-[[(phenyl)sulfonyl]oxy]-OHB[*f*]Q (8).** The HBr salt of *trans*-*N*-Benzyl-7-hydroxy-OHB[*f*]Q (70 mg, 0.19 mmol) was suspended in CH_2Cl_2 (4 mL) and layered with 10 % NaOH (4 mL). A catalytic amount of $Bu_4NH_4HSO_4$ was added. Subsequently, benzenesulfonyl chloride solved in CH_2Cl_2 (1.5 mL) was added dropwise. The mixture was stirred at RT for 30 hours, water was added and the organic layer was separated. The water layer was extracted with CH_2Cl_2, the combined organic layers was dried over Na_2SO_4, filtered and evaporated yielding a brownish oil. The oil was purified by flash chromatography on a silica column leaving a colorless oil (120 mg, 0.28 mmol). The HCl salt was prepared. mp 257–258 °C; ^{1}H NMR ($CDCl_3$) δ 7.90 (d, J=7.32, 2H), 7.69 (t, J=7.69, 1H), 7.55 (t, J=7.69, 2H), 7.31 (s, 3H), 7.27–7.19 (m, 3H), 7.09 (t, J=8.06, 1H), 6.83 (d, J=8.06, 1H), 4.10 (d, J=13.92, 1H), 3.32 (d, J=13.55, 1H), 2.85 (m, 2H), 2.63–2.60 (m, 2H), 2.30 (m, 2H), 2.08–2.01 (m, 2H), 1.73–1.71 (m, 2H), 1.26–1.21 (m, 2H); ^{13}C NMR ($CDCl_3$) δ 147.4, 142.5, 139.0, 136.2, 134.2, 130.0, 129.0, 128.1, 127.7, 126.8, 126.3, 126.1, 124.3, 119.2, 63.4, 57.2, 53.1, 42.2, 29.4, 26.5, 25.0, 23.2; FTIR (KBr) (s, $-SO_2O-$) cm^{-1}; MS (EIPI) 433; Anal: ($C_{26}H_{27}NO_3S\times HCl\times \frac{1}{2} H_2O$) C, H, N

***trans*-*N*-Ethyl-7-[[(methane)sulfonyl]oxy]-OHB[*f*]Q (9).** Triethylamine and methanesulfonyl chloride (0.013 mL, 0.17 mmol) were added to a suspension of *trans*-**21** (60 mg, 0.19 mmol) in CH_2Cl_2 and dioxan. The mixture was stirred at RT over night and quenched with 10 % NaOH. The organic layer was separated, the basic water layer was extracted once with CH_2Cl_2 and the organic layers were combined. Small solids in the organic layer were formed but GC and TLC indicated no product and were consequently filtered off. The mother liquor was dried over Na_2SO_4, filtered and evaporated leaving a solid (50 mg) which still contained starting-material. The reaction was repeated (mesyl chloride 0.015 mL) followed by the same work up procedure. The obtained product was purified on a silica column. mp 221–225 °C; ^{1}H NMR δ 7.23–7.11 (m, 3H), 3.19 (s, 3H), 3.18–2.93 (m, 2H), 2.91–2.58 (m, 3H), 2.48–2.43 (m, 2H), 2.36–2.23 (m, 2H), 2.18–2.10 (m, 2H), 1.86–1.80 (m, 2H), 1.58–1.52 (m, 1H), 1.03 (t, J=7.14, 3H); ^{13}C NMR ($CDCl_3$) δ 147, 142.5, 130, 126.7, 124.6, 119.0, 62.1, 51.9, 46.6, 42.4, 38.3, 29.6, 25.8, 25.7, 23.6, 9.5; FTIR (KBr) 2928 (s), 2451 (s), 1342, 1166 (s,-SO_2O-) cm^{-1}; MS (EIPI) 309; HR-MS Calcd (Obsd) for $C_{16}H_{23}NO_3S$ 309.140 (309.141)

***trans*-*N*-[2-(2-Thienyl)-ethyl]-7-[[(phenyl)sulfonyl]oxy]-OHB[*f*]Q (10).** *Trans*-*N*-[2-(2-Thienyl)-ethyl]-7-hydroxy-OHB[*f*]Q (24 mg, 0.07 mmol), TEA and dry CH_2Cl_2 were mixed before benzenesulfonyl chloride (30 µL, 0.24 mmol) was added. The mixture was stirred at RT for 5 hours. The reaction was quenched with 10 % Na_2CO_3 and the organic layer was separated. The water layer was extracted two times with CH_2Cl_2 and the combined organic layers was washed with brine, dried over Na_2SO_4, filtered and evaporated leaving a crude oil (63 mg, 0.14 mmol). mp 190–199 °C; ^{1}H NMR ($CDCl_3$) δ 8.0–7.5 (m, 5H), 7.3–7.0 (m, 3H), 6.9–6.7 (m, 3H), 3.8–3.3 (m, 5H), 3.1–2.5 (m, 5H), 2.4–2.0 (m, 2H), 1.5–1.3 (m, 4H); MS (EIPI) *m/z* (rel. intensity, 170 eV) 454 (M+, 100), 356 (13), 313 (37), 216 (18), 185 (4); MS (CI+) *m/z* (rel. intensity) 454 (M+, 100), 344 (6), 314 (26), 160 (8) for $C_{25}H_{27}NO_3S_2$

***trans*-7-[[(Methyl)sulfonyl]oxy]-OHB[*f*]Q (11).** *Trans*-**26** (147 mg, 0.7 mmol) was suspended in dioxan (8 mL) followed by addition of $NaHCO_3$ (4 mL half saturated in H_2O) and FMOC-chloride (300 mg, 1.2 mmol) solved in dioxan (5 mL). The suspension was left stirring over night at RT, quenched with water (5 mL) and extracted with EtOAc (3 × 10 mL). The combined organic layers was dried over Na_2SO_4, filtered and evaporated yielding an yellow oil which was purified by flash chromatography leaving a white solid (310 mg, 100 %).

The obtained *trans*-*N*-FMOC-7-methoxy-OHB[*f*]Q (310 mg, 0.7 mmol) was added to ethylmercaptane (5 mL) together with $AlCl_3$ (330 mg, 2.5 mmol). The mixture was stirred at RT for 3.5 hours and quenched with ice water. The reaction mixture was extracted with chloroform (3 × 15 mL) and ethylacetat (2 × 10 mL), the combined organic layers was dried over Na_2SO_4, filtered and evaporated leaving a white solid (211 mg, 71 %).

The intermediate, *trans*-*N*-FMOC-7-hydroxy-OHB[*f*]Q (103 mg, 0.24 mmol) was, without further purification, mixed with triethylamine (a few drops) and dry dichlorometane (9 mL) before methanesulfonyl chloride (32 µL, 0.4 mmol) was added. The mixture was stirred at RT for 2.5 hours and quenched with 10 % Na_2CO_3. The basic water layer was extracted with dichloromethane (3 × 10 mL). The combined organic layers was washed once with brine, dried over Na_2SO_4, filtered and evaporated leaving a white solid (110 mg, 91 %).

The final product, *trans*-**11**, was obtained from *trans*-*N*-FMOC-**7**-[[(methyl)sulfonyl]oxy]-OHB[*f*]Q (110 mg, 0.22 mmol) which was stirred in a 25 % piperidine/CH_2Cl_2 solution for 15 minutes. The solvents were evaporated and the remaining solids were purified by flash chromatography. mp 270–271 °C**;** ^{1}H NMR (CD_3OD) δ 7.31 (m, 3H), 3.48 (d, J=12.7), 3.3 (s, 3H), 3.25–3.09 (m, 2H), 3.04–2.86 (m, 2H), 2.68 (d, J=12.1), 2.28–2.09 (m, 2H), 2.01–1.84 (m, 2H), 1.56–1.48 (m, 1H); ^{13}C NMR ($CDCl_3$) δ 141.9, 129.7, 126.6, 124.2, 119.0, 63.1, 58.2, 46.4, 43.1, 38.1, 29.4, 26.5, 23.2, (147); FTIR (KBr) 2934 (s), 2532, 2362 (m), 1334, 1170 (s, -SO_2O-) cm^{-1}; MS (EIPI) 281; Anal: ($C_{14}H_{19}NO_3S \times HCl \times ½\ H_20$) C, H, N

***trans*-*N*-Phenylethyl-7-[[(4-toluoyl)sulfonyl]oxy]-OHB[*f*]Q (12).** *Trans*-**3** (31 mg, 0.08 mmol), TEA and dry CH_2Cl_2 were mixed before p-toluensulfonyl chloride (32 mg, 0.17 mmol) was added. The mixture was stirred at RT for 5 hours. The reaction was quenched with 10 % Na_2CO_3 and the organic layer was separated. The water layer was extracted two times with CH_2Cl_2 and the combined organic layers was washed with brine, dried over Na_2SO_4, filtered and evaporated leaving a crude oil which was converted to the HCl salt. Recrystallization from ethanol yielded brown crystals. mp 290 °C dec.; MS (EIPI) *m/z* (rel. intensity, 170 eV) 462 (M+, 11), 239 (100), 102 (76) for $C_{28}H_{31}NO_3S$

***trans*-*N*-(*n*-Propyl)-7-[[(4-toluoyl)sulfonyl]oxy]-OHB[*f*]Q (14).** To a solution of dioxan (8 mL), *trans*-**17** (60 mg, 0.18 mmol) and a few TEA drops, p-toluensulfonyl chloride (90 mg, 0.47 mmol) was added. The reaction was complete after one nights stirring as indicated by GC/MS. Dichloromethane and 10 % $NaHCO_3$ was added. The water layer was extracted three times with dichloromethane, and the combined organic layers was dried over Na_2SO_4, filtered and evaporated yielding a brownish oil. After purification by flash chromatography on a silica-column (MeOH/CH_2-Cl_2 1:15) the HCl salt was prepared yielding an oil (80 mg, 100 %) which crystallized while standing. The solid was recrystallized from ethanol. mp 217–219 °C; ^{1}H NMR ($CDCl_3$) δ 7.74 (d, J=8.30, 2H), 7.32 (d, J=8.06, 2H), 7.15 (t, J=8.30, 1H), 7.07 (d, 7.81, 1H), 6.80 (d, J=7.81, 1H), 3.07 (d, J=11.47, 1H), 2.83–2.51 (m, 4H), 2.45 (s, 3H), 2.40–2.21 (m, 2H), 2.19–2.07 (m, 2H), 1.87–1.78 (m, 2H), 1.60–1.41 (m, 3H), 1.27–1.19 (m, 2H), 0.88 (t, J=7.33, 3H); ^{13}C NMR ($CDCl_3$) δ 147.4, 145.2, 141.8, 133.1, 129.9, 129.6 (2C), 128.2 (2C), 126.3, 124.1, 119.3, 62.6, 54.7, 52.6, 41.5, 29.1, 25.2, 24.7, 23.0, 21.5, 17.4, 11.7; FTIR (KBr) 1187 (-SO_2O-) cm^{-1}; MS (EIPI) 399; Anal: ($C_{23}H_{29}NO_3S \times HCl \times 1\frac{1}{4}\ H_2O$) C, H (0.5%), N

***trans*-*N*-Allyl-7-[[(2-thienyl)sulfonyl]oxy]-OHB[*f*]Q (15).** *Trans*-**2** (80 mg, 0.33 mmol), TEA and dry CH_2Cl_2 were mixed before 2-thienyl-sulfonyl chloride (100 mg, 0.55 mmol) was added. The mixture was stirred at RT for 5 hours. The reaction was quenched with 10 % Na_2CO_3 and the organic layer was separated. The water layer was extracted three times with CH_2Cl_2 and the combined organic layers was washed with brine, dried over Na_2SO_4, filtered and evaporated leaving a crude oil. After purification by flash chromatography a brown/red oil was left. The HCl salt was prepared and recrystallized from ethanol leaving a white solid (49 mg, 36 %). mp 157–166 °C; ^{1}H NMR ($CDCl_3$) δ 7.76–7.73 (m, 1H), 7.62–7.60 (m, 1H), 7.23–7.22 (m, 1H), 6.90–6.83 (m, 1H), 6.03–5.84 (m, 1H), 5.46 (s, 1H), 5.39 (d, J=4.64, 1H), 3.78–3.53 (m, 2H), 3.37 (d, J=9.83, 1H), 3.26–3.08 (m, 1H), 2.91–2.43 (m, 5H), 2.35–2.10 (m, 2H), 2.01–1.80 (m, 2H), 1.41 (dd, (d, J=11.79), (d, J=13.76)); ^{13}C NMR ($CDCl_3$) δ 147.3, 139.6, 135.1, 134.8, 129.1, 127.7, 127.0, 126.4, 124.5, 124.1, 119.8, 63.4, 54.7, 52.1, 39.7, 27.8, 23.5, 22.4; FTIR (KBr) 3070, 2928, 2508, 1354, 1182 (-SO_2O-) cm^{-1}; Anal: ($C_{19}H_{23}NO_3S_2 \times HCl$) C, H, N

***trans*-*N*-(*n*-Butyl)-7-[[trifluoromethane)sulfonyl]oxy]-OHB[*f*]Q (16).** *Trans*-**6** (90 mg, 0.35 mmol) was mixed with CH_2Cl_2 (4 mL), 99 % *N*-phenyl-trifluoro-methane sulfonimide (210 mg, 0.59 mmol), $Bu_4NH_4HSO_4$ (catalytic amount) and layered with 8 % NaOH (4 mL). The mixture was stirred vigorously under N_2 over night and quenched with 10 % HCl. An attempt to perform an acid/base extraction failed because the product was too lipophilic and remained in the organic layer even after extraction with acid. Therefore, the water layer was extracted with CH_2Cl_2. The combined organic layers was dried over Na_2SO_4, filtered and evaporated. The product was purified with flash chromatography. mp 217 °C; ^1H NMR ($CDCl_3$) δ 7.31–7.24 (m, 2H), 7.17–7.13 (m, 1H), 3.74 (t, J=10.01, 1H), 3.58–3.46 (m, 1H), 3.24–3.09 (m, 3H), 2.88–2.75 (m, 3H), 2.63 (d, J=11.72, 2H), 2.40–2.38 (m, 2H), 2.07–2.03 (m, 1H), 1.83–1.24 (m, 5H), 0.99 (t, J=7.20, 3H); FTIR (KBr) 2966, 2416, 1411, 1209(C-F), 1140(-SO_2O-) cm^{-1}; HR-MS Calcd (Obsd) for $C_{18}H_{24}NO_3SF_3$ 391.143 (391.142).

***trans-N*-Allyl-7-methoxy-OHB[*f*]Q (18).** This compound was prepared from *trans*-**26** (180 mg, 0.83 mmol) following the general alkylation reaction (above) with allyl-bromide (86 μL, 1 mmol) leaving a yellowish oil (220 mg, 0.86 mmol). mp 200–209 °C; ^1H NMR ($CDCl_3$) δ 7.15 (t, J=7.94, 1H), 6.93 (d, J=7.81, 1H), 6.69 (d, J=7.81, 1H), 5.93–5.89 (m, 1H), 5.24–5.15 (m, 2H), 3.80 (s, 3H), 3.54 (d, J=14.16, 1H), 3.21–2.89 (m, 3H), 2.67 (m, 3H), 2.28–2.08 (m, 2H), 1.84–1.77 (m, 2H), 1.57–1.51 (m, 2H), 1.3–1.1 (m, 1H); ^{13}C NMR ($CDCl_3$) δ 156.7, 140.6, 134.4, 126.1, 124.7, 117.63, 117.59, 106.9, 62.9, 56.1, 55.0, 52.9, 42.4, 29.5, 25.8, 25.3, 22.8**;** MS (EIPI) 257 ; HR-MS Calcd (Obsd) for $C_{17}H_{23}NO$ 257.178 (257.180)

***trans-N*-(1-Prop-2-ynyl)-7-methoxy-OHB[*f*]Q (19).** This compound was prepared from *trans*-**26** (190 mg, 0.88 mmol) following the general alkylation procedure above with propargyl chloride (100 μL, 1.38 mmol) leaving a yellow oil (230 mg, 0.9 mmol). mp 213–222 °C; ^1H NMR ($CDCl_3$) δ 7.17 (t, J=7.94, 1H), 6.95 (d, J=8.05, 1H), 6.71 (d, J=8.05, 1H), 3.90 (d, J=17.71, 1H), 3.42 (d, J=1758, 1H), 3.05–2.84 (m, 2H), 2.74–2.27 (m, 6H), 2.21 (t, J=2.44, 1H), 1.92–1.79 (m, 2H), 1.59–1.41 (m, 1H), 1.38–1.15 (m, 1H); ^{13}C NMR ($CDCl_3$) δ 156.8, 140.3, 126.1, 124.7, 117.7, 107.0, 73.1, 61.1, 55.0, 42.7, 42.3, 29.4, 25.6, 25.4, 22.7; FTIR (KBr) 3156 (s), 2962 (m), 2365, 2307 (w), 1584 (s), 1468 (s) cm^{-1}; MS (EIPI) 255; Anal: ($C_{17}H_{21}NO \times HCl$) C, H, N

***trans-N*-(1-Prop-2-ynyl)-7-hydroxy-OHB[*f*]Q (20).** A mixture of 1M BBr_3 (1.9 mL) and dry CH_2Cl_2 (3 mL) were cooled to -60°C before a solution of *trans*-**19 (**100 mg, 0.39 mmol) and CH_2Cl_2 (3.6 mL) was added dropwise. After addition, the temperature was allowed to reach RT and was left stirring over night. The reaction was cooled to 0°C before MeOH (4 mL) was added, followed by 15 minutes refluxing. The solvents were evaporated the HCl salt was prepared. mp 251–252 °C; ^1H NMR ($CDCl_3$) δ 6.94 (t, J=7.82, 1H), 6.84 (d, J=7.81, 1H), 6.58 (d, J=7.81, 1H), 3.81 (d, J=17.58, 1H), 3.40 (d, J=17.58, 1H), 3.34 (s, 1H), 2.95–2.87 (m, 2H), 2.64–2.29 (m, 6H), 1.88–1.81 (m, 2H), 1.45–1.10 (m,2H); ^{13}C NMR ($CDCl_3$) δ 154.1, 139.5, 125.8, 122.2, 116.3, 111.3, 61.7, 52.7, 41.8, 41.7, 29.2, 24.9, 24.8, 22.2; FTIR (KBr) 3194 (s, -OH), 2930 (s), 2556 (s), 1585 (s), 1466 (s), 1272 (s) cm^{-1}; MS (EIPI) 241; HR-MS Calcd (Obsd) for $C_{16}H_{19}NO$ 241.147 (241.147)

***trans-N*-Ethyl-7-hydroxy-OHB[*f*]Q (21).** The intermediate *trans-N*-Ethyl-7-methoxy-OHB[*f*]Q was prepared from *trans*-**25** (310 mg, 1.4 mmol) following the general alkylation procedure (above) with iodoethane (190 μL, 2.4 mmol) leaving a white solid (310 mg, 89 %). ^1H NMR ($CDCl_3$) δ 7.13 (t, J=8.06, 1H), 6.90 (d, J=8.05, 1H), 6.68 (d, J=8.06, 1H), 3.79 (s, 3H), 3.08–2.92 (m, 3H), 2.81–2.32 (m, 6H), 1.87–1.83 (m, 2H), 1.48–1.41 (m, 3H), 1.07 (t, J=7.14, 3H);

^{13}C NMR ($CDCl_3$) δ 156.8, 140.7, 126.2, 124.8, 117.8, 107.1, 62.4, 55.2, 52.1, 46.6, 42.5, 29.8, 25.9, 25.5, 23.1, 9.3; MS (EIPI) 245

Trans-*N*-Ethyl-7-methoxy-OHB[*f*]Q (220 mg, 0.9 mmol) was used without further purification to prepare *trans*-**21,** following the general demethylation procedure (above) yielding a white/brownish solid (270 mg, 96 %). The solid was recrystallized from MeOH. mp 300–302 °C; ^1H NMR (CD_3OD) δ 7.03 (t, J=7.33, 1H), 6.85 (d, 1H), 6.65 (d, J=7.8, 1H), 3.47–3.57 (m, 2H), 3.16–3.30 (m, 6H), 3.03 (m, 2H), 2.60–2.66 (m, 2H), 2.10 (m, 2H), 1.80 (m, 1H), 1.33–1.37 (m, 3H); FTIR (KBr) 3192, 2924, 2743, 2679, 1585, 1468, 1273 cm^{-1}; HR-MS Calcd (Obsd) for $C_{15}H_{21}NO$ 231.162 (231.162)

ː *In vitro* Pharmacology

Cell Culture. CHO K1 cells stably transfected with the genes of wild type D_{2L}, D_3, and $D_{4.2}$ receptors were grown and harvested as previously described.[34-36]

Radioligand Binding. CHO K1 cells expressing the human DA D_{2L}, D_3, and $D_{4.2}$ receptors were removed by replacement of growth medium with PBS-EDTA (0.02 % EDTA in phosphate buffered saline). After swelling for 5-10 min, the cells were scraped from the flasks, and centrifuged at about 1000 x g for 5 min. The cells were then resuspended in 50 mM Tris-HCl binding buffer pH 7.4 at room temperature (50 mM Tris-HCl, 1 mM EDTA, 1.5 mM CaCl2, 5 mM KCl, 120 mM NaCl and 5 mM $MgCl_2$). The membranes were pelleted by centrifugation at 20,000 × g at 4°C for 20 min. The supernatant fluid was removed and the pellets were resuspended and homogenized with a Brinkman Polytron (setting 5 for 15 sec) in the binding buffer and 1 mL aliquots stored at -80°C until used in the binding assay.

Binding assays were carried out in duplicate in 1.4 mL microtubes (Marsh Biomedical Products, Inc.). Each tube received 50 µL of competing drug or binding buffer, 50 µL of [^{3}H]spiperone (final concentration was 0.2 nM for D_{2L} and $D_{4.2}$ and 0.5 nM for D_3) and 0.4 mL membranes (15–30 µg protein) to give a final volume of 0.5 mL. After 60 min incubation at 25°C, the incubations were terminated by rapid filtration over GF/B filters presoaked in 0.5 % polyethylenimine and washed rapidly with 3 × 1 mL ice-cold buffer. Filters were put in scintillation vials, 4 mL of Beckman Ready Gel Scintillation fluid was added and the radioactive content determined by liquid scintillation spectrophotometry. Non-specific binding was defined in presence of 1 µM haloperidol. Data for IC_{50} values were analyzed using the iterative nonlinear least square curve-fitting program LIGAND. The dissociation constant, K_i, was derived from the concentration, C, for 50 % inhibition of binding, using $K_i = C/(1 + C^*/K_d)$ where C* was the concentration of [^{3}H]spiperone and the K_d was 0.116 nM, 0.152 nM and 0.093 nM for DA D_{2L}, D_3 and $D_{4.2}$ receptors, respectively.[37] Experimental compounds were made up as stock solutions in dimethyl sulfoxide (DMSO). The final concentration of 0.1 % DMSO in the incubation mixture had no effect on specific binding.

3.8 References

1. Cannon, J.G.; Khonje, P.R.; Long, J.P. Centrally Acting Emetics. 9. Hofmann and Emde Degradation Products of Nuciferine. *J. Med. Chem.* **1975**, *18,* 110-112.
2. Cannon, J.G. and Hatheway, G.J. Centrally Acting Emetics. 10. Rigid Dopamine Congeners Derived from Octahydrobenzo[*f*]quinoline. *J. Med. Chem.* **1976**, *19,* 987-993.
3. Cannon, J.G.; Suarez Gutierrez, C.; Lee, T.; Long, J.P.; Costall, B.; Fortune, D.H.; Naylor, R.J. Rigid Congeners of Dopamine Based on Octahydrobenzo[*f*]quinoline: Peripheral and Central Effects. *J. Med. Chem.* **1979**, *22,* 341-347.
4. Cannon, J.G.; Lee, T.; Goldman, D.; Long, J.P.; Flynn, J.R.; Verimer, T.; Costall; Naylor, R.J. Congeners of the Beta Conformer of Dopamine Derived from *cis*- and *trans*-Octahydrobenzo[*f*]quinoline and *trans*-Octahydrobenzo[*g*]quinoline. *J. Med. Chem* **1979**, *23,* 1-5.
5. Wikström, H.; Sanchez, D.; Lindberg, P.; Arvidsson, L.; Hacksell, U.; Johansson, A.M.; Nilsson, J.L.; Hjorth, S.; Carlsson, A. Monophenolic Octahydrobenzo[f]quinolines: Central Dopamine- and Serotonin-Receptor Stimulating Activity. *J. Med. Chem.* **1982**, *25,* 925-931.
6. Sonesson, C.; Lin, C.H.; Hansson, L.; Waters, N.; Svensson, K.; Carlsson, A.; Smith, M.W.; Wikström, H. Substituted (*S*)-Phenylpiperidines and Rigid Congeners as Preferential Dopamine Autoreceptor Antagonists: Synthesis and Structure-Activity Relationships. *J. Med. Chem.* **1994**, *37,* 2735-2753.
7. Katerinopoulos HE, Tagmatarchis N, Thermos K. *N*-Iodopropenyl-Octahydrobenzo[*f*]- and -[*g*]quinoline Analogs with Adrenergic and Dopaminergic Activity. [Abstract] 214th ACS National meeting, Las Vegas, 7-11 september, 1997.
8. Wikström, H.; Andersson, B.; Sanchez, D.; Lindberg, P.; Arvidsson, L.E.; Johansson, A.M.; Nilsson, J.L.; Svensson, K.; Hjorth, S.; Carlsson, A. Resolved Monophenolic 2-Aminotetralins and 1,2,3,4,4a,5,6,10b-octahydrobenzo[*f*]quinolines: Structural and Stereochemical Considerations for Centrally Acting Pre- and Postsynaptic Dopamine-Receptor Agonists. *J. Med. Chem.* **1985**, *28,* 215-225.
9. Hjorth, S.; Svensson, K.; Carlsson, A.; Wikström, H.; Andersson, B. Central Dopaminergic Properties of HW-165 and its Enantiomers; *Trans*-octahydrobenzo[*f*]quinoline Congeners of 3-PPP. *Arch Pharmacol* **1986**, *333,* 205-218.
10. Wikström, H.; Andersson, B.; Elebring, T.; Svensson, K.; Carlsson, A.; Largent, B. *N*-Substituted 1,2,3,4,4a,5,6,10b-octahydrobenzo[*f*]quinolines and 3-Phenylpiperidines: Effects on Central Dopamine and Sigma Receptors. *J. Med. Chem.* **1987**, *30,* 2169-2174.
11. Liljefors, T.; Bøgesø, K.P.; Hyttel, J.; Wikström, H.; Svensson, K.; Carlsson, A. Pre- and Postsynaptic Dopaminergic Activities of Indolizidine and Quinolizidine Derivatives of 3-(3-Hydroxyphenyl)-*N*-(n-propyl)piperidine (3-PPP). Further Developments of a Dopamine Receptor Model. *J. Med. Chem.* **1990**, *33,* 1015-1022.
12. Froimowitz, M.; Deng, Y.; Jacob, J.N.; Li, N.; Cody, V. Drug Design and Discovery. Dopaminergic (4a*R*,10b*S*)-*cis*- and (4a*S*,10b*S*)-*trans*-Octahydrobenzoquinolines have Similiar Pharmacophores. Harward Academic Publishers GmbH.: **1995**; pp. 73-81.
13. Dewar, M.J.S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107,* 3902-3909.
14. SYBYL- Molecular Modeling Software, *6.3,* Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA
15. Pallas 1.2, *1.2,* CompuDrug Chemistry Ltd. Program for calculation of logP, logD, and pKa, MS-Windows 3.1.
16. Rekker, R.F. and De Kort, H.M. The Hydrophobic Fragmental Constant: an Extension to a 1000 Data Point Set. *Eur. J. Med. Chem* **1979**, 479-488.
17. Clementi, S.; Cruciani, G.; Baroni, M.; Skagerberg, B. QSAR: Rational Approaches to the Design of Bioactive Compounds. Selection of Informative Structures for QSAR Studies. Silipo, C. and Vittoria, A. Eds. Elsevier Science Publishers B. V. Amsterdam, **1991**; pp. 217-222.
18. Casey, D.E. Clozapine: Neuroleptic-Induced EPS and Tardive Dyskinesia. *Psychopharmacology* **1989**, *99,* S47-S53.
19. Wold, S., Albano, C., Dunn III, W.J., Esbenssen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., Sjöström, M., Skagerberg, B., Wikström, C., Öhman, J. Multivariate Data Analyses: Converting Chemical Data Tables to Plots. 1985 Jun 10; 1985;
20. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185,* 1-17.
21. Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons: New York, **1989**;
22. Morgan, E. Chemometrics: Experimental Design. Chadwick, N. Ed.; John Wiley and Sons Ltd: Chichester, **1991**;
23. Froimowitz, M. and Jacob, J.N. inventors. Octahydrobenzo[*f*]quinoline-Based Receptor Agonists and Antagonists. WO 95/14006. MA. Date Filed: **1993/11/19.**
24. Cannon, J.G.; Chang, Y.-a.; Amoo, V.E.; Walker, K.A. Stereospecific Route to *trans*-1,2,3,4,4a,5,6,10b-Octahydrobenzo[*f*]quinolines. *Synthesis* **1986**, 494-496.

25. McDermed, J.D.; Freeman, H.S.; Ferris, R.M. Catecholamines: Basic and Clinical Frontiers. Usdin, E.; Kopin, I.J.; Barchas, J. Eds.; Pergamon Press: New York, **1979**; pp. 568-577.
26. Liljefors, T. and Wikström, H. A Molecular Mechanics Approach to the Understanding of Presynaptic Selectivity for Centrally Acting Dopamine Receptor Agonists of the Phenylpiperidine Series. *J. Med. Chem.* **1986**, *29*, 1896-1904.
27. Teeter, M.M.; Froimowitz, M.; Stec, B.; DuRand, C.J. Homology Modeling of the Dopamine D_2 Receptor and its Testing by Docking of Agonists and Tricyclic Antagonists. *J. Med. Chem.* **1994**, *37*, 2874-2888.
28. Trumpp-Kallmeyer, S.; Hoflack, J.; Bruinvels, A.; Hibert, M. Modeling of G-protein-Coupled Receptors: Application to Dopamine, Adrenaline, Serotonin, Acetylcholine, and Mammalian Opsin Receptors. *J. Med. Chem.* **1992**, *35*, 3448-3462.
29. Dijkstra, D.; Mulder, T.B.; Rollema, H.; Tepper, P.G.; Van der Weide, J.; Horn, A.S. Synthesis and Pharmacology of *trans*-4-n-propyl-3,4,4a,10b-tetrahydro-2*H*,5*H*-1-benzopyrano[4,3-b]-1,4-oxazin-7- and -9-ols: the Significance of Nitrogen pK_a Values for Central Dopamine Receptor Activation. *J. Med. Chem.* **1988**, *31*, 2178-2182.
30. Sheppard, W.A. The Effect on Fluorine Substitution on the Electronic Properties of Alkoxy, Alkylthio and Alkylsulfonyl Groups. *J. Am. Chem. Soc.* **1962**, *85*, 1314-1318.
31. Sonesson, C.; Boije, M.; Svensson, K.; Ekman, A.; Carlsson, A.; Romero, A.G.; Martin, I.J.; Duncan, J.N.; King, L.J.; Wikström, H. Orally Active Central Dopamine and Serotonin Receptor Ligands: 5-, 6-, 7-, and 8-[[Trifluoromethyl)sulfonyl]oxy]-2-(di-n-propylamino)tetralins and the Formation of Active Metabolites *in vivo*. *J. Med. Chem.* **1993**, *36*, 3409-3416.
32. Gerding, T.K.; Drenth, B.F.; de Zeeuw, R.A.; Tepper, P.G.; Horn, A.S. The Metabolic Fate of the Dopamine Agonist 2-(*N*-Propyl-*N*-2-thienylethylamino)-5-hydroxytetralin in Rats after Intravenous and Oral Administration. II. Isolation and Identification of Metabolites. *Xenobiotica.* **1990**, *20*, 525-536.
33. Clark, M.; Cramer III, R.D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comp. Chem.* **1989**, *10*, 982-1012.
34. Pugsley, T.A.; Davis, M.; Akunne, H.C.; MacKenzie, R.G.; Shih, Y.H.; Damsma, G.; Wikström, H.; Whetzel, S.Z.; Georgic, L.M.; Cooke, L.W.; DeMattos, S.B.; Corbin, A.E.; Glase, S.; Wise, L.D.; Dijkstra, D.; Heffner, T.G. Neurochemical and Functional Characterization of the Preferentially Selective Dopamine D_3 Agonist PD 128907. *J. Pharmacol. Exp. Ther.* **1995**, *275*, 1355-1366.
35. Shih, Y.H.; Chung, F.-Z.; Pugsley, T.A. Cloning, Expression and Characterization of a Human Dopamine $D_{4.2}$ Receptor (CHO K1 cells) and Various $D_{4.2}/D_{2L}$ Chimeras (COS-7 cells). *Prog. Neuro-Psychopharmacol. & Biol. Psychiat.* **1997**, *21*, 153-167.
36. MacKenzie, R.G.; Van Leeuwen, D.; Pugsley, T.A.; Shih, Y.H.; DeMattos, S.B.; Tang, L.; Todd, R.D.; O'Malley, K.L. Characterization of Human D_3 Receptor Expressed in Transfected Cell Lines. *Eur. J. Pharmacol.* **1993**, *266*, 785-789.
37. Cheng, Y. and Prusoff, W.H. Relationship Between the Inhibition Constant (K_i) and the Concentration of Inhibitor which Causes 50 Percent Inhibition (IC_{50}) of an Enzymatic Reaction. *Biochem. Pharm.* **1973**, *22*, 3099-3108.

A GRID/GOLPE 3D QSAR Study on a Set of Benzamides and Naphthamides

4

This chapter is based on the article: Nilsson, J.; Wikström, H.; Smilde, A. K..; Glase, S.; Pugsley, T.; Cruciani, G.; Pastor, M.; Clementi, S. A GRID/GOLPE 3D QSAR Study on a Set of Benzamides and Naphthamides, with Affinity for the Dopamine D_3 Receptor Subtype. *J. Med. Chem*. **1997**, 40, 833-840

Summary

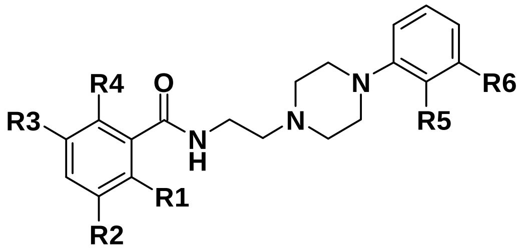
In the pursuit of drugs effective in the treatment of schizophrenia without extrapyramidal side-effects, compounds that selectively block the dopamine D_3 receptor are thought to be of interest. In order to create a model with which the D_3 affinity, of compounds not yet synthesised, can be predicted, a Comparative Molecular Field Analysis (CoMFA) was performed. The data set consisted of 30 compounds which were described quantitatively with three different probes in the GRID program. The multivariate statistical analyses were performed using the GOLPE program. The predictive ability of the model was found to increase significantly when the number of variables was reduced from 25110 to 784. A crossvalidated Q^2 of 0.65 was obtained with the final model, confirming the predictability of the model.

It was concluded that naphthamides in general were more potent than benzamides, and a compound with an unsubstituted arylpiperazine phenyl ring was less potent than a substituted one, in this series of compounds.

4.1 Introduction

The dopamine D_3 receptor is characterised by its selective expression in mesolimbic dopaminergic projection areas (see Figure 1.2) of the rat and human brains and its high affinity for antipsychotic drugs, suggesting a role of this receptor in the control of locomotion and motivation, as well as in the pathogenesis of disorders such as drug abuse and schizophrenia.[1] Existing drugs against schizophrenia cause major movement disorders called extrapyramidal syndrome (EPS), proposed to be caused by blockade of D_2 receptors in striatum. The 30 ligands[2] included in this study (Tables 4.1–4.3) were synthesised with the aim to achieve ligands that selectively could antagonise the D_3 receptor, and thereby also avoid the EPS. The ligands belong to two structural different classes, benzamides and naphthamides, both having an arylpiperazine tail connected to the amide nitrogen. The objective with this study was to create a 3D QSAR model able to predict the activities of compounds not yet synthesised and which could serve as an aid in the design of new compounds. Traditionally, 3D QSAR models are created using CoMFA as implemented in the SYBYL program.[3,4] Today other methods are available and it was decided to use the GRID program[5] for the generation of molecular descriptors and the GOLPE program[6] for the multivariate statistical analysis.

Table 4.1 The benzamide ligands and their experimental and fitted affinities for the dopamine D_3 receptor subtype.



Compd	R1	R2	R3	R4	R5	R6	exp.[a,b]	fitted
1	-OMe		-Br				2.5	2.6
2	-OMe		-Br	-OH			2.6	2.8
3	-OMe	-Et	-Cl	-OH			3.2	3.1
4	-OMe	-Cl	-Cl	-OH			3.0	2.9
5	-OMe	-Cl	-Cl	-OH	-Cl	-Cl	2.7	2.7

[a] [^{3}H]Spiperone, human DA D_3 receptors expressed in CHO K1 cells; K_i values were obtained from four to six concentrations, run in triplicate, by a non linear regression analysis.; [b] $\log_{10}$ of K_i(nM) values

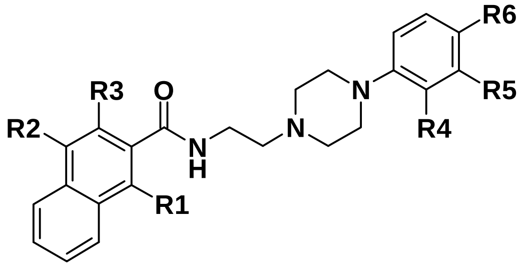
4.2 Molecular Descriptors Generated in the GRID program

In both SYBYL/CoMFA[3,4] and GRID,[5] as was mentioned in Chapter 1, a 3D QSAR model utilises a grid, large enough to enclose all the aligned ligands. In each grid point interactions between a probe atom and the target molecules are calculated. The programs SYBYL/CoMFA[3,4] and GRID[5] use different force fields, different types of probe atoms and the interactions are calculated differently. Interactions accounted for in the GRID force field are steric, electrostatic and hydrogen bonding interactions represented by the Lennard-Jones energy (E_{ste}), the Coloumbic energy (E_{ele}) and a hydrogen bonding (E_{hb}) term, respectively. In contrast to SYBYL/CoMFA where the interaction energies, *i.e.* E_{ste} and E_{ele}, are considered separately, the sum of all the different interaction energies (Equation 4.1) is calculated in each grid point with GRID. An attractive interaction between the probe atom and the ligand produces a negative field (E_{tot}) while a repulsive interaction is positive.

$$E_{tot} = E_{ele} + E_{ste} + E_{hb} \tag{4.1}$$

Different probes reflect different types of interactions and may selectively be included to mimic specific interactions between the ligand and the receptor.[7] Often more than one probe is necessary for a complete description of the different interaction types.

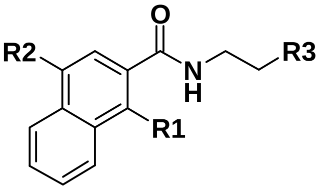
Table 4.2 The naphthamide ligands and their experimental and fitted affinities for the dopamine D_3 receptor subtype.

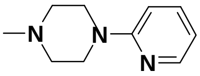
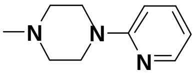
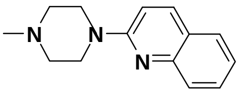
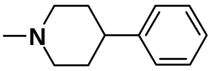
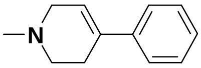


Compd	R1	R2	R3	R4	R5	R6	exp.[a,b]	fitted
6	-OMe	-Br					1.4	1.9
7	-OMe	-Cl					1.7	1.9
8	-OEt	-Br					2.4	2.0
9	-OMe	-Br		-OMe			1.9	1.7
10	-OEt	-Br		-OMe			1.9	1.8
11		-Br	-OMe				2.5	2.3
12	-OMe	-Br			$-CF_3$		2.4	2.2
13	-OEt	-Br			$-CF_3$		2.1	2.3
14	-OMe	-Br		-CN			1.7	1.5
15	-OMe	-Br		-Me			1.8	1.6
16	-OMe	-Br			-Me		2.5	2.2
17	-OMe	-Br				-Me	2.4	2.3
18	-OMe	-Br		-Me	-Me		1.7	1.7
19	-OMe	-Br		-Cl			1.6	1.5
20	-OMe	-Br			-Cl		1.6	2.0
21	-OMe	-Br			-Cl	-Cl	2.3	2.4
22	-OMe	-Br		-Cl		-Cl	2.0	2.0
23	-OMe	-Cl		-Cl	-Cl		1.6	1.7
24	-OMe	-Br		-F		-F	2.3	2.0
25	-OMe	-Br		-Me	-Cl		1.7	1.7

[a] [^{3}H]Spiperone, human DA D_3 receptors expressed in CHO K1 cells; K_i values were obtained from four to six concentrations, run in triplicate, by a non linear regression analysis.; [b] $\log_{10}$ of K_i(nM) values

Table 4.3 Naphthamide ligands with and their experimental and fitted affinities for the dopamine D_3 receptor subtype.



Compd	R1	R2	R3	exp.[a,b]	fitted
26	-OMe	-Br		1.9	1.9
27	-OEt	-Br		1.9	2.0
28	-OMe	-Br		2.4	2.5
29	-OMe	-Br		1.9	1.8
30	-OMe	-Br		1.3	1.7

[a] [^{3}H]Spiperone, human DA D_3 receptors expressed in CHO K1 cells; K_i values were obtained from four to six concentrations, run in triplicate, by a non linear regression analysis.; [b] $\log_{10}$ of K_i(nM) values

4.3 Data Pretreatment

An important and sometimes crucial step in Comparative Molecular Field Analysis (CoMFA)[4,8] is the pretreatment of the descriptor matrices. In Chapter 2 it was suggested that auto-scaling is appropriate if the data set comprise several different types of descriptors, *e.g.* physicochemical descriptors. In a 3D QSAR data set with descriptors from GRID,[5] all variables are measured similarly and auto-scaling is not suitable anymore. However, the GOLPE[6] program offers several effective pretreatment options:

1) A probe very close to the target molecule may produce unrealistically high positive (repulsive) interactions, caused mainly by the Lennard-Jones (E_{ste}) contribution to E_{tot} (Equation 4.1), that may influence the PLS solution detrimentally. Therefore, it is wise to introduce a positive maximum cut-off value. The negative interaction values (attractions) decline smoothly as the distance between the probe and the target molecule increases and a cut-off value for negative values is not necessary.

2) Interaction values in grid points located in the periphery of the grid tend to be low with variation more similar to noise than true reflections of the variations in the field. The zeroing option corrects for this by replacing absolute values, lower than a specified cut-off value, with zero.

3) Grid points with a too low standard deviation may be assumed insignificant and consequently be omitted from the analysis. This option is identical to the "MINIMUM_SIGMA" option in SYBYL/CoMFA.[3]

4) In some grid points it is possible that all ligands but one, have identical interaction values (*e.g.* maximum cut-off) and, as a consequence, PLS[9] adjust its solution for this single variable.[6] These types of variables called 2-level variables may produce spurious PLS solutions and can optionally be omitted with hopefully a more stable PLS model as a result. Similarly, also 3- and 4-level variables can be omitted.

5) In addition to traditional auto-scaling, block-scaling is also possible. In auto-scaling, the sum of squares within each variable (column) is normalised whereas in block-scaling, the sum of squares of whole grids are normalised. This is identical to the "CoMFA_std" scaling option in SYBYL/CoMFA.

4.4 Statistical Tools

A typical 3D QSAR data set comprise a two-way matrix with a lot more columns than rows, which disqualifies MLR[9] (see Chapter 2) as the regression method. It is also reasonable to assume that the variables are collinear. Therefore, PLS[9,10] normally is utilised for the analysis of this type of data.

The part of **X,** not accounted for by the PLS model, is assumed to consists only of insignificant variation. Consequently, a PLS model generally has better predictability but explains less of the variance in **y** as compared with for example a MLR model, where 100 % of **y** is explained. As a measure of the explained variation in **y** the fit, *i.e.* the R^2 is used:

$$R^2 = 1 - \sum_{i=1}^{I} \left(y_i - \hat{y}_i \right)^2 \Big/ \mathrm{SSY} \tag{4.2}$$

The R^2 gives the fraction of the total variation in **y** accounted for by the model, where y_i is the biological activity of the *i*th compound measured, $\hat{y}_i$ the model estimation of y_i and SSY the sum of squares of **y**.

In 3D QSAR, models with high predictability are desired. Crossvalidation[9-11] is used as an internal measurement of the predictability. With crossvalidation, a model is calculated with a group of objects omitted which subsequently are predicted with the reduced model. This is repeated until all objects have been omitted once. The predictability is quantified with the crossvalidated Q^2:

$$Q^2 = 1 - \sum_{i=1}^{I} \left(y_i - \hat{y}_{(i)} \right)^2 \Big/ \mathrm{SSY} \tag{4.3}$$

The predicted y in Equation 4.3 is denoted $\hat{y}_{(i)}$, *i.e.*, a prediction of y_i using a model with the *i*th object omitted and SSY is the sum of squares of **y**. Models with high predictability have a crossvalidated Q^2 close to one, while models with low or negative Q^2 will predict no better than random. Other methods of quantifying the predictability, *e.g.* SDEP[12,13] have been suggested.

4.5 D-Optimal Preselection of Variables

D-optimal[14,15] preselection of variables and variable selection guided by a fractional factorial design (FFD),[16] form the basis of the GOLPE (Generating of Optimal Linear PLS Estimations)[17] variable selection procedure.

The data generated in GRID, and 3D QSAR in general, contains a large number of variables where only a fraction of them contain information correlated with the biological activity. In GOLPE, only the most informative variables are selected by a D-optimal preselection in the weight space $\mathbf{W}$ $[(\mathbf{w}_1|\mathbf{w}_2|\ldots|\mathbf{w}_A)$; A is the number of PLS components$]$ from an initial PLS model. The dimensionality (A) of the PLS model is determined by an initial leave-one-out crossvalidation[9] experiment of the complete data set. The selection procedure is iterative and not more than 50 % of the variables should be omitted each time.[17] Each iteration begins by establishing a new PLS model including only the previously selected variables and is repeated until the R^2 starts to decrease. (During the preparation of this work, the D-optimal preselection procedure was used to reduce the number of variables from roughly tens of thousands to thousands of variables before the R^2 started to decrease.)

4.6 Variable Selection Following a Fractional Factorial Design (FFD) Procedure

At this point, most of the redundant variables have been eliminated and the predictability of the model can be optimised. The influence of each variable on the predictability is estimated by a number of crossvalidation experiments where variables are included and excluded, alternately. A design matrix,[17] with the number of columns equal to the number of variables left after the D-optimal preselection and with two times as many rows, is created. Each row represents an experiment were "plus" and "minus" signs mean include and exclude a variable in the experiment, respectively. Obviously, models with different combinations of variables have different predictability and, by means of Yates' algorithm,[16] the influence of each individual variable on the predictability can be estimated. In order to separate a variable that significantly improves predictability from one that does not, a number of dummy variables are introduced in the design matrix. A dummy variable must, by definition, have no influence on the predictability of the model. Therefore, the estimated average effect of the dummy variables may serve as a limit, on the basis of a Student t-tailoring at the 95 % confidence level, for the estimated effects of the true variables. A true variable with significantly higher estimated effect than the limit will be excluded. A true variable with significantly lower estimated effect than the limit will be kept fixed and a true variable with an estimated effect within the limit interval may optionally be fixed or excluded.

4.7 Results and Discussion

Molecular Modelling

Initially, 30 molecules (Tables 4.1–4.3) were built and minimised using the MM2* force field, as implemented in the molecular modelling package Macromodel 4.5.[18] In order to simplify the following calculations, two general assumptions concerning the conformations were made. First, the benzamide part of the ligands was fixed in a planar conformation (τ(5-6-7-8) in Figure 4.1), as supported from several X-ray structures present in the Cambridge Crystallographic Structural Database (CCSD). Additionally, the oxygen from the ortho-methoxy (Figure 4.1, atom 5) stabilises the conjugated benzamide part by forming an internal hydrogen bond with the amide *N-H*.

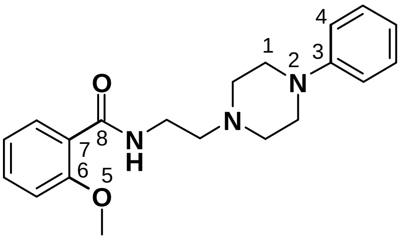


Figure 4.1 Benzamide-phenylpiperazine skeleton. The highlighted torsional angles τ(1-2-3-4) and τ(5-6-7-8) were fixed at 84° and 0°, respectively, with support form X-ray structures.

The second assumption concerns the arylpiperazine tail, which conformation also was fixed with support from X-ray structures. The piperazine ring in a chair conformation was present in all arylpiperazines found in CCSD, suggesting that this conformation is energetically preferable, as compared to the boat conformation. A conformational search experiment in SYBYL (Tripos force field[19]) confirmed this, by finding the chair conformation in all low energy conformations (not presented). The torsional angle between the piperazine ring and the phenyl ring (τ(1-2-3-4)) was more flexible. An unsubstituted phenyl ring, can according to the X-ray structures, be found in almost any angle, while an ortho substituted phenyl ring is more or less always somewhat twisted. Due to these findings the torsional angle between the piperazine ring and the phenyl ring, *e.g.* τ(1-2-3-4), was fixed at 84°. As a consequence, the overlap of the phenyl rings from compound **30** (Table 4.3) and the rest of the ligands was improved.

The conformational space of all ligands was investigated by conformational searches using the Monte-Carlo procedure, as implemented in Macromodel 4.5, while keeping the above discussed torsional angles fixed. New conformations were randomly generated, and their energy were minimised and subsequently compared with the previously saved conformations. If a new conformation was lower in energy than the present "global minimum energy conformation" the new conformation replaced the old one. If a new conformation was identical with the "global minimum energy conformation" a variable was increased by one, and when this conformation was found a sufficient number of times the conformational search was considered converged. All conformations within 20 kJ/mol from the "global minimum energy conformation" were saved and considered as even likely to be the conformation interacting with the receptor.

The most crucial step when preparing a CoMFA study is the alignment of the ligands. Since the homology between the ligands in the model is very high the main goal with the alignment procedure was to achieve maximum overlap between the ligands. The pharmacophore below, was chosen with this in mind. The midpoint of the aromatic benzamide ring, the amide *O*, amide *N-H* and a dummy atom in the direction of the lone-pair of electrons from the basic nitrogen of the arylpiperazine part were identified as possible interaction points with the receptor. Dummy vectors and midpoints were added to each conformation from all ligands using vecadd, a sub-program in the pharmacophore program Apollo.[20] Due to lack of a rigid template the "global minimum energy conformation" from ligand **1** (Table 4.1) was used as a template. Subsequently, all the other ligands were fitted on this template, using the pharmacophore points identified above. Each fitting procedure (performed in Apollo) considered all conformations (from each ligand) within 20 kJ/mol from the global minimum as determined by the conformational search procedure. The output was a number of fits ranked in decreasing order of RMS (Root Mean Square). Optionally, the energy of a fitted conformation can be selected to affect the ranking and consequently, the lower the energy of a fitting conformation the higher the rank of the fit becomes. The highest ranked conformation, from each ligand, was included in the final model.

Finally, the 30 conformations selected were converted into the Tripos mol2 format using the file converting program Babel.[21]

Receptor Binding

The *in vitro* affinity of ligands for the human dopamine D_3 receptor subtype was the dependent variable considered in this study. In the antagonist binding study, the affinity of the compounds was determined by their ability to displace [^3H]-spiperone from the dopamine D_3 receptor.[22] Receptor binding affinities are mostly expressed as K_i values calculated from IC_{50} values, as described by Cheng and Prusoff.[23] In order to get a more homogenous distribution of the dependent variable the $\log_{10}$ of the K_i (nM) was used in this analysis. Consequently a compound with high affinity has lower $\log_{10}(K_i)$ than a compound with low affinity.

Probe selection

The grid was designed in the GRID program and was large enough to enclose the aligned ligands with 4 Å in all directions with a resolution of 1 Å. Each grid consisted of 8370 grid points and interactions between the eight probes in Table 4.4 and all 30 ligands were calculated in each grid point, forming a matrix with 240 rows and 8370 columns.

Table 4.4 Description of the eight different probes from Grid. The three selected probes are marked with bold face characters.

probe	description	probe	description
H	hydrogen atom	OH	OH with acidic H
C3	**sp3 C atom**	**OH2**	**water**
O::	sp2 O in C=O	NA+	sodium cation
O1	sp3 O in O-H	**CA+2**	**calcium cation**

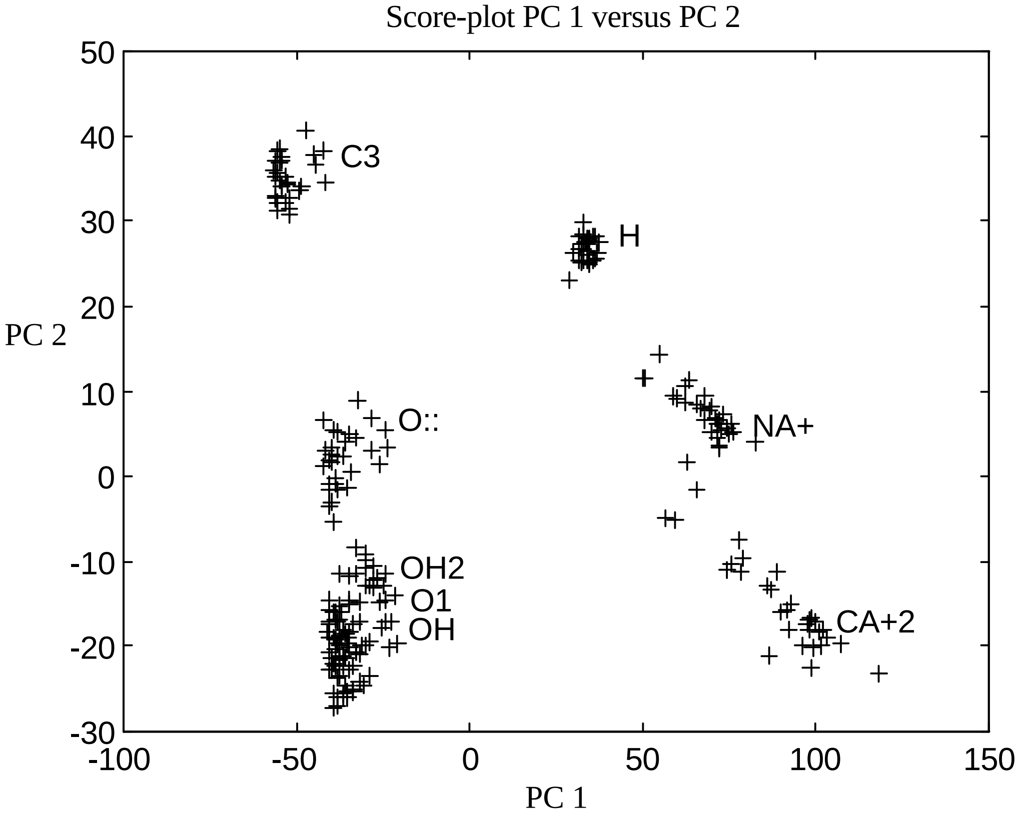


Figure 4.2 Plot of the first two Principal Components explaining 64 % of the variation. Six clusters of probes are clearly identified.

Subsequently, a principal component analysis (PCA, see Chapter 2)[6,9,24] was performed, where the first two components described 64 % of the variation. In Figure 4.2, six clusters were identified from the eight different probes and it can be concluded that the CA+2, the C3 and the OH2 probes contained the most diverse information. Interestingly, the O1, the OH and the OH2 probes were not possible to separate in the score-plot, indicating that no extra information would be added to the model if more than one of them were included. Therefore, only the OH2 probe together with the C3 and the CA+2 probes were selected for the final analysis.

Hydrogen bonding is one of the more important interactions in the ligand-receptor interaction.[5,7] Figure 4.3(a) represents the OH2 interaction energies contoured at the -2.6 kcal/mol level for the template molecule (**1**). It is clear that **1** can interact as a hydrogen bond acceptor with the receptor at both basic nitrogens from the piperazine moiety. The CA+2 probe mimics the electrostatic interactions and in Figure 4.3(b) it is clear that significant negative electrostatic fields are generated around the benzamide part of the template molecule (**1**) on the -3 kcal/mol level. Finally, the field generated by the C3 probe (Figure 4.3(c)) on the 0.02 kcal/mol level indicates the smallest distance a non charged molecule may approach the template without causing repulsive interactions. The energy cut-offs used above, are of no importance since they were chosen in order to optimally describe the specific differences between the three molecular fields.

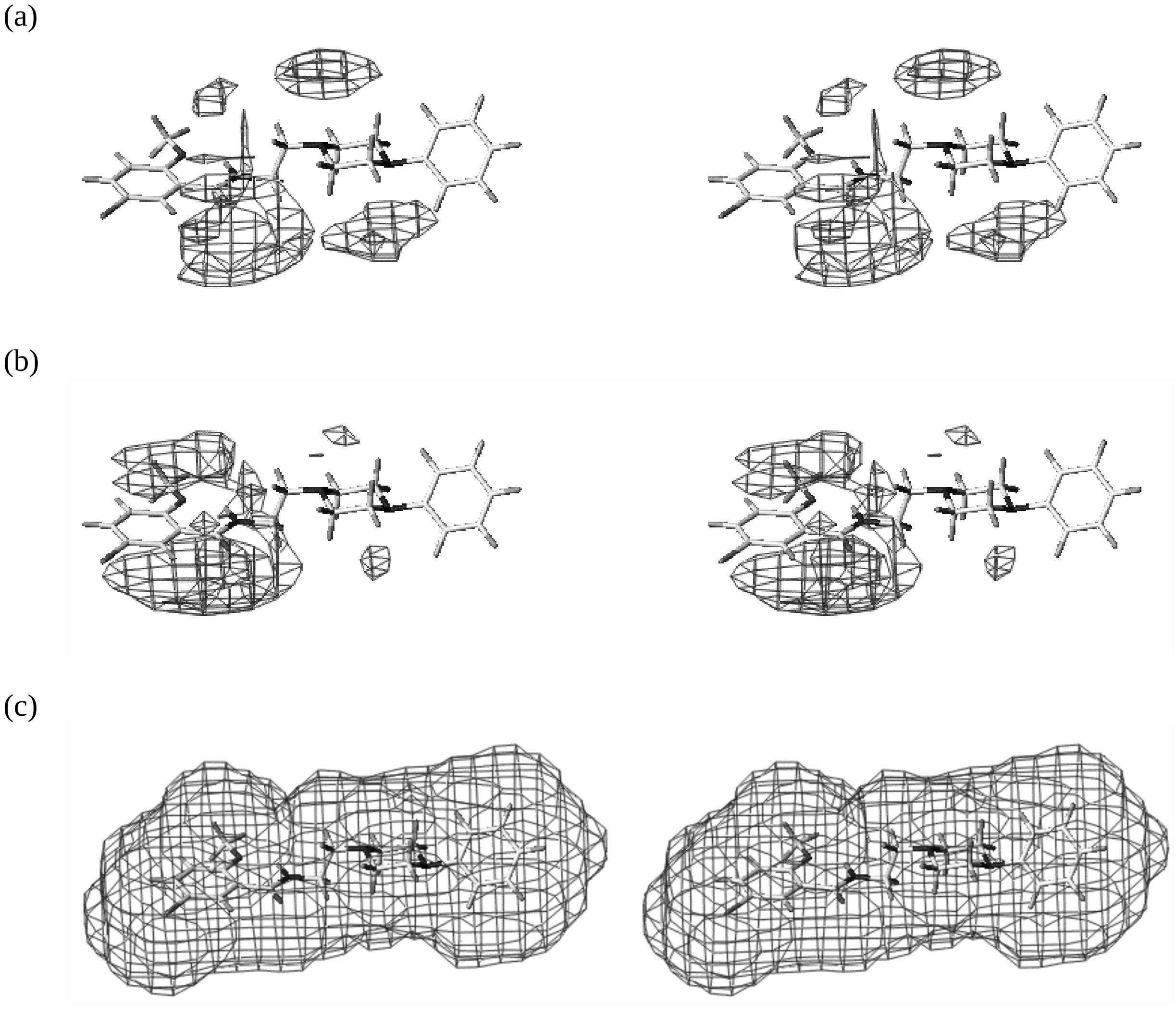


Figure 4.3 The actual fields from (a) the OH2 probe, (b) the CA+2 probe and (c) the C3 probe surroundings of the template molecule (**1**) at the -2.6 kcal/mol, -3.0 kcal/mol and 0.02 kcal/mol levels, respectively.

In the final data set (Figure 4.4) each ligand is represented by its interactions with the three selected probes unfolded to form a row, leaving a matrix **X** (30 × 25110).

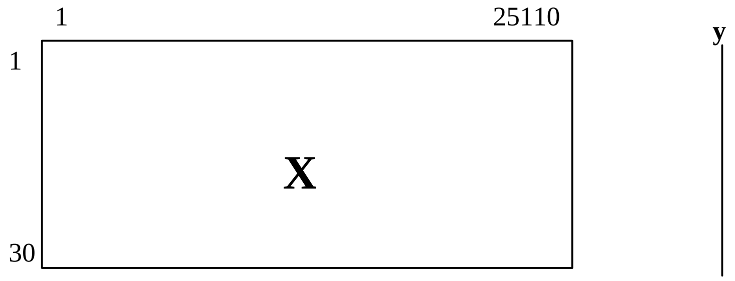


Figure 4.4 The final data set (**X**) consisting of 30 compounds and 25110 variables (30 × 25110) with one dependent variable **y** (30 × 1).

Important for the following analysis is the fact that the standard deviation in grid points close to substituent groups present in all ligands, are very low (Figure 4.5(a)–(c)). For instance, the strong hydrogen bonding field present around the amide group of **1** in Figure 4.3(a), is present in all ligands and consequently, the standard deviation in grid points located in this region become low (Figure 4.5(a)). Since PLS focuses on variables (grid points) with high standard deviation it may therefore turn out that these regions, most certainly involved in the ligand-receptor interaction, have low weights in the final PLS solution.

: Variable Pretreatment

Our data set consists of 25110 variables, 8370 from each of the three probes OH2, C3 and CA+2. The positive maximum cut-off was set to 5 kcal/mol already during the generation of the descriptors in the GRID program. In order to correct for round off errors,[6] GOLPE automatically rejects columns (variables) having a total sum of square (SS) lower than 10^{-7}. Absolute values lower than 0.01 kcal/mol were set to zero and as a consequence another 2000 variables were omitted from further modelling due to a too low variation ($SS < 10^{-7}$). By introducing a lower standard deviation limit for the columns the number of variables may be reduced significantly and render the absolute value cut-off almost useless.[17] This action was considered as a variable selection method in itself and was not utilised here. Further, all 2, 3 and 4-level variables were removed followed by a block-scaling procedure, as described above.

The pretreatment procedure reduced the number of active variables from 25110 to 19180.

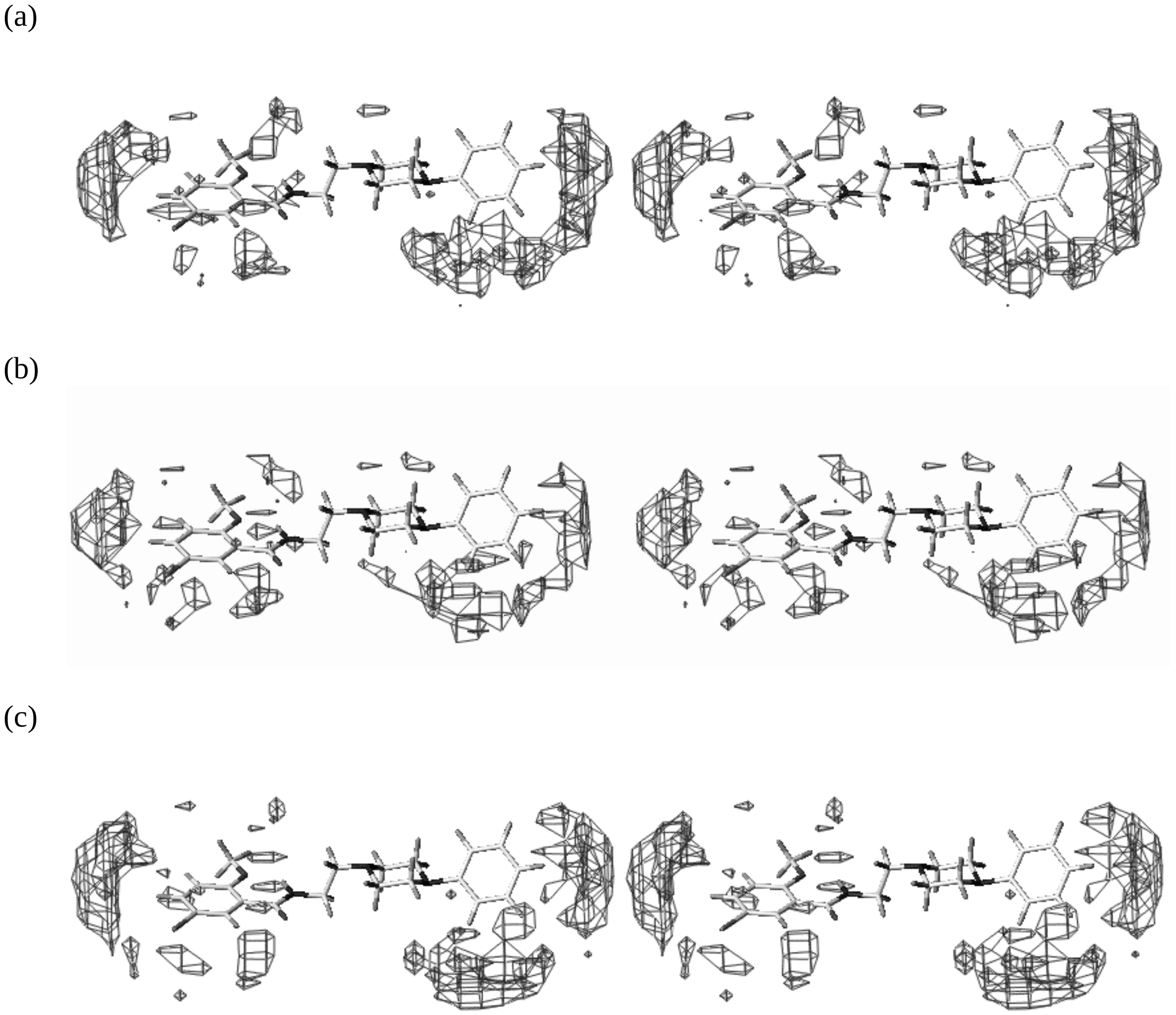


Figure 4.5 Contour maps of the standard deviations, all 30 compounds considered, from (a) the OH2 probe, (b) the CA+2 probe and the C3 probe. For clarity, only standard deviations larger than unity are depicted.

D-optimal Variable Preselection

The most informative variables span the weight space from an initial PLS model assuming that a sufficient number of components are considered. If too few components are considered, information may be lost due to the fact that significant variables may not have been selected. If too many components are considered, variables not correlated with the biological activity may be selected and introduce random variation in the model. At this early stage of modelling, the only interest is to make sure that a sufficient number of components is considered, not the predictability. A leave-one-out crossvalidation experiment gave the highest crossvalidated Q^2 (0.45) after two components (Table 4.5). Consequently, it was decided to perform the D-optimal preselection procedure using three components, assuming three would be enough to capture all the significant information in the data.

In the iterative process, 50 % of the variables were selected each time. Each iteration started by calculation of a new PLS model, including only the previously selected variables. The selection procedure was repeated four times before the R^2 started to decrease, reducing the number of variables from 19180, 9543, 4771, 2385 to 1192.

Table 4.5 The impact of pretreatment, D-optimal variable preselection and FFD variable selection on the fitted R^2 and the crossvalidated Q^2.

	# of variables	# lv[a]	R^2	Q^2
after pretreatment	19180	2	0.76	0.45
after D-opt. selection	1192	3	0.85	0.49
after FFD[b] selection	784	2	0.80	0.65

[a] The number of components determined with leave-one-out crossvalidation.
[b] After the Fractional Factorial Design selection.

Variable selection

The matrix, containing only the 1192 variables left after the D-optimal preselection procedure was used as input for the final step in GOLPE. A design matrix with 4096 number of experiments (rows) and 299 number of dummy variables was created and each experiment was validated with crossvalidation (five random groups repeated 20 times). The Fractional Factorial Design procedure resulted in 313 and 408 variables with significant positive and significant negative effect on the predictability, respectively. Accordingly, 408 variables were omitted and 313 were maintained together with 471 variables with non-significant effect on the predictability, leaving 784 variables for the final model. Variables from a 3D QSAR study differ from classic variables in the sense that each variable represents a definite spatial coordinate in the grid. It can be shown[25] that the variables selected are located in regions where the standard deviations were high in Figure 4.5.

Since an external test set (see Chapter 5) not was utilised, the final model was validated with crossvalidation in three different ways (Table 4.6): First, with the leave-one-out procedure, followed by leave-two-out and finally, groups of five were left out and repeated 20 times.[24] The last validation experiment is a mixture between crossvalidation[9] and bootstrapping[26,27] in the sense that each crossvalidation experiment was repeated a number of times, as in bootstrapping, and the objects where included only once, as in crossvalidation.

Table 4.6 Different crossvalidation experiments performed with the final model. A more elaborated description over the different experiments is provided in the text.

experiment	# lv[a]	Q^2
leave-one-out	2	0.65
leave-two-out	2	0.65
5-random groups[b]	2	0.63

[a] The number of components with maximum Q^2
[b] Average from 20 experiments

Two components were sufficient to explain most of the variation in **y** ($R^2 = 0.80$), and the experimental $\log_{10}(K_i)$ and the fitted $\log_{10}(K_i)$ is plotted in Figure 4.6(a).

The GOLPE variable selection procedure has indeed improved the predictability of the model by increasing the crossvalidated Q^2 from 0.45 to 0.65 (Table 4.5). A crossvalidated Q^2 of 0.65 may in 3D QSAR be considered sufficient and in Figure 4.6(b) the experimental $\log_{10}(K_i)$ is plotted against the predicted $\log_{10}(K_i)$.

In order to be able interpret the PLS model, the PLS coefficients ($\mathbf{b}_{\mathrm{PLS}}$; Chapter 2) after the second PLS component were plotted as contour plots connecting grid points with similar values (Figures 4.7(a)–(c)). The negative and positive coefficients are represented by dark and light grey iso-contours, respectively. In order to simplify the interpretation of Figure 4.7 Equation 4.4 may be of good help, where b_r is the $\mathrm{b}_{\mathrm{PLS}}$ coefficient in the rth grid point, x_r is the actual field in the rth grid point, $\hat{y}$ the estimation of y and R the number of grid points. The $\mathbf{b}_{\mathrm{PLS}}$ ($= b_1,\ldots,b_R$) coefficients are plotted in Figures 4.7(a)–(c) for the OH2 probe, the CA+2 probe and the C3 probe, respectively.

$$\hat{y} = b_1x_1 + b_2x_2 + \ldots + b_Rx_R \tag{4.4}$$

Basically, the $\mathbf{b}_{\mathrm{PLS}}$ coefficients in Equation 4.4 are needed for predictions of the biological activity $\hat{\mathbf{y}}$ of new molecules, but since the sizes and signs of the coefficients reveal the relative influence of each grid point on $\mathbf{y}$, they are also suitable for the interpretation. That is, a new compound with a substituent protruding into a region with positive b_r will produce a positive (repulsive) field x_r in this region and consequently $b_r \times x_r$ is also positive, indicating a negative influence on $\hat{y}$. (A high $\log_{10}(K_i)$ corresponds to low affinity.) The opposite is valid if the region has negative b_rs.

In the position para to the amide group of the benzamide moiety there is space for substituents (Figure 4.7(c), arrow I) unable to form hydrogen bonds (Figure 4.7(a), arrow II) and that produce repulsive interactions with the CA+2 probe (Figure 4.7(b), arrow III). This is the case with the second phenyl ring present in the naphthamide moiety. This series generally has lower K_i than the benzamide ligands.

In one of the ortho positions (Figure 4.7(c), arrow IV), a methoxy group is necessary in order to make an internal hydrogen bond interaction possible and fix the benzamide and the naphthamide moieties in planar conformations. Accordingly, it can also be concluded from the model that there is no room for substituents (Figure 4.7(c), arrow V) in the second ortho position.

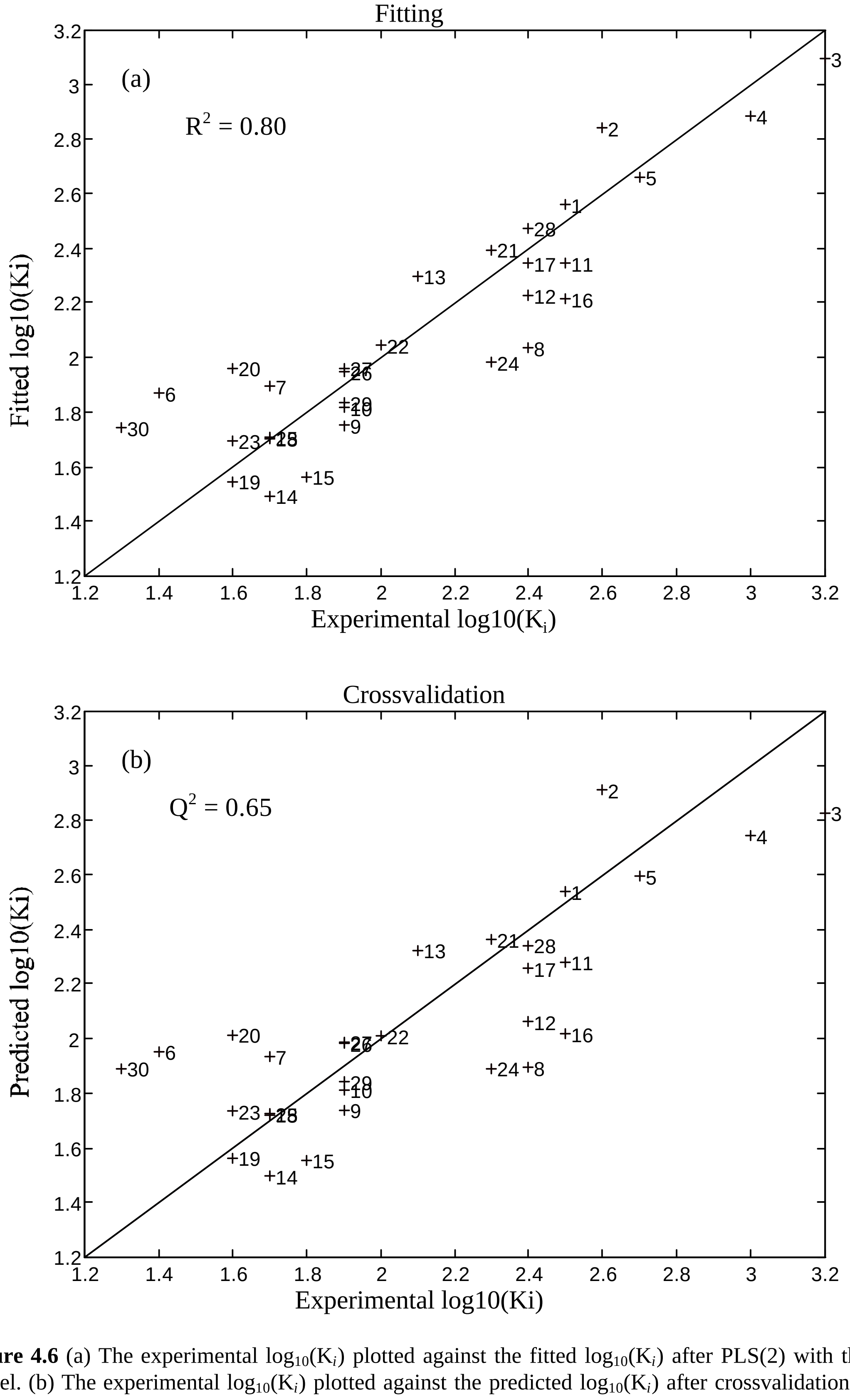


Figure 4.6 (a) The experimental $\log_{10}(K_i)$ plotted against the fitted $\log_{10}(K_i)$ after PLS(2) with the final model. (b) The experimental $\log_{10}(K_i)$ plotted against the predicted $\log_{10}(K_i)$ after crossvalidation (LOO) with PLS(2) with the final optimal PLS model.

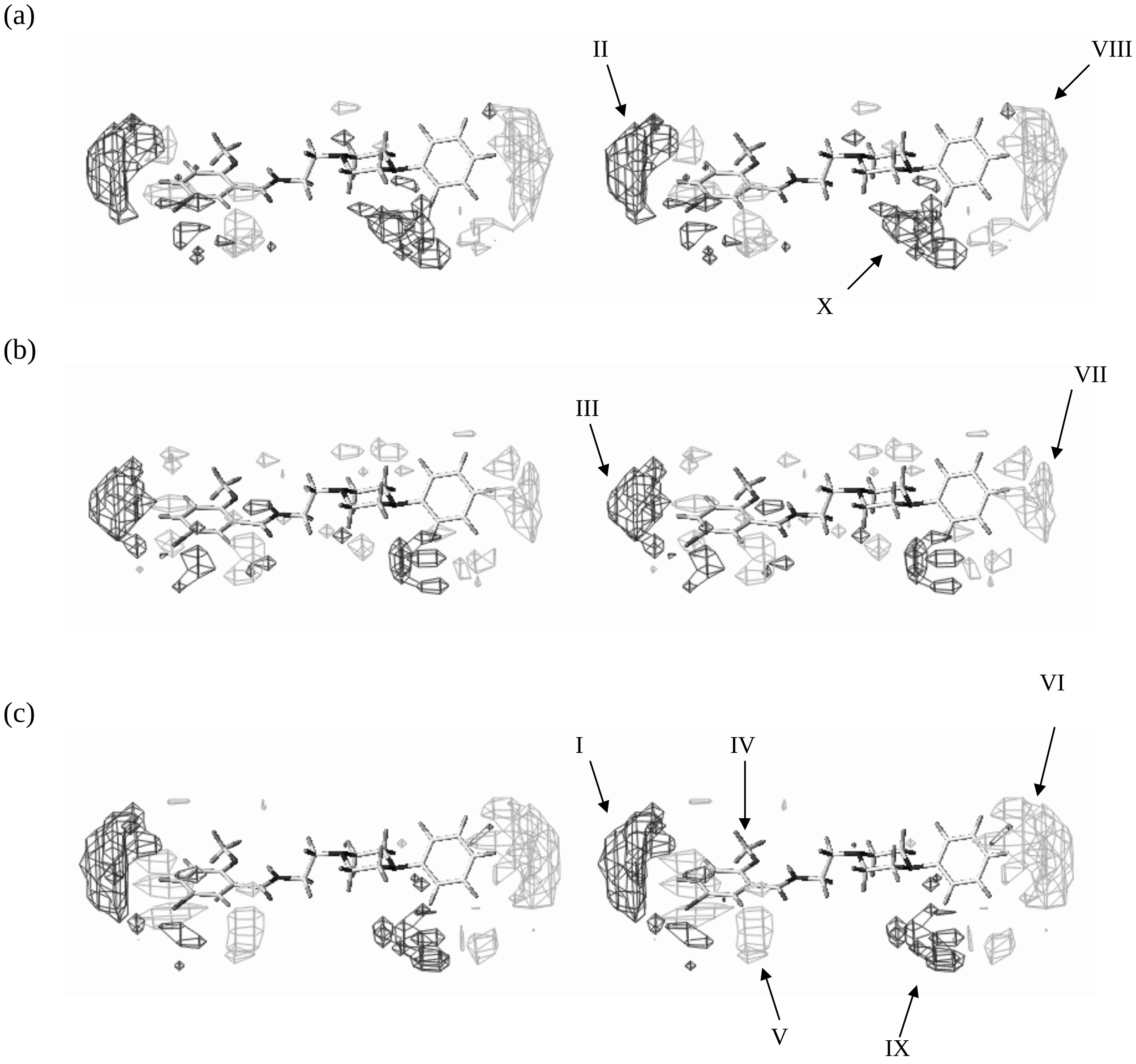


Figure 4.7 Contour plots of the PLS coefficients ($\mathbf{b}_{PLS}$) after the second PLS component: (a) from the OH2 probe, (b) from the CA+2 probe and (c) from the C3 probe. Only coefficients larger than |0.0005| are shown for clarity. Negative and positive iso-contours are represented by dark and light grey tones, respectively.

The phenyl ring in the phenylpiperazine tail may not be substituted in the para position (Figure 4.7(c), arrow VI) due to steric reasons, however, an attractive interaction with the CA+2 probe (Figure 4.7(b), arrow VII) promotes binding. Therefore, a small substituent with a negative electrostatic potential and with the ability to form hydrogen bonds[28] (Figure 4.7(a), arrow VIII) like a fluorine atom, may be appropriate. The ortho position on the phenylpiperazine phenyl (Figure 4.7(c), arrow IX) may very well be substituted, but substituents able to form hydrogen bonds will not improve binding (Figure 4.7(a), arrow X). Additionally, a ligand with an unsubstituted phenylpiperazine moiety is less potent than a substituted one. A speculative explanation for this could be that an unsubstituted ligand in solution more often has a planar phenylpiperazine tail than what could be expected from X-ray structures. This, in turn, could sterically hinder the ligand to interact with the receptor where a more twisted conformation may be preferable, which always is the most likely conformation for an ortho substituted ligand.

Finally, as explained above, the hydrogen bonding properties of the amide part and the basic nitrogen of the arylpiperazine tail are important and should definitely be taken into account if one, for instance, want to create a mini-receptor model[29] with this study as a reference. Several arguments against this are possible, but it is especially stressed that a CoMFA model measures the differences between the ligands in the training set. Therefore, the alignment used in the model does not necessary need to fit into an active site of a receptor model. Hence, a successful alignment is achieved when the differences between the molecular fields of the ligands are reflected optimally in the model.

4.8 Conclusions

The present 3D QSAR data set consisting of 30 compounds was analysed using the GRID and the GOLPE programs for the description of the ligands and the regression analysis, respectively. The original 25110 descriptors were reduced to 784 by GOLPE variable selection, with increased predictability as the result. As the number of variables was reduced the predictability, *i.e.* the crossvalidated Q^2 of the model increased, indicating the importance of a thorough variable selection procedure. This confirms what previously have been found by others.[13,30] The final model had a crossvalidated Q^2 of 0.65 which can be considered sufficient for a 3D QSAR model. The importance of focusing not only on the crossvalidated Q^2 when validating a 3D QSAR model, but also to study the grid plots of the PLS coefficients in combination with the actual field plots, is stressed.

This model has been validated internally with different crossvalidation experiments but real validation can only be performed by prediction of an external test set; a test set with compounds that have had no influence on the calibration process. In the following chapter a test set with 21 compounds has been added for validation purposes.

4.9 References

1. Sautel, F. and Griffon, N. A Functional Test Identifies Dopamine Agonists Selective for D_3 versus D_2 Receptors. *NeuroReport* **1995**, *6*, 329-332.
2. Glase, S.; Akunne, H.C.; Heffner, T.G.; Johnson, S.J.; Kesten, S.R.; MacKenzie, R.G.; Manley, P.J.; Pugsley, T.A.; Wright, J.L.; Wise, L.D. 4-Bromo-1-methoxy-*N*-[2-(4-aryl-1-piperazinyl)ethyl]-2-naphtalenecarboxamides: Selective Dopamine D_3 Receptor Partial Agonists. *Bioorg. & Med. Chem. Lett.* **1996**, *6*, 1361-1366.
3. SYBYL- Molecular Modeling Software, *6.3*, Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA,
4. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analyses (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
5. GRID, Goodford, P.J. Molecular Discovery Ltd, University of Oxford, England, SGI.
6. GOLPE, *3.0*, Clementi, S. Multivariate Infometric Analyses(MIA), Perugia, Italy, SGI.
7. Goodford, P.J. Multivariate Characterization of Molecules for QSAR Analyses. *J. of Chemometrics* **1996**, *10*, 107-117.
8. Cramer III, R.D. and Wold, S. inventors. Comparative Molecular Field Analyses (COMFA). 5025388. United States. Date Filed: **1988/08/26.**

9. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185,* 1-17.
10. Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons: New York, **1989**
11. Computer-Assisted Lead Finding and Optimization. Current tools for Medicinal Chemistry. van de Waterbeemd, H.; Testa, B.; Folkers, G. Eds.; VHCA and Wiley: Weinheim, **1997**;
12. Cruciani, G.; Baroni, M.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part 1: Standard Deviation of Prediction Errors (SDEP). *J. of Chemometrics* **1992**, *6,* 335-346.
13. Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model. *J. of Chemometrics* **1992**, *6,* 347-356.
14. Mitchell, T.J. An Algorithm for the Construction of "D-optimal" Experimental Designs. *Technometrics* **1974**, *16,* 203-210.
15. Steinberg, D.M. and Hunter, W.G. Experimental Design Review and Comment. *Technometrics* **1984**, *26,* 71-76.
16. Morgan, E. Chemometrics: Experimental Design. Chadwick, N. Ed.; John Wiley and Sons Ltd: Chichester, **1991**
17. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct. -Act. Relat.* **1993**, *12,* 9-20.
18. Mohamadi, F.; Richards, N.G.J.; Guida, W.C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W.C. MacroModel-An integrated Software System for Modeling Organic and Bioorganic Molecules Using Molecular Mechanics. *J. Comp. Chem.* **1990**, *11,* 440-467.
19. Clark, M.; Cramer III, R.D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comp. Chem.* **1989**, *10,* 982-1012.
20. APOLLO - Automated PharmacOphore Location Through Ligand Overlap, Koehler, K.F. The APOLLO program available upon request from Konrad F. Koehler (koehler@irbm.it), SGI.
21. Babel- fileconversion, *1.1,* Walters, P. and Stahl, M. Dolata Research Group,
22. Van Vliet, L.A.; Tepper, P.G.; Dijkstra, D.; Damsma, G.; Wikström, H.; Pugsley, T.A.; Akunne, H.C.; Heffner, T.G.; Glase, S.A.; Wise, L.D. Affinity for Dopamine D_2, D_3 and D_4 Receptors of 2-Aminotetralins. Relevance of D_2 Agonist Binding for Determination of Receptor Subtype Selectivity. *J. Med. Chem.* **1996**, *39,* 4233-4237.
23. Cheng, Y.-C. and Prusoff, W.H. Relationship Between the Inhibition Constant (K_i) and the Concentration of Inhibitor Which Causes 50 % Inhibition (IC_{50}) of an Enzymatic Reaction. *Biochem. Pharm.* **1973**, *22,* 3099-3108.
24. Wold, S.; Albano, C.; Dunn III, W.J.; Edlund, U.; Esbenssen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. Proceedings of the NATO Advanced Study on Chemometrics. Chemometrics Mathematics and Statistics in Chemistry. Kowalski, B.R. Ed. D. Reidel Publishing Company.: Dordrecht, **1984**; pp. 1-79.
25. Nilsson, J.; Wikström, H.; Smilde, A.K.; Glase, S.; Pugsley, T.A.; Cruciani, G.; Pastor, M.; Clementi, S. A GRID/GOLPE 3D-QSAR Study on a Set of Benzamides and Naphthamides, with Affinity for the Dopamine D_3 Receptor Subtype. *J. Med. Chem.* **1997**, *40,* 833-840.
26. Wold, S. Validation of QSAR's. *Quant. Struct. -Act. Relat.* **1991**, *10,* 191-193.
27. Cramer III, R.D.; Bunce, J.D.; Patterson, D.E.; Frank, I.E. Crossvalidation, Bootstrapping and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct. -Act. Relat.* **1988**, *7,* 18-25.
28. March, J. Advanced Organic Chemistry: Reactions, Mechanisms and Structure. March, J. Ed.; John Wiley & Sons: New York, **1996**; pp. 75-79.
29. YAK, *3.8-GL,* SIAT Biographics Laboratory, YAK- minireceptor building software,
30. Cruciani, G. and Watson, K.A. Comparative Molecular Field Analysis Using GRID force field and GOLPE Variable Selection Methods in a Study of Inhibitors of Glycogen Phosphorylase b. *J. Med. Chem.* **1994**, *37,* 2589-2601.

Multilinear PLS Analysis Application to a 3D QSAR Data Set

5

This chapter is based on the article: Nilsson, J.; De Jong, S.; Smilde, A. K. Multiway Calibration in 3D QSAR. *J. of Chemometrics* **1997**,11, 511-524.

Summary

The multilinear PLS method has been employed for the analysis of a set of benzamides with affinity for the dopamine D_3 receptor subtype, synthesised as potential drugs against schizophrenia. The key issue in 3D QSAR modelling is to obtain a predictive model that is easy to interpret. Each component in the multilinear PLS model accounts for clearly defined spatial regions, e.g., substituent positions, while the bilinear PLS solution is general and more difficult to interpret. The best models were obtained after four components with multilinear PLS (Q^2 = 51 %) and after only one component with bilinear PLS (Q^2 = 50 %). The external test set was better predicted with multilinear PLS (Q^2 = 31 %) as compared with bilinear PLS (Q^2 = 25 %). Additionally, with multilinear PLS one loses in fit, but gains in stability and simplicity due to the smaller number of parameters that need to be estimated, as compared with bilinear PLS. Finally, multilinear PLS is also less influenced by insignificant variation in the descriptor block, which stabilises the 3D QSAR model.

5.1 Introduction

Since Cramer *et al.*[1] presented the Comparative Molecular Field Analysis (CoMFA)[1,2] procedure in 1988, it has frequently been utilised by medicinal[3-5] and environmental chemists,[6] as implemented in the SYBYL molecular modelling package.[7] Today, other similar approaches are available, *e.g.*, the GRID[8] program in combination with GOLPE variable selection[9] (see also Chapter 4). Rational drug design with 3D QSAR comprises several subsequent steps: conformational analyses, alignment of the molecules, generation of molecular descriptors and regression analysis. Optionally, one or more biological response(s) can be used as the independent variable(s).

First, low energy conformations of the molecules are aligned by superimposition of mutual and possible interaction points with the target receptor protein (Chapters 4 and 6). This is by far the most crucial step in order to achieve reliable 3D QSAR models.

A molecular field is a three dimensional grid, large enough to enclose all the aligned molecules, where in each grid point interactions between a probe atom and each molecule are calculated (see Chapter 1). The interaction values in the grid points are thus utilised as variables in the subsequently following regression analysis.

Since multicollinearity among the descriptor variables may affect the regression analysis detrimentally, PLS[10] is frequently used as the regression method in 3D QSAR. Recently, Bro[11] presented the multilinear PLS algorithm (N-PLS, Chapter 2) and demonstrated some additional advantages; multilinear PLS was less influenced by noise, more stable, increased the predictability[12] and improved the interpretation of the result as compared to other methods applied to his data set. Accordingly, the multilinear PLS algorithm was implemented for the analysis of a 3D QSAR data set comprising a set of benzamides and naphthamides[13,14] (Figure 5.1, Tables 4.1–4.3) characterised in the GRID program. The same data set was also analysed in the previous chapter, utilising the GRID/GOLPE approach. In this chapter, the performances of the N-PLS[11] and the bilinear PLS[10,15] methods have been scrutinised and compared.

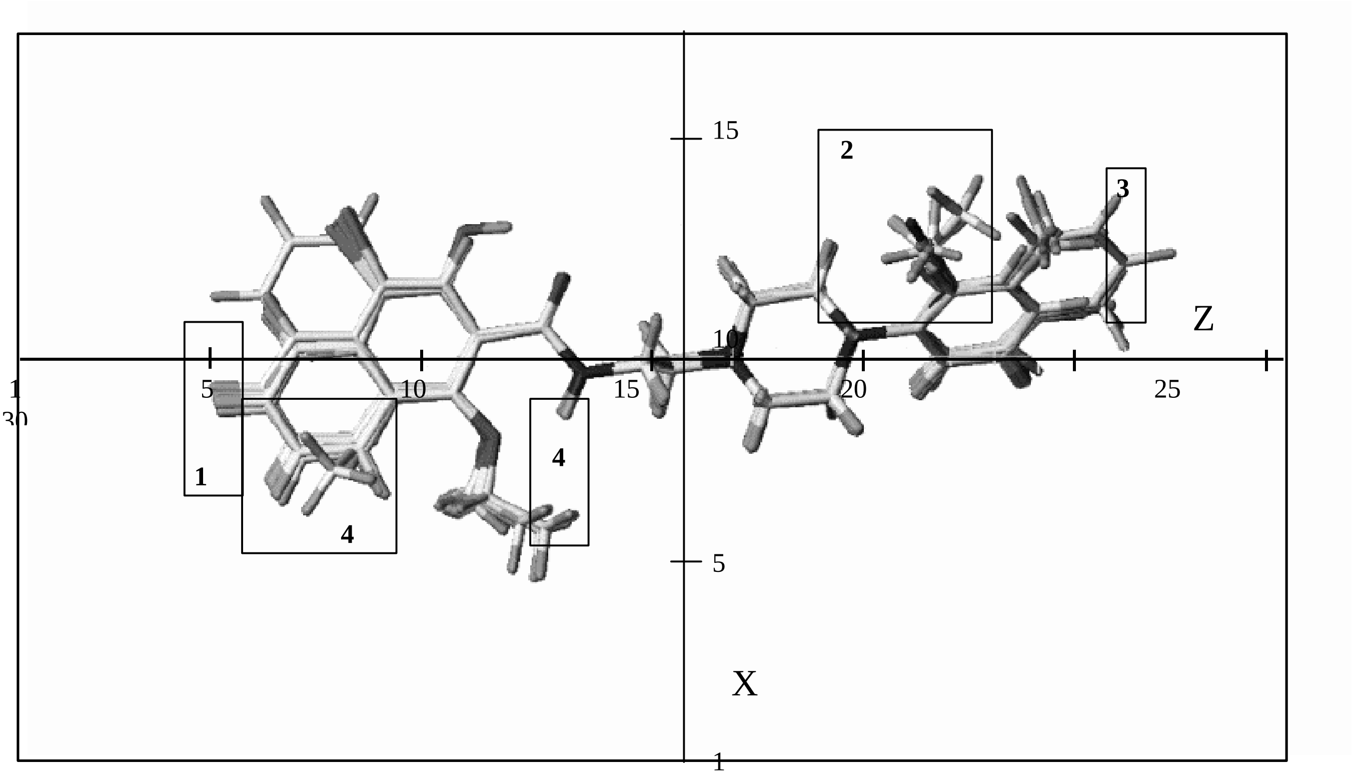


Figure 5.1 The 30 aligned molecules included in the training set viewed in the x and the z mode. The squares indicates the regions where the first four N-PLS components are focused.

It is well known that redundant variables may affect the regression analysis detrimentally and, consequently, several methods to reduce the number of variables[9,16,17] have been proposed. Multilinear PLS has been employed for the variable reduction of the present data set and the performance of the reduced model was compared with that of the complete model.

5.2 Theory

The multilinear PLS algorithm[11,18] and the bilinear PLS algorithm[10,15] have both been, thoroughly, described in Chapter 2. Accordingly, the theory described below are tools, *i.e.*, partial PLS coefficients and leverages, used for the interpretation of the multilinear PLS models. In addition to the data set analysed in the previous chapter, a test set consisting of 21 compounds has been added, to the present investigation, for validation purposes.

ː Partial PLS Coefficients

For the purpose of interpretation, the results from CoMFA studies are often presented with contour plots of the partial regression coefficients $\mathbf{b}_{\mathrm{PLS}}$.[7] Basically, the coefficients $\mathbf{b}_{\mathrm{PLS}}$ are needed for predictions of new samples, but since the sizes and the signs of the coefficients reveal the relative importance of the variables, they are also suitable for the interpretation.

A direct relationship between $\mathbf{X}^{(0)}$ and $\hat{\mathbf{y}}$ is searched for:

$$\hat{\mathbf{y}} = \mathbf{T}\mathbf{b}_{\mathrm{A}} = \mathbf{X}^{(0)}\mathbf{b}_{\mathrm{PLS}} \tag{5.1}$$

where $\mathbf{X}^{(0)}$ ($I \times R$) is the unfolded original $\underline{\mathbf{X}}$, $\hat{\mathbf{y}}$ ($I \times 1$) is the fitted $\mathbf{y}$, $\mathbf{b}_{\mathrm{A}}$ ($A \times 1$) are the regression coefficients as defined in Equation 5.2 and $\mathbf{T}$ ($I \times A$) is the score matrix. The derivation of the full and closed predictions with multilinear PLS has been presented by Smilde,[18] but since the PLS coefficients are frequently utilised in 3D QSAR, it is essential to repeat the derivation also in this context.

Since the scores from different components are not orthogonal the regression coefficients $\mathbf{b}_{\mathbf{A}}$, in Equation 5.1, have to be calculated taking all the score vectors into account:

$$\mathbf{b}_{\mathrm{A}} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{5.2}$$

Additionally, the weights obtained with multilinear PLS are also not orthogonal and need to be taken into account when the $\mathbf{b}_{\mathrm{PLS}}$ coefficients are derived (below).

For clarity, $\mathbf{X}$ is updated after the *a*th component with $\mathbf{X}^{(a)} = \mathbf{X}^{(a-1)} - \mathbf{t}_a\mathbf{w}_a^{\mathrm{T}}$, as in Martens' non-orthogonalized PLS algorithm.[15] If $\underline{\mathbf{X}}$ is three-way, $\mathbf{w}_a = \mathbf{w}_k^K \otimes \mathbf{w}_j^J$, where $\otimes$ represents the Kronecker product then

$$\mathbf{t}_1 = \mathbf{X}^{(0)}\mathbf{w}_1 \tag{5.3}$$

$$\mathbf{t}_2 = \mathbf{X}^{(\mathbf{1})}\mathbf{w}_2 = \left(\mathbf{X}^{(\mathbf{0})} - \mathbf{t}_1\mathbf{w}_1^{\mathrm{T}}\right)\mathbf{w}_2 = \left(\mathbf{X}^{(\mathbf{0})} - \mathbf{X}^{(\mathbf{0})}\mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\mathbf{w}_2 = \mathbf{X}^{(\mathbf{0})}\left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\mathbf{w}_2 \tag{5.4}$$

$$\cdots$$

$$\mathbf{t}_A = \mathbf{X}^{(0)}\left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\ldots\left(\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A\text{-}1}^{\mathrm{T}}\right)\mathbf{w}_A \tag{5.5}$$

With $\mathbf{T} = (\mathbf{t}_1|\mathbf{t}_2|\ldots|\mathbf{t}_A)$ the following holds:

$$\mathbf{T} = \mathbf{X}^{(0)}\left[\mathbf{w}_1\middle|\left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\mathbf{w}_2\middle|\ldots\middle|\left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\left(\mathbf{I} - \mathbf{w}_2\mathbf{w}_2^{\mathrm{T}}\right)\ldots\left(\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A\text{-}1}^{\mathrm{T}}\right)\mathbf{w}_A\right] \tag{5.6}$$

Insertion of Equation 5.6 in Equation 5.1 followed by rearrangement gives:

$$\mathbf{b}_{\mathrm{PLS}} = \left[\mathbf{w}_1 \middle| \left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\mathbf{w}_2 \middle| \ldots \middle| \left(\mathbf{I} - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\left(\mathbf{I} - \mathbf{w}_2\mathbf{w}_2^{\mathrm{T}}\right)\ldots\left(\mathbf{I} - \mathbf{w}_{A-1}\mathbf{w}_{A\text{-}1}^{\mathrm{T}}\right)\mathbf{w}_A\right]\mathbf{b}_A \tag{5.7}$$

When the number of variables is large, as in 3D QSAR, computing the outer product of the weights can be a problem. However, computational shortcuts are possible (see below).

If $\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = 0 \quad (i \neq j)$ then Equation 5.7 reduces to:

$$\mathbf{b}_{\mathrm{PLS}} = [\mathbf{w}_1|\mathbf{w}_2|\ldots|\mathbf{w}_A]\mathbf{b}_A = \mathbf{W}\mathbf{b}_A \tag{5.8}$$

which resembles the solution obtained with Martens' non-orthogonalized PLS algorithm.[15]

: Leverages

In order to determine which variables that have influenced the model most, the variables were ranked by their leverages[15] (**h**). The leverages are determined by first calculating an overall weight matrix, $\mathbf{W} = (\mathbf{w}_1|\mathbf{w}_2|...|\mathbf{w}_A)$, in which $\mathbf{w}_a$ ($R \times 1$; $R = JKLM$) combines the weights from the different modes as

$$\mathbf{w}_a = \mathbf{w}_a^M \otimes \mathbf{w}_a^L \otimes \mathbf{w}_a^K \otimes \mathbf{w}_a^J \quad (a = 1,\ldots,A) \tag{5.9}$$

The ⊗ sign represents the Kronecker product and a denotes the component number. The leverage vector[15] **h** ($R \times 1$) after A components is then expressed as

$$\mathbf{h} = \mathrm{diag}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}}\right) \tag{5.10}$$

A variable with a leverage h_r close to zero has not affected the model very much while a variable with a h_r close to one is very important for the model. The average h_r is A/R and variables with leverage exceeding $h_{\mathrm{cut}} \times A/R$ (h_{cut} being an integer, normally 2 or 3) may, according to Martens and Næs,[15] be considered significant.

: Model Validation

In the present investigation crossvalidation and external predictions have been utilised for the validation of the obtained models. The results from the validation experiments are quantified with the crossvalidated Q^2 and the predicted Q^2 as calculated in Equation 5.11. The quality of the calibrations are given by the multiple regression coefficient R^2 in Equation 5.12:

$$\mathrm{Q}^2 = \left[1 - \left(\sum_{i=1}^{I}\left(y_i - \hat{y}_{(i)}\right)^2 \Big/ \sum_{i=1}^{I}\left(y_i - \bar{y}\right)^2\right)\right] \times 100 \tag{5.11}$$

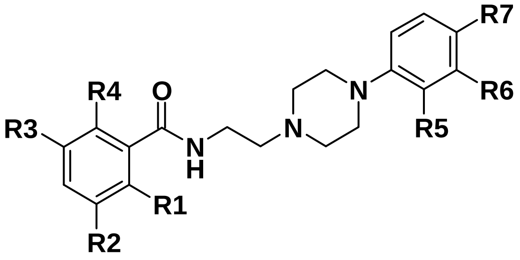
$$\mathrm{R}^2 = \left[1 - \left(\sum_{i=1}^{I}\left(y_i - \hat{y}_i\right)^2 \Big/ \sum_{i=1}^{I}\left(y_i - \bar{y}\right)^2\right)\right] \times 100 \tag{5.12}$$

The predicted y in Equation 5.11 is denoted $\hat{y}_{(i)}$, *i.e.*, in the case of crossvalidation an estimation of y_i using a model with the ith object excluded. In the case of external predictions, y_i is the response of the ith test object estimated with the complete calibration model. The fitted y from the calibration in Equation 5.12 is denoted $\hat{y}_i$.

The Test Set

The molecules analysed in this investigation were synthesised by Glase *et al.*[13] In addition to the 30 compounds from the previous chapter a test set, consisting of 21 compounds, was introduced for validation purposes (Tables 5.1–5.3).

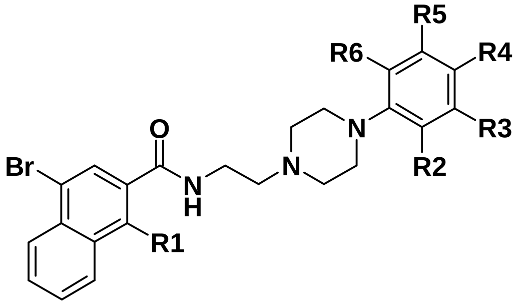
Table 5.1 The benzamides included in the test set used for the validation of the models obtained in this chapter



Compd	R1	R2	R3	R4	R5	R6	R7	$\log_{10}(K_i)^a$
t1								2.5
t2	-OMe	-Cl	-Cl	-OH	-OMe			2.9
t3	-OMe	-Cl	-Cl	-OH		-Cl	-F	3.1
t4	-OMe	-Cl	-Cl	-OH		$-CF_3$		3.3

[a] $\log_{10}$ was performed on the K_i(nM)

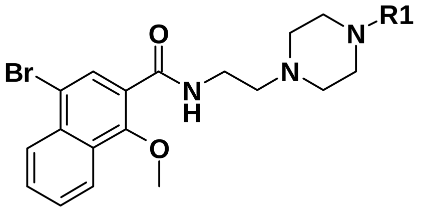
Table 5.2 Naphthamides included in the test set used for the validation of the models obtained in this chapter.

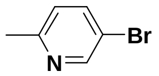
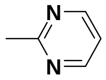


Compd	R1	R2	R3	R4	R5	R6	$\log_{10}(K_i)^a$
t5	-OMe	-Cl	-Cl				0.9
t6	-OH	-Cl	-Cl				1.7
t7	-OMe	-Cl			-Cl		2.4
t8	-OMe			-Cl			2.1
t9	-OMe		-Cl	-F			2.9
t10	-OMe	-F	-F				1.7
t11	-OMe	-F			-F		1.5
t12	-OMe		-F				2.1
t13	-OMe		-F	-F			2.6
t14	-OMe	-Br					2.3
t15	-OMe		-Br				1.6
t16	-OMe			-Br			3.1
t17	-OMe		-CN				0.9
t18	-OMe			-CN			3.2
t19	-OMe	-Me				-Me	2.7

[a] $\log_{10}$ was performed on the K_i(nM)

Table 5.3 Naphthamides included in the test set used for the validation of the models obtained in this chapter.



Compd	R1	log$_{10}$(K$_i$)[a]
t20		2.4
t21		3.0

[a] log$_{10}$ was performed on the K$_i$(nM)

Low energy conformations of all the molecules were initially aligned as described in Chapter 4 and subsequently surrounded by a three dimensional grid large enough to enclose all the aligned molecules with four Å in all directions (Figure 5.1). The directions x, y and z in the grid were divided into 31, 15 and 18 steps of 1 Å, respectively, yielding a total of 8370 grid points. The surroundings of each molecule were mapped by calculating the interactions between probe atoms and each molecule at each grid point. The resulting grid, filled with interaction values, is called a molecular field. Three different probes[8] were used, a carbon atom (the C3 probe), a water molecule (the OH2 probe) and a plus two charged calcium ion (the CA+2 probe), reflecting the steric field, the hydrogen bonding field and the electrostatic field, respectively. In CoMFA, the differences in these fields are correlated with, *e.g.*, the affinities for a certain receptor subtype. The complete model is described graphically in Figure 5.2.

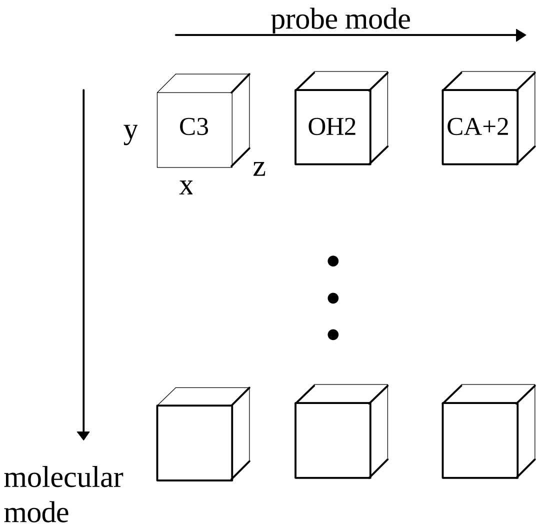


Figure 5.2 The data set comprises five different modes. The molecular, x, y, z and the probe modes consist of 30, 31, 15, 18 and 3 dimensions, respectively.

Prior to bilinear PLS analysis, the data set is unfolded to form a two-way matrix which is decomposed into scores **t** ($I \times 1$) and loadings **p** ($JKLM \times 1$) as described in Figure 2.4. With multilinear PLS, however, the unfolding step is omitted and the one-component decomposition consists of a score vector **t** ($I \times 1$) and four weight vectors $\mathbf{w}^J$ ($J \times 1$), $\mathbf{w}^K$ ($K \times 1$), $\mathbf{w}^L$ ($L \times 1$) and $\mathbf{w}^M$ ($M \times 1$), as

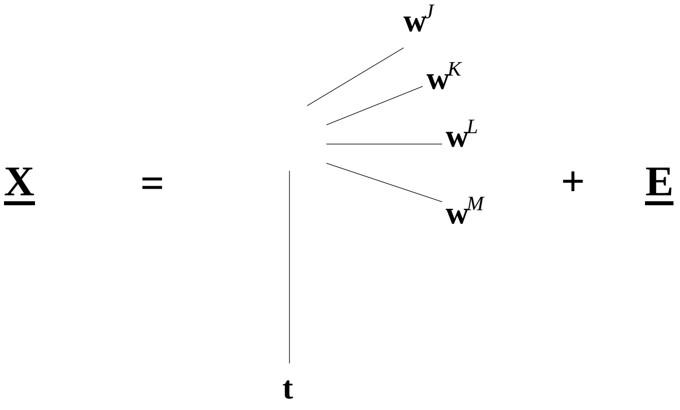


Figure 5.3 The multiway decomposition of $\underline{\mathbf{X}}$ ($I \times J \times K \times L \times M$) into a score vector **t** ($I \times 1$) and four weight vectors $\mathbf{w}^J$ ($J \times 1$), $\mathbf{w}^K$ ($K \times 1$), $\mathbf{w}^L$ ($L \times 1$) and $\mathbf{w}^M$ ($M \times 1$). $\underline{\mathbf{E}}$ is the part of $\underline{\mathbf{X}}$ not accounted for by the model.

in Figure 5.3. The vectors **t**, $\mathbf{w}^J$, $\mathbf{w}^K$, $\mathbf{w}^L$ and $\mathbf{w}^M$ correspond directly to the molecular, x, y, z and the probe mode, respectively, as described in Figure 5.2.

5.3 Results

: Model I

The only data pre-processing applied was mean-centering in the molecular mode. In bilinear PLS, scaling is often performed column-wise, *e.g.*, auto-scaling[10] whereas in multilinear PLS scaling is not that straightforward.[19]

The objective of this investigation is to introduce the multilinear PLS method in 3D QSAR modelling and compare its solution with the bilinear PLS solution. Accordingly, the complete model (Model I) was calibrated and validated with both regression methods, presented in Tables 5.4 and 5.5, respectively. With multilinear PLS (Table 5.4) maximum crossvalidated Q^2 was obtained after four components ($Q^2 = 51$ %), where 17 % of the variation in $\underline{\mathbf{X}}$ explained 73 % of the variation in **y**. With bilinear PLS, however, maximum crossvalidated Q^2 was found after only one component ($Q^2 = 48$ %), where 22 % of the variation in **X** explained 62 % of the variation in **y**. The weights from the different modes obtained with multilinear PLS are useful for the interpretation of the result. The weights from the first four components are plotted in Figure 5.4. For comparison, the weight vector from the first component with the reduced bilinear PLS model is plotted in Figure 5.5.

Table 5.4 Calibration and validation of Model I (30 × 25110) with multilinear PLS.[a]

#LV	R^2 ($\underline{\mathbf{X}}$)	R^2 ($\mathbf{y}$)	Q^2 (LOO)	Q^2 (Pred)[b]
1	7	48	39	19
2	12	58	43	18
3	15	64	45	29
4	**17**	**73**	**51**	**31**
5	18	76	34	34

[a] all values in percentage; [b] predictions of the external test set (21 × 25110)

Table 5.5 Calibration and validation of Model I (30 × 25110) with bilinear PLS.[a]

#LV	R^2 ($\mathbf{X}$)	R^2 ($\mathbf{y}$)	Q^2 (LOO)	Q^2 (Pred)[b]
1	**22**	**62**	**48**	**26**
2	34	76	47	21
3	43	86	46	32
4	53	89	42	31
5	59	93	37	32

[a] all values in percentage; [b] predictions of the external test set (21 × 25110)

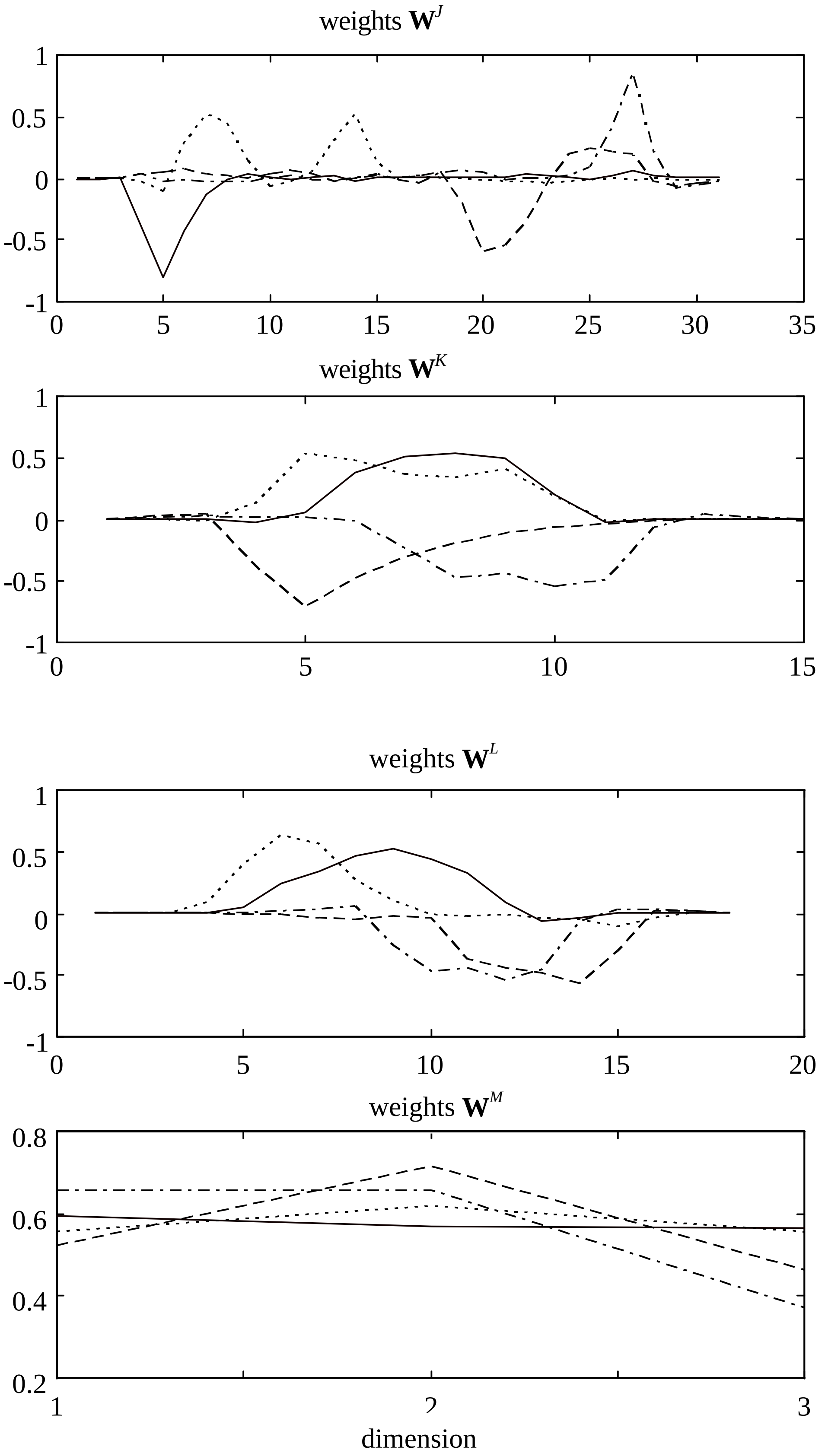


Figure 5.4 The weights $\mathbf{W}^J$, $\mathbf{W}^K$,$\mathbf{W}^L$ and $\mathbf{W}^M$:——— , component one;- - - - - - , component two; -·-·-·-·, component three; ··········, component four.

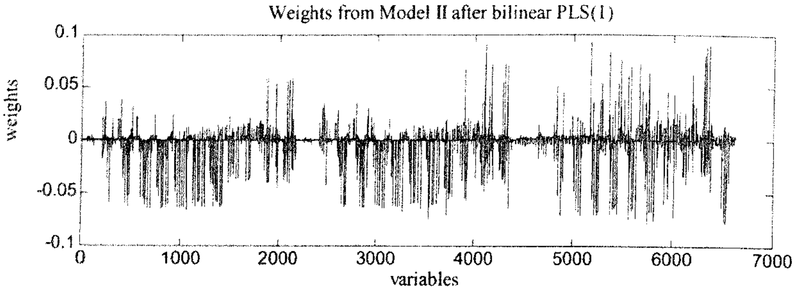


Figure 5.5 Weight vector ($\mathbf{w}_1$) after first component with bilinear PLS.

The number of significant components was estimated by leave-one-out (LOO) crossvalidation and maximum crossvalidated Q^2 was found after four components (Table 5.4) with multilinear PLS. In order not to lose information during the variable reduction step, the variable reduction was performed from a model one component more complex than optimal. Accordingly, the absolute sum of the weights, after the first five components, was calculated for each mode separately. A position in a mode was considered significant and selected only if it exceeded a lower cut-off value. An arbitrary cut-off value of 0.2 generated a reduced data set with 6624 number of variables, called Model II. Stated differently, only variables with high weights from Model I were selected and included in Model II. The probe mode was left intact, hence variables from all three probes were included in the reduced data set.

■ Model II

The results from Model II are summarised in Tables 5.6 and 5.7 which was validated thoroughly (Table 5.7) with crossvalidation and external predictions. In addition to traditional 'leave-one-out' crossvalidation also 'leave-three-out' and 'leave-five-out' crossvalidations were performed, where in each experiment objects were left out randomly but only once. The results are reported as the average Q^2 of 20 crossvalidation experiments.[20,21]

In order to simplify the interpretations of a PLS model in 3D QSAR, the partial PLS coefficients $\mathbf{b}_{PLS}$ in Equation 5.8 are often presented as comprehensive iso-contour plots. That is, each $\mathbf{b}_{PLS}$ is transferred back to its original position in the grid, where grid points with similar coefficients are connected. In Figure 5.6, the $\mathbf{b}_{PLS}$ contour is plotted in stereo from the C3 probe after the fourth multilinear PLS component.

Table 5.6 Crossvalidations and external predictions of Model II (30 × 6624).

# LV	N-PLS[a]				PLS[a]			
	LOO[b]	L3O[b]	L5O[b,c]	Pred.[d]	LOO[b]	L3O[b]	L5O[b,c]	Pred.[d]
1	39	43	42	19	**50**	**50**	**50**	**25**
2	43	44	41	18	48	46	47	23
3	45	43	41	29	48	46	48	33
4	**51**	**53**	**49**	**31**	44	41	42	30
5	43	42	38	34	39	37	40	31

[a] all values in percentage; [b] LOO is short for leave-one-out, L3O for leave-three-out and L5O for leave-five-out; [c] average from 20 Q^2s; [d] predictions of the external test-set (21 × 6624)

Table 5.7 Calibration of Model II with bilinear PLS and multilinear PLS for the first five components.[a]

# LV	N-PLS		PLS	
	R^2 ($\underline{\mathbf{X}}$)	R^2 ($\mathbf{y}$)	R^2 ($\mathbf{X}$)	R^2 ($\mathbf{y}$)
1	8	48	22	64
2	13	58	32	79
3	16	64	41	86
4	19	73	51	90
5	20	76	58	93

[a] all values in percentage

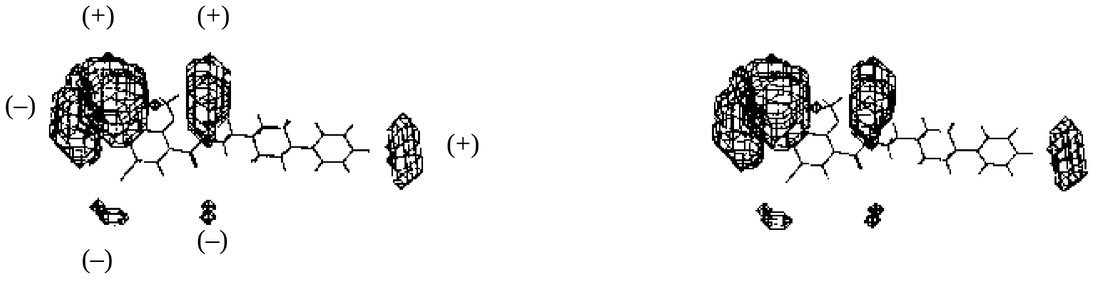


Figure 5.6 The $\mathbf{b}_{\text{PLS}}$ coefficients from the final multilinear PLS model and C3 probe after four components.

5.4 Discussion

The key issue in 3D QSAR modelling is to find a predictive model which can be used as a tool in the design of new compounds. The solution should also be simple and straightforward, since also the non-expert must be able to interpret the model.

The initial complete model (Table 5.4) indicated four significant components with leave-one-out crossvalidation. With help from Figure 5.4 it can be determined, with good precision, which regions are accounted for by the components. The full lines in Figure 5.4 represent the weights from the first component, the broken curves the second component, the chain curves the third and the dotted curves the fourth component. For clarity, the weights $\mathbf{W}^J$ and $\mathbf{W}^L$ correspond to the x and z modes in Figure 5.1, respectively. The first component has high weights $\mathbf{w}^J$ in position 5 and high weights $\mathbf{w}^L$ between positions 5 and 10 (Figure 5.4), which correspond to the region where the naphthalene moiety protrudes (Figure 5.1). Thus the first component accounts for the differences between naphthamides and benzamides. Similarly, it can be concluded that the second component mainly deals with the ortho and meta positions on the arylpiperazine phenyl ring, the third component the para position and, finally, the fourth component with substituents on the benzamide phenyl ring.

In contrast with the weights from multilinear PLS (Figures 5.4), the weights from bilinear PLS (Figure 5.5) are difficult to interpret.

Striking is the significantly lower percentage variance explained with multilinear PLS method (Table 5.4) as compared with the bilinear PLS (Table 5.5) method. A speculative explanation for this is the fewer parameters that need to be estimated with multilinear PLS.[11,22] Additionally, each component in multilinear PLS focuses on small specific items, *e.g.*, regions in the grid, while bilinear PLS searches for more general directions for its components and is more flexible.

It is well known that many of the variables in a 3D QSAR model are more or less redundant and may affect the predictability detrimentally. From Figure 5.4 it is clear that positions corresponding to grid points in the periphery of the grid have low weights and also limited influence on the model. By omitting these variables, as described above, a reduced model with 6624 variables was obtained which was validated with crossvalidation and external predictions. Variable selection must be performed very carefully, otherwise problems with overfitting may occur. Norinder[17] and Cho[16] reported that the crossvalidated Q^2 increased in their models, while the ability to predict external test sets decreased when the number of variables was reduced. In the present investigation the number of variables were reduced from 25110 to 6624, which speeded up further calculations, but with no improvement in the predictability (nor decrease) as the result.

From Model II (Table 5.6) it can be concluded that the model is homogenous and stable, since the crossvalidated Q^2 was not affected very much when larger groups of molecules were left out each time. Each crossvalidation experiment was repeated 20 times[20] and, accordingly, reported as the average Q^2.

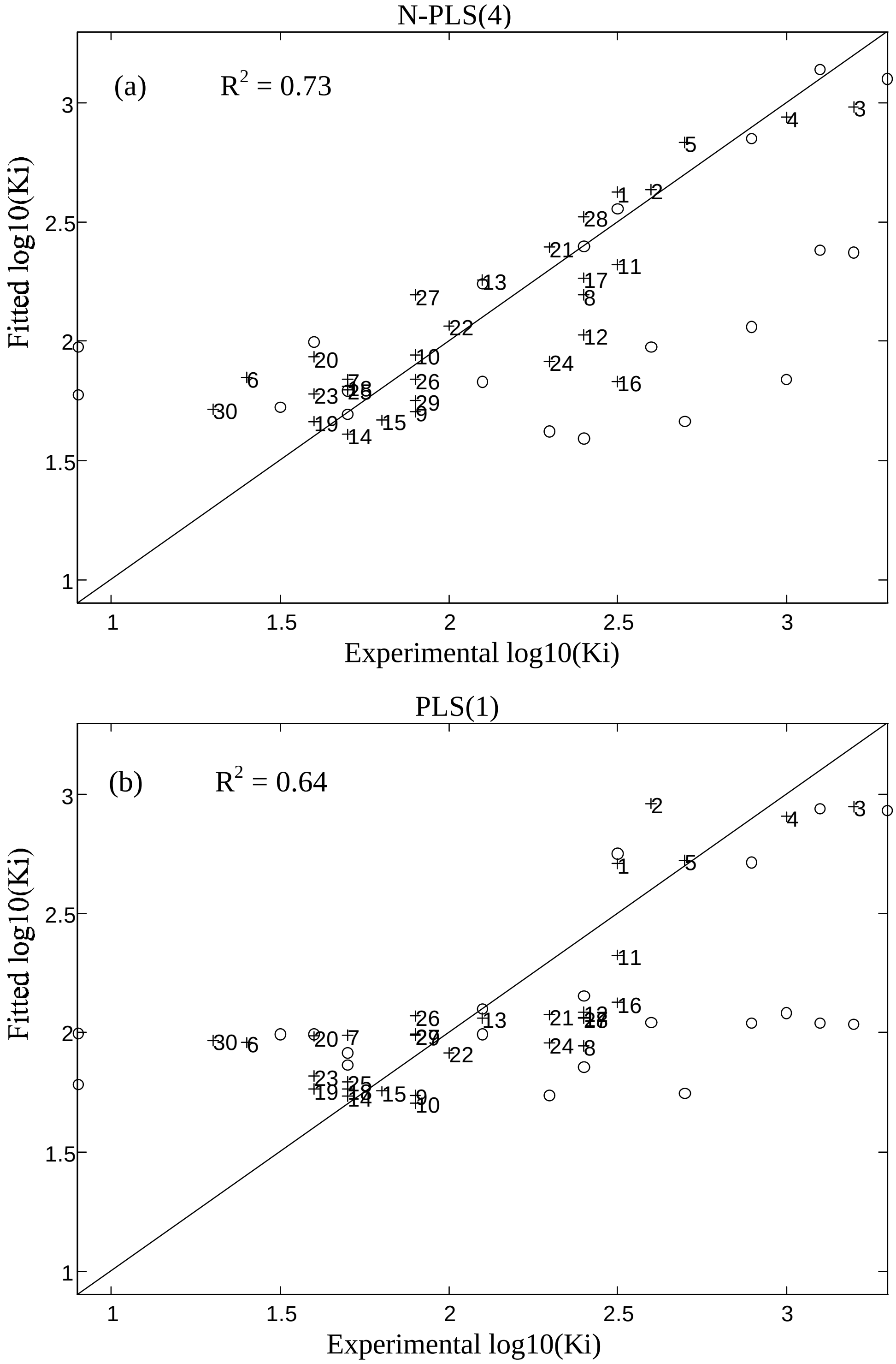


Figure 5.7 Experimental $\log_{10}(K_i)$ versus fitted $\log_{10}(K_i)$ after (a) four component with multilinear PLS and (b) one component with bilinear PLS. The small rings represents the predictions of the external test set (21 × 6624).

In Figure 5.7(a) and 5.7(b) the experimental $\log_{10}(K_i)$ are plotted against the fitted $\log_{10}(K_i)$ from model II for the training set with multilinear PLS and bilinear PLS, respectively. The 21 test compounds have been predicted and plotted on the same figures as small circles. The four-component model with multilinear PLS ($R^2 = 73$ %) explains more of the variation in **y** as compared to the one-component bilinear PLS model ($R^2 = 64$ %). The test compounds were also better predicted with multilinear PLS ($Q^2 = 31$ %) than with bilinear PLS ($Q^2 = 25$ %). In fact, the bilinear

PLS model (Figure 5.7(b)) more or less distinguishes between two groups of compounds, *i.e.*, between benzamides and naphthamides, while the multilinear PLS model is much better fitted (Figure 5.7(a)).

The iso-contour plot of the $\mathbf{b}_{\mathrm{PLS}}$ coefficients after the fourth component, in Figure 5.6, is probably the most comprehensible tool for the interpretation of the model:

$$y = x_1 b_1 + \ldots + x_i b_i + \ldots + x_R b_R + e \tag{5.13}$$

If a novel molecule is designed with a substituent protruding in a negative $\mathbf{b}_{\mathrm{PLS}}$ region then x_i in Equation 5.13 will be positive and consequently $x_i b_i$ will be negative. This substituent will thus have a negative effect on *y*. If low values of *y* is desirable, new substituents must be added in regions where the $\mathbf{b}_{\mathrm{PLS}}$ for the C3-probe (steric-field) is negative, and *vice versa*. For a more elaborated explanation of how to interpret the iso-contour plots the SYBYL-manual[7] or Chapter 4 in this thesis are recommended.

In Figure 5.8 the 6624 variables are ranked by their leverages. Even after variable reduction a lot of variables with low influence on the model are present.

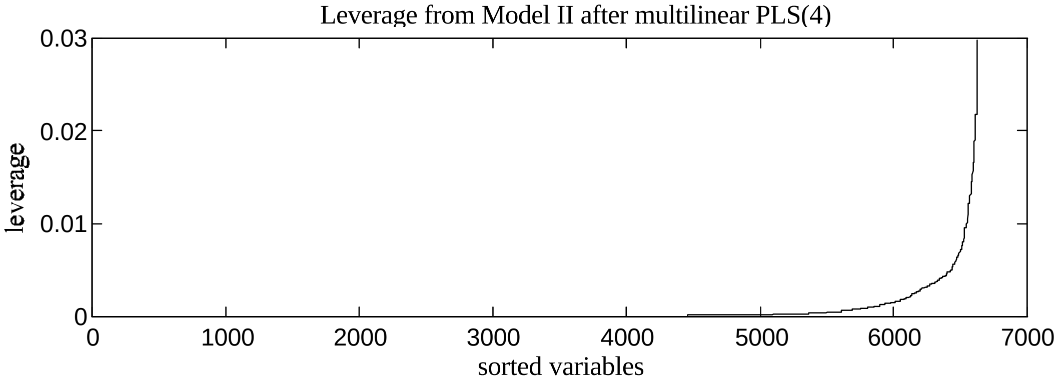


Figure 5.8 Leverage from Model II after four multilinear PLS components ordered in increasing order of size.

5.5 Conclusions

The multilinear PLS method has successfully been introduced as regression method in 3D QSAR. The main improvement lies in the interpretation of the result and the slightly better predictive ability as compared with bilinear PLS. The multilinear PLS model is also superior to bilinear PLS with regard to simplicity and stability, since fewer parameters need to be estimated.

The number of variables were effectively reduced with help from the multilinear PLS weights. The variable selection did not improve the predictability but speeded up the calculations significantly. The number of high leverage variables was quite low even after variable reduction.

5.6 Matlab Code for Regression Coefficients in Multilinear PLS

Smilde[18] gives the following explicit expression for the regression coefficients in multilinear PLS1 calibration based on A components:

$$\mathbf{b}_{\mathrm{PLS}} = \mathbf{W}^{*}\mathbf{b}_{A} \tag{5.14}$$

where

$$\mathbf{W}^{*} = \left[\mathbf{w}_1 \middle| \left(\mathbf{I}_R - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\mathbf{w}_2 \middle| \ldots \middle| \left(\mathbf{I}_R - \mathbf{w}_1\mathbf{w}_1^{\mathrm{T}}\right)\left(\mathbf{I}_R - \mathbf{w}_2\mathbf{w}_2^{\mathrm{T}}\right)\ldots\left(\mathbf{I}_R - \mathbf{w}_{A-1}\mathbf{w}_{A-1}^{\mathrm{T}}\right)\mathbf{w}_A\right] \tag{5.15}$$

In Equation 5.15 $\mathbf{w}_a$ is the vectorized (unfolded) form of the rank-1 N-way tensor product obtained from the mode-specific weight vectors $\mathbf{w}^J$, $\mathbf{w}^K$, etc. that define the ath PLS component.

Equation 5.15 is not suitable for implementation in predictive CoMFA computations using N-PLS regression since it involves very large matrices $\mathbf{I}_R - \mathbf{w}_a\mathbf{w}_a^{\mathrm{T}}$ $(R \times R)$. For example, in the current application ($R = JKLM = 31 \times 15 \times 18 \times 3 \approx 25000$) one such matrix occupies 5 Gb. Merely multiplying two such matrices takes 31 Tflops!

Let us consider the second column of $\mathbf{W}^{*}$. The expression $(\mathbf{I}_R\text{-}\mathbf{w}_1\mathbf{w}_1^{\mathrm{T}})\mathbf{w}_2$ represents the projection of $\mathbf{w}_2$ onto $\mathbf{w}_1^{\perp}$, the orthogonal complement of $\mathbf{w}_1$. It is more efficient, with respect to both space and time, to compute this as $\mathbf{w}_2 - \left(\mathbf{w}_1^{\mathrm{T}}\mathbf{w}_2\right)\mathbf{w}_1$. The same approach can be used recursively in each of the subsequent columns, starting from the back. MATLAB[23] code implementing this procedure is given below as Algorithm I. It requires little additional storage and involves $2A^2R$ flops.

The speed may be increased even further by starting at the last column of $\mathbf{W}^{*}$, *i.e.*, computing $b_A\mathbf{w}_A$, projecting this onto $\mathbf{w}_{A-1}^{\perp}$, adding the result to $b_{A-1}\,\mathbf{w}_{A-1}$, projecting this onto $\mathbf{w}_{A-2}^{\perp}$, adding the result to $b_{A-2}\,\mathbf{w}_{A-2}$, and so forth. In this way an alternative Algorithm II is obtained. It requires $(4A-3)R$ flops; hence Algorithm II is about $A/2$ times faster than Algorithm I.

Other approaches to compute N-PLS regression coefficients for prediction purposes are discussed elsewhere.[24]

ALGORITHM I

```
function bPLS = getbpls1(W, b)

% function bPLS = getbpls1(W, b)
% gives explicit b_PLS in trilinear
% PLS
%   (i.e. y^hat = X * b_PLS )
%    from W(JKxA) and b(Ax1)

A = size(W,2);
bPLS = 0;
for a=1:A
  v = W(:,a);
  for j=a-1:-1:1
    v = v-(v'*W(:,j))*W(:,j);
  end
  bPLS = bPLS + b(a)*v;
end
```

ALGORITHM II

```
function bPLS = getbpls1(W, b)

% function bPLS = getbpls1(W, b)
% gives explicit b_PLS in trilinear
% PLS
% ( i.e. y^hat = X * b_PLS )
% from W(JKxA) and b(Ax1)

A = length(b);
bPLS = b(A)*W(:,A);
for a=A-1:-1:1
  bPLS=bPLS+(b(a)-PLS'*W(:,a))*W(:,a);
end
```

5.7 References

1. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analyses (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110,* 5959-5967.
2. Cramer III, R.D. and Wold, S. inventors. Comparative Molecular Field Analyses (COMFA). 5025388. United States. Date Filed: **1988/08/26.**
3. Agarwal, A.; Pearson, P.P.; Taylor, E.W.; Li, H.B.; Dahlgren, T.; Herslof, M.; Yang, Y.; Lambert, G.; Nelson, D.L.; Regan, J.W.; et al Three-Dimensional Quantitative Structure-Activity Relationships of 5-HT Receptor Binding Data for Tetrahydropyridinylindole Derivatives: a Comparison of the Hansch and CoMFA Methods. *J. Med. Chem.* **1993**, *36,* 4006-4014.
4. Raghavan, K.; Buolamwini, J.K.; Fesen, M.R.; Pommier, Y.; Kohn, K.W.; Weinstein, J.N. Three-Dimensional Quantitative Structure-Activity Relationship (QSAR) of HIV Integrase Inhibitors: a Comparative Molecular Field Analysis (CoMFA) Study. *J. Med. Chem.* **1995**, *38,* 890-897.
5. Oprea, T.I.; Waller, C.L.; Marshall, G.R. 3D-QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. III. Interpretation of CoMFA Results. *Drug Des. Discov.* **1994**, *12,* 29-51.
6. Briens, F.; Bureau, R.; Rault, S.; Robba, M. Applicability of CoMFA in Ecotoxicology: a Critical Study on Chlorophenols. *Ecotoxicol. Environ. Saf.* **1995**, *31,* 37-48.
7. SYBYL- Molecular Modeling Software, *6.3,* Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA,
8. GRID, Goodford, P.J. Molecular Discovery Ltd, University of Oxford, England, SGI.
9. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct. -Act. Relat.* **1993**, *12,* 9-20.
10. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185,* 1-17.
11. Bro, R. Multiway Calibration. Multilinear PLS. *J. of Chemometrics* **1996**, *10,* 47-61.
12. Bro, R. and Heimdal, H. Enzymatic Browning of Vegetables. Calibration and Analysis of Variance by Multiway Methods. *Chemom. and Intell. Lab. Syst.* **1996**, *34,* 85-102.
13. Glase, S.; Akunne, H.C.; Heffner, T.G.; Johnson, S.J.; Kesten, S.R.; MacKenzie, R.G.; Manley, P.J.; Pugsley, T.A.; Wright, J.L.; Wise, L.D. 4-Bromo-1-methoxy-*N*-[2-(4-aryl-1-piperazinyl)ethyl]-2-naphtalenecarboxamides: Selective Dopamine D_3 Receptor Partial Agonists. *Bioorg. &Med. Chem. Lett.* **1996**, *6,* 1361-1366.
14. Nilsson, J.; Wikström, H.; Smilde, A.K.; Glase, S.; Pugsley, T.A.; Cruciani, G.; Pastor, M.; Clementi, S. A GRID/GOLPE 3D-QSAR Study on a Set of Benzamides and Naphthamides, with Affinity for the Dopamine D_3 Receptor Subtype. *J. Med. Chem.* **1997**, *40,* 833-840.
15. Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons: New York, **1989**
16. Cho, S.J. and Tropscha, A. Crossvalidated R^2-guided Region Selection for Comparative Molecular field Analyses: A Simple Method to Achieve Consistent Results. *J. Med. Chem* **1995**, *38,* 1060-1066.
17. Norinder, U. Single Domain Mode Variable Selection in 3D QSAR Applications. *J. of Chemometrics* **1996**, *10,* 95-105.
18. Smilde, A.K. Comments on Multilinear PLS. *J. of Chemometrics* **1997**, *11,* 367-377.
19. Smilde, A.K. Three-way Analyses. Problems and Prospects. *Chemom. and Intell. Lab. Syst.* **1992**, *15,* 143-157.
20. Cruciani, G.; Baroni, M.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part 1: Standard Deviation of Prediction Errors (SDEP). *J. of Chemometrics* **1992**, *6,* 335-346.
21. Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model. *J. of Chemometrics* **1992**, *6,* 347-356.
22. Smilde, A.K. and Doornbos, D.A. Three-way Methods for the Calibration of Chromatographic Systems: Comparing PARAFAC and Three-way PLS. *J. of Chemometrics* **1991**, *5,* 345-360.
23. Matlab, *4.2c,* Simulink Inc.
24. De Jong, S. Regression Coefficients in Multilinear PLS. *submitted* **1997**

A Multiway 3D QSAR Analysis of a Series of (*S*)-*N*-[(1-Ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides

6

This chapter is based on the article: Nilsson, J.; Homan, E. J.; Smilde, A. K.; Grol, C. J.; Wikström, H. A Multiway 3D QSAR Analysis of a Series of (*S*)-*N*-[(1-Ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides. *J. of Comp.-Aided Mol. Design*, **1997**, in press

Summary

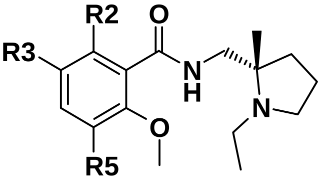
In the previous chapter the multilinear PLS algorithm (N-PLS) was implemented as a regression method in 3D QSAR. Here the method has been validated on a well-known set of (S)-N-[(1-ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides, with affinity for the dopamine D_2 *receptor subtype. After exhaustive conformational analyses on the ligands, the active analogue approach was employed to align them in their presumed pharmacologically active conformations, using (–)-piquindone as a template. Descriptors were then generated in the GRID program, and 40 calibration compounds and 18 test compounds were selected by means of a Principal Component Analysis (PCA) in the descriptor space. The final model was validated with different types of crossvalidation experiments, e.g., leave-one-out, leave-three-out and leave-five-out. The crossvalidated* Q^2 *was 62 % for all experiments, confirming the stability of the model. The prediction of the test set, with a predicted* Q^2 *of 62 %, confirmed the predictive ability. In conclusion, the conformation analysis and the alignment of the ligands in combination with multilinear PLS played an important role for the success of our model. Hence, it was shown that multilinear PLS certainly is suitable as regression method for the analysis of this type of data.*

6.1 Introduction

Ever since Cramer III *et al.*[1] introduced the CoMFA methodology it has been available as implemented in the SYBYL molecular modelling package.[2] During the last years, new methods for the alignment of the ligands[3,4] have evolved, and other methods to generate 3D descriptors[5] have become available. Additionally, several attempts have been made to reduce the number of variables[6,7] but less efforts have been undertaken to improve the multivariate analyses. In Chapter 5,[8] multilinear PLS[9] (N-PLS) was introduced as a regression method in 3D QSAR and several advantages were demonstrated, as compared to bilinear PLS. It was shown that multilinear PLS is more stable, increases the predictive ability, and improves the interpretation of the results. However, still it is necessary to evaluate whether multilinear PLS really is suitable for 3D QSAR data, or if it just happened to be successful, for the series of compounds used in Chapters 4 and 5.[8,10] Therefore, a well-known series of substituted *N*-[(1-ethyl-2-pyrrolidinyl)methyl]-derived benzamides[11,12] (Table 6.1) was chosen for a re-evaluation of this method.

Compounds of this chemical class have been shown to possess high affinity and selectivity towards dopamine D_2 receptors, and as a consequence, several compounds of this series have become valuable tools for *in vivo* and *in vitro* receptor binding studies [*e.g.*, ^{3}H-raclopride (**1**) and ^{125}I-NCQ298 (**9**)],[13] as well as for *in vivo* visualisation techniques like PET [*e.g.*, ^{11}C-raclopride (**1**) and ^{11}C-eticlopride (**36**)] and SPECT [*e.g.*, ^{123}I-NCQ298 (**9**)].[13]

Table 6.1 Aromatic substitution patterns and dopamine D_2 receptor binding properties of compounds **1–58**.



Compound	R_2	R_3	R_5	Activity[a] Exp[b]	Fitted[c]	Pred[d]
1	OH	Cl	Cl	7.49	7.75	–
2	OH	OMe	Cl	7.15	7.44	–
3	H	Br	Br	8.10	7.94	–
4	H	Et	Br	7.96	8.25	–
5	H	I	OMe	9.17	9.14	–
6	OH	*n*-Pr	Me	8.30	8.27	–
7	OH	Cl	*n*-Pr	6.96	7.16	–
8	OH	H	Et	6.91	6.82	–
9	OH	I	OMe	9.54	9.43	–
10	H	SMe	OMe	8.96	9.09	–
11	OH	Et	OMe	8.89	9.61	–
12	H	*n*-Bu	OMe	8.57	8.57	–
13	H	*n*-Pr	H	7.17	7.45	–
14	H	Cl	H	6.59	6.98	–
15	H	Cl	Cl	7.70	7.72	–
16	H	Cl	Br	8.25	7.53	–
17	H	Br	H	7.34	7.38	–
18	H	Br	OMe	8.92	8.64	–
19	H	Et	Cl	8.38	8.41	–
20	OH	H	Cl	7.19	6.96	–
21	OH	H	OMe	8.06	7.83	–
22	OH	F	H	6.44	6.70	–
23	OH	Cl	H	7.41	7.26	–
24	OH	Cl	Br	7.24	7.45	–
25	OH	Cl	Et	7.92	7.63	–
26	OH	Br	H	8.08	7.70	–

[a] Activity values are expressed in pIC50 molar units.; [b] Experimental values obtained from ref.11; [c] Fitted values (training set).; [d] Predicted values (test set).

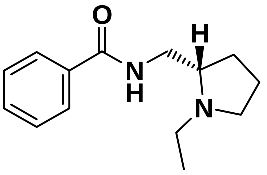
Table 6.1 Continued

				Activity[a]		
Compound	R_2	R_3	R_5	Exp[b]	Fitted[c]	Pred[d]
27	OH	Br	Cl	7.77	8.11	–
28	OH	Br	Br	7.59	7.80	–
29	OH	Br	OMe	8.85	8.90	–
30	OH	Br	NO_2	6.73	6.70	–
31	OH	I	H	8.52	8.12	–
32	OH	Me	Cl	7.59	7.96	–
33	OH	Me	Me	8.11	7.98	–
34	OH	Et	H	8.54	8.32	–
35	OH	Et	F	8.82	8.88	–
36	OH	Et	Cl	9.04	8.62	–
37	OH	Et	Br	8.64	8.32	–
38	OH	*n*-Pr	H	8.30	7.91	–
39	OH	OMe	H	6.69	7.09	–
40	OH	OMe	Br	7.17	7.12	–
41	H	Br	OH	8.00	–	8.22
42	OH	*n*-Pr	Cl	8.49	–	8.20
43	OH	Me	Br	8.26	–	7.65
44	H	Me	OMe	8.28	–	8.34
45	OH	Br	Et	7.77	–	8.01
46	OH	Et	Et	8.75	–	8.58
47	H	Et	OMe	8.89	–	9.12
48	H	H	OMe	7.28	–	7.68
49	H	Et	H	7.40	–	7.82
50	OH	H	H	6.50	–	6.52
51	OH	H	Br	7.25	–	6.67
52	OH	Cl	Me	7.96	–	7.76
53	OH	Cl	OMe	8.77	–	8.45
54	OH	Br	F	8.15	–	8.33
55	OH	Br	Me	7.96	–	8.09
56	OH	Me	H	7.72	–	7.52
57	OH	Me	*n*-Pr	6.85	–	7.41
58	OH	NO_2	H	5.52	–	7.30

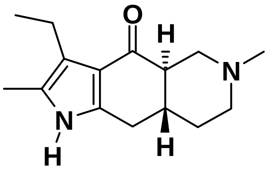
[a] Activity values are expressed in pIC50 molar units.; [b] Experimental values obtained from ref.11; [c] Fitted values (training set).; [d] Predicted values (test set).

6.2 Theory and Methods

Prior to the multilinear PLS analysis, the conformational space accessible to the ligands was sampled by performing an exhaustive conformational analysis (see below) on the basic skeleton (**59**) common to all ligands. The active analogue approach[14] was then employed to derive the presumed pharmacologically active conformation, using the rigid pyrrolo-isoquinoline (–)-piquindone (**60**) as a template (see below). The best matching conformation of **59** then served as a starting point for the construction of compounds **1–58**, which were subsequently optimised with respect to the orientations of the aromatic substituents. Finally, all ligands were superimposed in their presumed pharmacologically active conformations (see below).



59



60 Piquindone

Conformational Analyses

Conformational analyses and pharmacophore identification (see below) were performed essentially as described by Jansen *et al.*[15] Thus, conformational analyses were performed within MacroModel version 4.5,[16] using the MM2* force field and the Monte Carlo (MC) search protocol. (*S*)-*N*-[(1-Ethyl-2-pyrrolidinyl)-methyl]benzamide (**59**) was built from standard fragments and subsequently energy-minimised with default options. During all minimisation’s, the benzamide torsional angle was kept fixed at 0°. For (–)-piquindone (**60**), its X-ray crystal structure (BANKIK)[17] served as the starting conformation. The conformational space accessible to both compounds was sampled by submitting the input structures to 1000 MC steps. Starting geometry’s were generated systematically (SUMM option) and the number of torsional angles to be adjusted in each MC step was randomly varied between 2 and n–1, n being the total number of variable torsional angles. Ring closure bonds were defined in 5- and 6-membered non-aromatic rings in order to allow torsional angles in these rings to be varied during the MC search procedure. Ring closure distances were limited between 0.5 and 2.0 Å. Default values were used for testing high-energy nonbonded contacts. Duplicate minimum energy conformations (all pairs of equivalent atoms separated by less than 0.25 Å) were determined by least squares superposition of all non-hydrogen atoms and rejected. Chiral centres were checked for conservation of the original stereochemistry before saving conformations. An energy cut-off of 21 kJ/mol was applied. Minimisation’s were performed using the Truncated Newton Conjugate Gradient (TNCG) minimiser, allowing for 250 iterations per structure. A gradient of 0.01 kJ $Å^{-1}$ mol^{-1} was set as the initial convergence criterion. The local minimum energy conformations thus obtained were submitted to a final minimisation, using the Full Matrix Newton Raphson (FMNR) minimiser, allowing for line searching and 1000 iterations per structure. A gradient of 0.002 kJ $Å^{-1}$ mol^{-1} was set as the final convergence criterion.

ː Pharmacophore Identification

The pharmacophore-identifying program APOLLO[3] was used to align **59** and **60** in their presumed pharmacologically active conformations. The output files of the MC searches, containing multiple minimum energy conformations of each compound, served as input for the VECADD module of APOLLO. In each conformation of **59** and **60**, extension vectors from the carbonyl *O* atom and the basic nitrogen atom pointing towards putative receptor points, as well as a centroid and a normal through the aromatic ring were defined. In all cases, minimum densities of vectors were specified, representing ideal hydrogen bonding positions. The RMSFIT module was used to determine the conformation of each ligand which gave the best overall fit with respect to the specified fitting points. When receptor points emanating from the carbonyl oxygens were fitted, the two possible points were defined as choices. When fitting aromatic rings, both extremes of the normal and the centroid were defined as choices. All points were weighed equally. Conformational energies were taken into account for determining the root mean square (RMS) deviations. The RMS cut-off was set to 0.5 Å. The MMDFIT module was used to extract the conformations which gave the best matches.

ː Substituent Geometries and Ligand Alignment

The best fitting conformation of **59** identified by APOLLO was used as a starting point for building the compounds **1–58** in Table 6.1. The appropriate substituents were attached to the aromatic ring and the conformational space of substituents with conformational freedom was probed using the MC procedure in MacroModel as described above. For each newly introduced torsional angle 100 MC steps were performed. All other torsional angles were fixed during these MC searches in order to maintain the overall geometry of the pharmacophore pattern for all compounds. The lowest energy conformations resulting from these MC searches were used for the final alignment, which was performed within SYBYL 6.3.[2] The basic nitrogen atoms were protonated and centroids were defined in the aromatic rings. Then all compounds were superimposed with respect to the centroids, the carbonyl *O* atoms, the amide *H* atoms, and the protons on the protonated tertiary nitrogen atoms, using the *Fit Atoms* procedure as implemented in SYBYL.

ː The Data Set

The descriptors used in this paper were generated in the GRID program[18] from three different probe atoms, the C3 probe, the CA+2 probe and the OH2 probe, reflecting the steric field, the electrostatic field and the hydrogen-bonding field, respectively. The selection of probes was based on the PCA performed in Chapter 4.[10]

$$E_{tot} = E_{ste} + E_{ele} + E_{hb} \tag{6.1}$$

In contrast to SYBYL/CoMFA,[2] the same interactions (Equation 6.1) are calculated in each grid point independent of the type of probe atom. Thus, there are no separate steric or electrostatic fields generated but instead the type of probe will reflect different type of fields due to the fact that the relative size of the terms in Equation 6.1 differs from one probe to another. For example, a charged probe, *e.g.*, CA+2, reflects predominantly the electrostatics since the E_{ele} will be the dominating term, but still, the other two terms are not excluded. In section 1.5, molecular descriptors are discussed more thoroughly.

ː Multilinear PLS Analysis

In the previous chapter the multilinear PLS (N-PLS) algorithm[9] was scrutinised and implemented as regression method for 3D QSAR data.[8] Consequently, the derivation of the N-PLS algorithm is not repeated here, however, the same notations will be used also in this chapter.

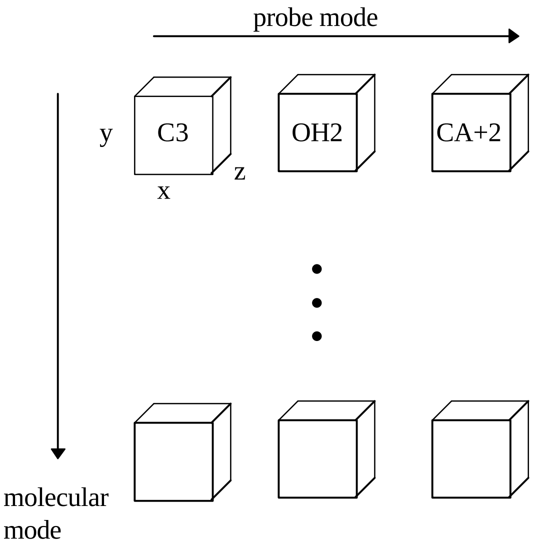


Figure 6.1 The complete data set ($\underline{\mathbf{X}}$) defining five directions. The molecular direction comprises 40 steps, the x direction 20 steps, the y direction 22 steps, the z direction 20 steps and the probe direction 3 steps.

The definition of this data set ($\underline{\mathbf{X}}$) is identical to the data set in Chapter 5, hence, the 58 compounds included are characterised with three different probes as described in Figure 6.1. Also here five different directions or modes can be defined: the molecular mode, the grid x mode, the grid y mode, the grid z mode and the probe mode. Consequently, in the decomposition of the multiway matrix ($\underline{\mathbf{X}}$) performed during the N-PLS analysis (see Figure 6.2(b) in the previous chapter) one score vector ($\mathbf{t}$) and four loading vectors ($\mathbf{w}^J$, $\mathbf{w}^K$, $\mathbf{w}^L$ and $\mathbf{w}^M$) are formed.

: Model Validation

The crossvalidation experiments and the external predictions were quantified with crossvalidated Q^2 and predicted Q^2, and the quality of the calibrations with R^2, respectively. The Q^2 and R^2 are already defined in Chapter 2.

In addition to the traditional ‘leave-one-out’ (LOO) crossvalidation also ‘leave-three-out’ (L3O) and ‘leave-five-out’ (L5O) crossvalidation were performed, where in each experiment the objects were left out randomly, but only once. The results were reported as the average Q^2 of 20 crossvalidation experiments.[19,20]

: Partial PLS Coefficients

The results from 3D QSAR studies are often presented as comprehensive iso-contour plots of the partial regression coefficients $\mathbf{b}_{\mathrm{PLS}}$ $(b_1,\ldots, b_R;\ R = JKLM)$[2,21] in Equation 6.2, where x_r is the rth grid point, *i.e.*, variable and b_r is the corresponding coefficient.

$$\hat{y} = b_1x_1 + \ldots + b_rx_r + \ldots + b_Rx_R \qquad (6.2)$$

Basically, the coefficients $\mathbf{b}_{\mathrm{PLS}}$ are needed for the predictions of the biological activity $\hat{y}$ of new molecules, but since the sizes and the signs of the coefficients reveal the relative influence of each grid point on $\mathbf{y}$, they are also suitable for the interpretation. That is, an external compound, not included in the training set, with a substituent protruding into a region with positive b_rs will produce a positive (repulsive) field x_r in this region and, consequently, have positive influence on $\hat{y}$. If the region has negative b_rs, however, the opposite is valid.

6.3 Results

: Conformational Analyses

The MC procedure identified 32 and 3 minimum energy conformations for **59** and **60**, respectively. The two lowest energy conformations of **60** were equal in energy, conformation 1 being identical to the minimised X-ray crystal structure, while in conformation 2 the ethyl side chain points in the opposite direction.

: Pharmacophore Identification

APOLLO identified conformation 6 of **59** ($\Delta E = 5.80$ kJ/mol) and conformation 2 ($\Delta E = 0.00$ kJ/mol) of **60** as best matches with respect to the indicated fitting points, with an RMS deviation of 0.26 Å (Figure 6.2). In conformation 6 of **59**, the (1-ethyl-2-pyrrolidinyl)methyl side chain adopts a half-folded conformation.

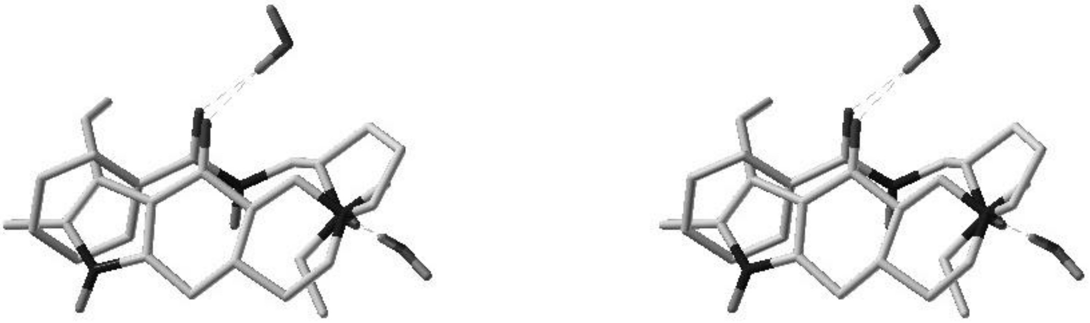


Figure 6.2 Stereo representation of the superposition of the best matching conformations of **59** (conformation 6) and **60** (conformation 2), as identified by APOLLO. For clarity purposes, alkyl hydrogens have been omitted. The water molecules mimic amino acid residues of the receptor, capable of forming hydrogen bonds with the ligands.

: Substituent Geometry’s and Ligand Alignment

The orientation of the 6-OMe group in the lowest energy conformations depended on the 5-substituent. When R_5 was *H*, OMe, OH or *F*, the 6-OMe group was oriented coplanar with respect to the aromatic ring. In all other cases, *i.e.*, R_5 being alkyl, *Cl*, *Br* or NO_2, the 6-OMe group adopted a conformation perpendicular to the plane of the aromatic ring. In both orientations an intramolecular hydrogen bond between the 6-OMe *O* atom and the amide *H* atom was formed. All alkyl substituents longer than methyl at the 3- and 5-positions adopted all-trans conformations perpendicular to the plane of the aromatic ring, while OMe and OH groups at these positions had a coplanar orientation. All OH substituents at the 2-position adopted coplanar conformations, forming an intramolecular hydrogen bond between the 2-OH *H* atom and the carbonyl *O* atom. The final alignment of compounds **1–58** is shown in Figure 6.3.

Figure 6.3 Stereo representation of the final alignment of compounds **1–58**. For clarity purposes, alkyl hydrogens have been omitted.

: The Data Set

The grid was created large enough to enclose all the aligned molecules with at least 4 Å in all directions, where the x, the y and the z direction were divided into 20 ($J = 20$), 22 ($K = 22$) and 20 ($L = 20$) parts with a step size of 1 Å, respectively. Three different probes ($M = 3$) were considered: the C3 probe, the OH2 probe and the CA+2 probe. The total data set consisted of 58 molecules and 26400 number of variables. As described above and in Figure 6.1, five different modes have been defined: the molecular direction, the grid x mode, the grid y mode, the grid z mode and finally the probe mode.

The 58 molecules were divided into a training set and a test set, selected with help from a Principal Component Analysis (PCA) of **X** (58 × 26400). Six main clusters were identified in the three first Principal Components describing 60 % of the variation in **X**. The number of compounds selected from each cluster depended on the size of the cluster. This resulted in 18 compounds which were selected for the test set (Table 6.1, compounds **41–58**) and the remaining 40 compounds comprised the training set (Table 6.1, compounds **1–40**).

: Multilinear PLS Analysis

Different data pre-processing methods have been discussed in the literature,[5] but the only data pre-processing applied on our data set was column mean-centering in the molecular direction.[8] Thus, the multiway matrix $\underline{\mathbf{X}}$ was unfolded, column mean-centred and, subsequently, 'back-folded' before the regression analysis was performed.

Table 6.2 Calibration[a] of the complete model Model I (40 × 26400).

#LV	R^2 ($\underline{X}$)	R^2 (y)	Q^2 (LOO)
1	5	47	29
2	11	59	44
3	14	69	52
4	17	75	56
5	20	79	56
6	**21**	**81**	**57**
7	22	85	56
8	24	87	62

[a] All values are expressed as percentages.

The number of significant N-PLS components was estimated by leave-one-out crossvalidation of the complete model, Model 1, and we found six N-PLS components (Table 6.2; Q^2 = 57 %) to be optimal. Important, in order not to lose information during the variable reduction step, we performed the variable reduction starting from a model one component more complex than optimal. Accordingly, the absolute sum of the weights from the first seven N-PLS components was calculated, each mode apart. A position in a direction was considered significant and selected only if it exceeded a lower cut-off value. An arbitrary cut-off value of 0.1, generated a reduced data set with 2940 number of variables, called Model 2. Stated differently, only variables with high weights from Model 1 were selected and included in Model 2. The probe mode was left intact, hence, variables from all three probes were included in the reduced data set. (One may argue why the variable selection not was performed from a model with eight N-PLS components having a Q^2 of 62 %. The rationale for this decision was that the absolute sum of the weights barely changed when going from seven to ten N-PLS components.)

Table 6.3 Calibration[a] and validation[b] of Model II (40 × 2940).

#LV	R^2 ($\underline{X}$)	R^2 (y)	Q^2 (LOO)	Q^2 (L3O)[c]	Q^2 (L5O)[c]	Q^2 (Pred)[d]
1	5	44	23	20	19	36
2	8	64	41	39	38	45
3	13	73	48	47	46	64
4	15	79	57	54	53	56
5	18	81	59	59	58	62
6	**21**	**87**	**63**	**62**	**62**	**62**
7	25	88	64	63	63	57
8	26	90	64	64	63	63
9	30	91	66	64	64	62
10	30	92	68	63	63	59

[a] All values are expressed as percentages.; [b] LOO stands for leave-one-out, L3O for leave-3-out and L5O for leave-5-out.; [c] Average from 20 crossvalidation experiments.; [d] Predictions of the external test set (18 × 2940).

Table 6.3 summarises the calibration and validation results, from the first ten N-PLS components of Model 2. The six first N-PLS components used 21 % of the variation in $\underline{\mathbf{X}}$ to explain 87 % of the variation in **y** and the Q^2s from the three crossvalidation experiments LOO, L3O and L5O were calculated to be 63 %, 62 % and 62 %, respectively. The calculated pIC_{50} values and the predicted pIC_{50} values after the sixth N-PLS component are tabulated in Table 6.1, and plotted against the experimental pIC_{50} values in Figures 6.4(a) and 6.4(b) for the training set and the test set, respectively.

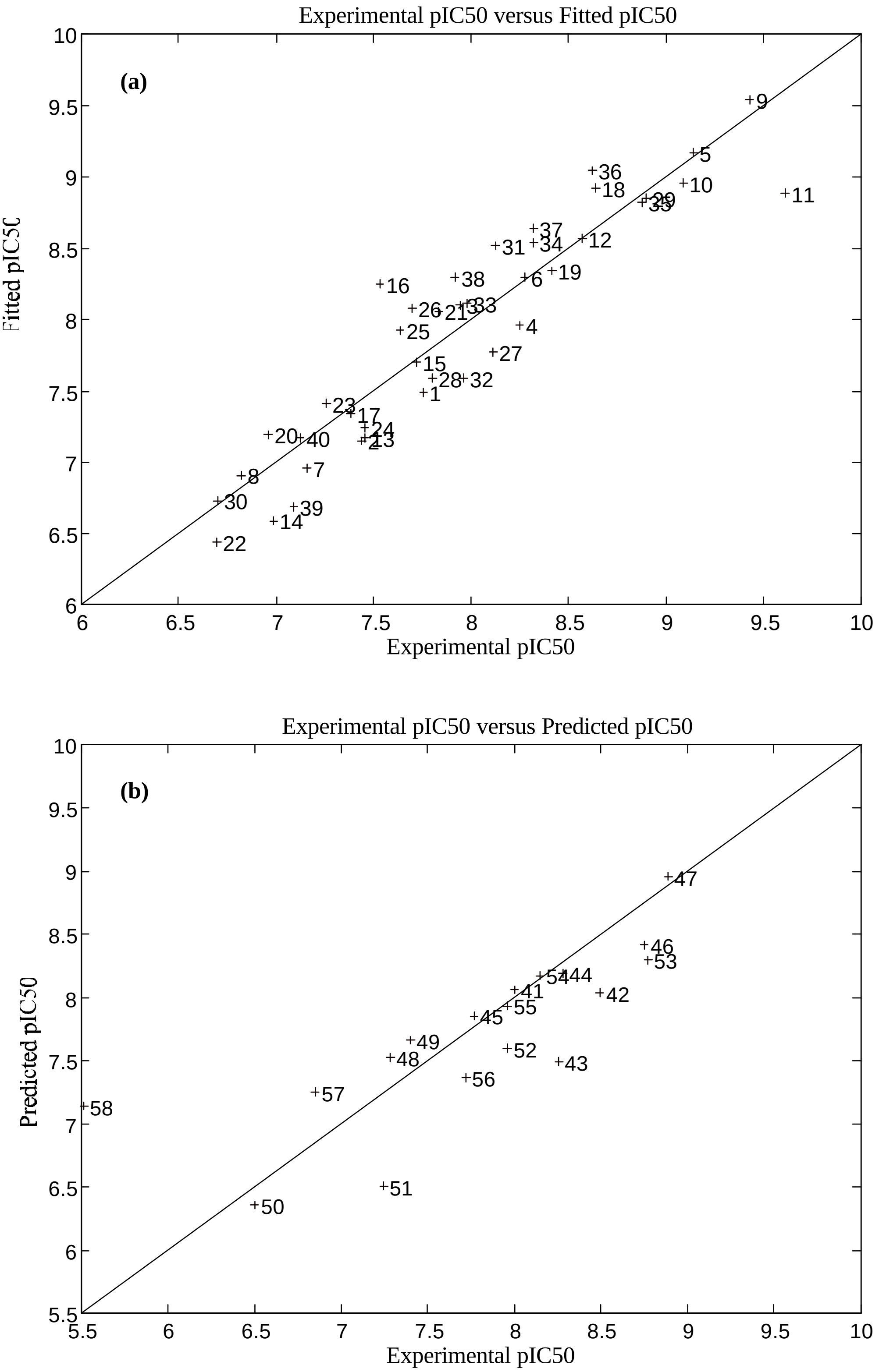


Figure 6.4 The performance of Model 2, showing in (a) Experimental pIC_{50} values versus Fitted pIC_{50} values and in (b) Experimental pIC_{50} values versus Predicted pIC_{50} values, after six N-PLS components.

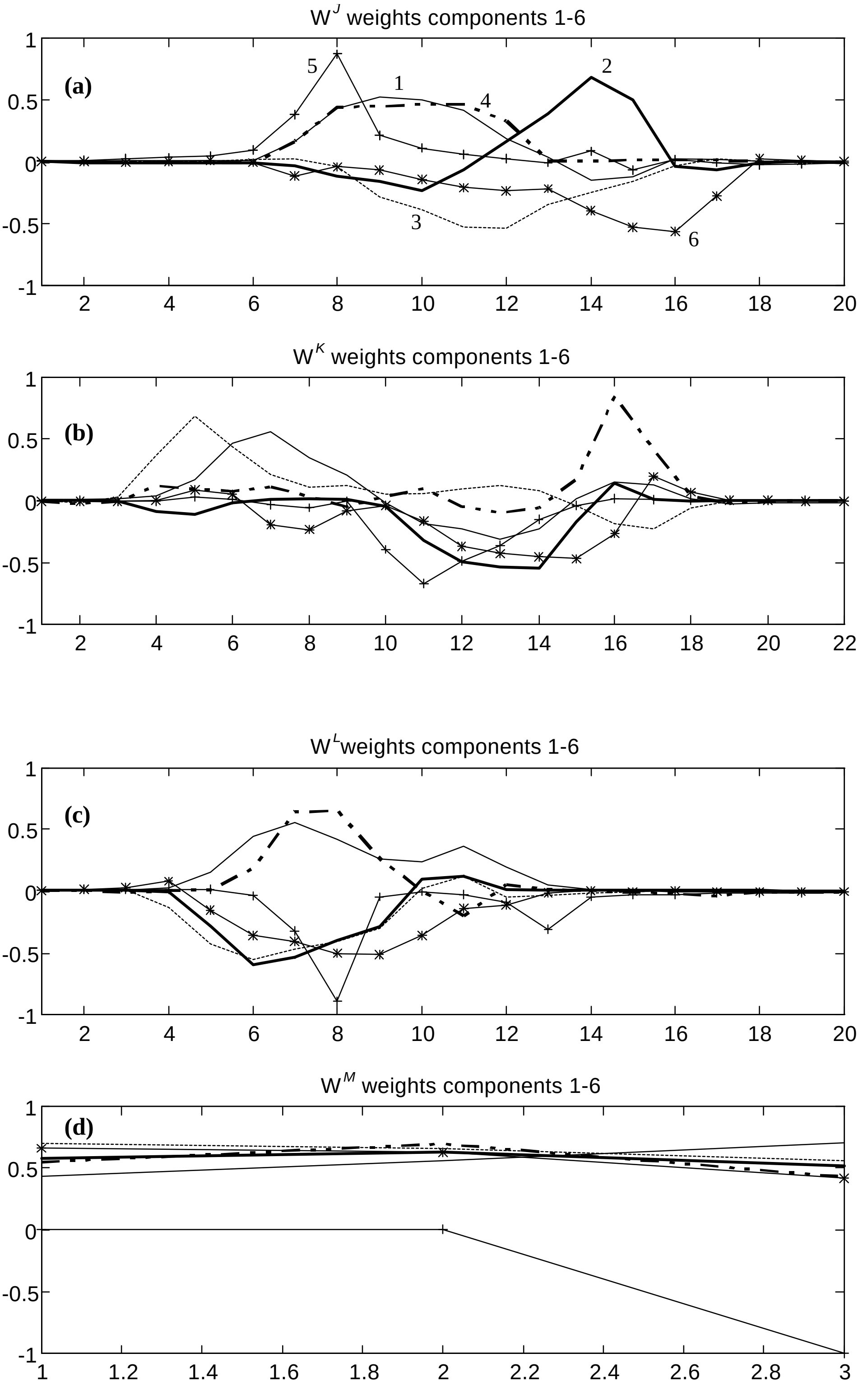


Figure 6.5 The weights from the first six N-PLS components of the complete model from (a) the x-mode, (b) the y-mode, (c) the z-mode and (d) the probe mode. N-PLS(1) is represented by a thin full line, N-PLS(2) a thick full line, N-PLS(3) a dotted line, N-PLS(4) a thick broken/dotted line, N-PLS(5) a + full line and N-PLS(6) a * full line.

Each subsequent component focuses on different variables, *i.e.*, different regions in the grid, which can be determined with help from Figure 6.5. The weights in Figures 6.5(a) and 6.5(b) correspond to positions in the x and y mode in Figure 6.6. For the first N-PLS component the weights in Figures 6.5(a) and 6.5(b) are high in positions 7–13 and 5–9, respectively, corresponding to the substituents on the benzamide 3-position (Figure 6.6, area 1). Analogously, N-PLS component two focuses on substituents on the benzamide 5-position (area 2), N-PLS component three on large substituents on the benzamide 3-position (area 3), N-PLS component four on the out-of-plane 6-OMe (area 4), N-PLS component five on the amid part and substituents on the benzamide 2-position (area 5), and N-PLS component six on large substituents on the benzamide 5-position (area 6). All components account for variation from all three probes except for the fifth, which accounts for variation only from the CA+2 probe (Figure 6.5(d)). In multilinear PLS consecutive components are not orthogonal which implies that some overlap of information may exist between different components. However, it has been shown[8] in multilinear PLS that each N-PLS component accounts for smaller easy to interpret regions while in bilinear PLS each PLS component accounts for more variation but not that well defined regions.

The partial PLS coefficients $\mathbf{b}_{PLS}$ (Equation 6.2) are presented as stereo iso-contour plots in Figures 6.7(a)–6.7(c) for the three probes after six N-PLS components.

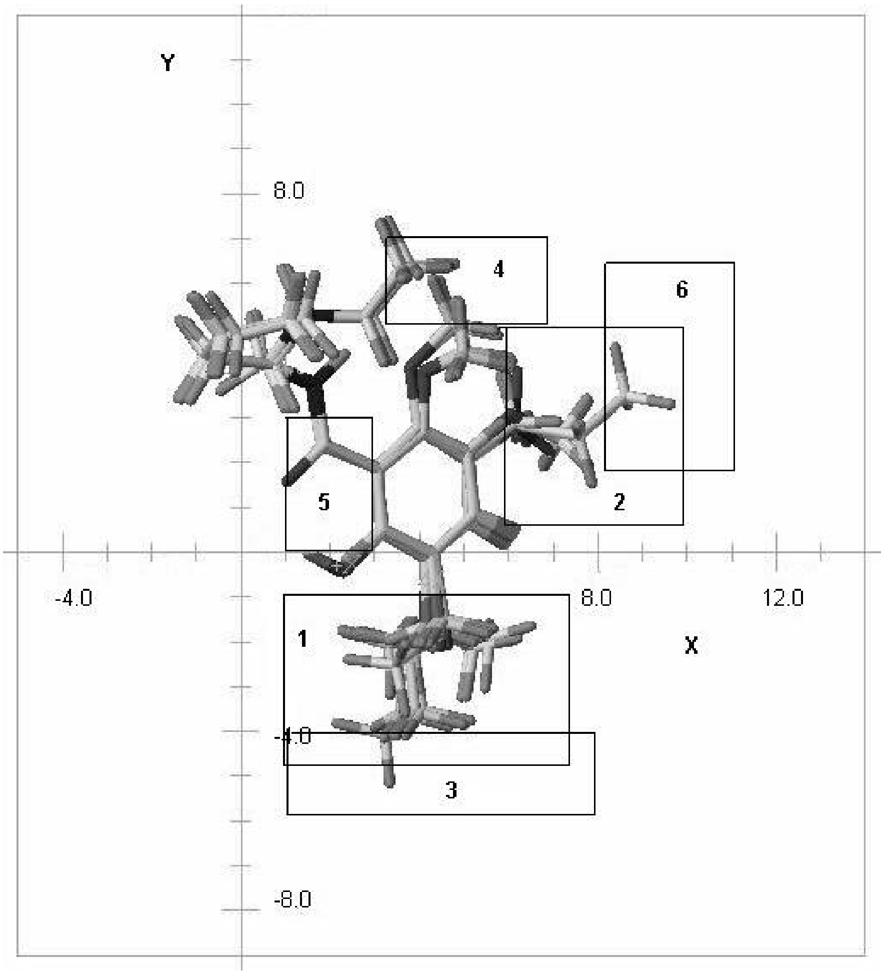


Figure 6.6 The 40 aligned training molecules, enclosed in the grid, viewed in the x and y mode.

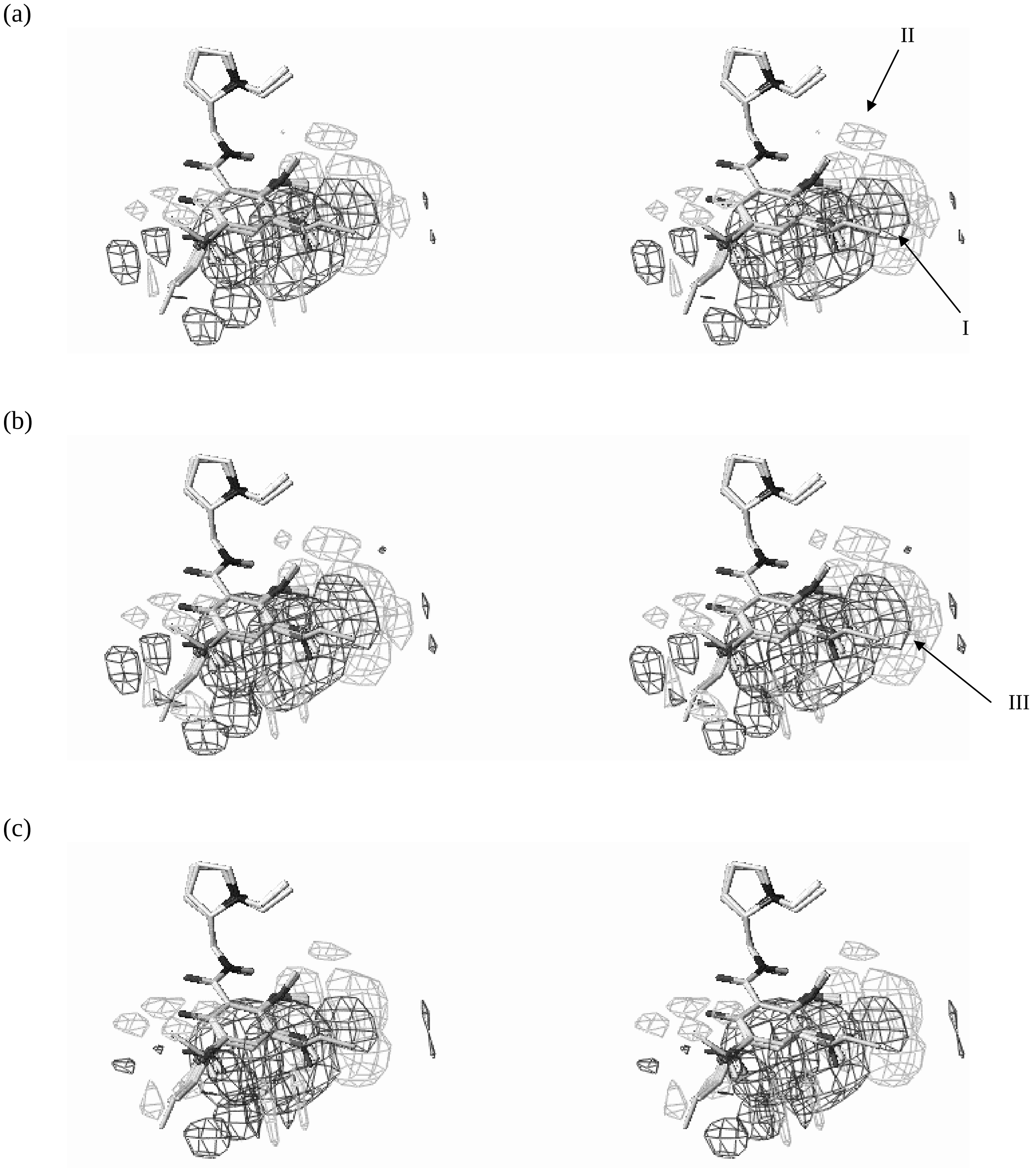


Figure 6.7 The $\mathbf{b}_{\mathrm{PLS}}$-coefficients from Model 2 after six N-PLS components, showing in (a) the C3 probe, in (b) the OH2 probe and in (c) the CA+2 probe on the |0.001| level. Regions with negative and positive $\mathbf{b}_{\mathrm{PLS}}$ coefficients are pictured with dark and light grey contours, respectively.

6.4 Discussion

The key issue in 3D QSAR modelling is to find a simple and straightforward model with high predictive ability. Traditionally, bilinear PLS is utilised as regression method but recently it has been shown that multilinear PLS[8-10] is more stable, simpler and offers improved predictability.

An important step in 3D QSAR analyses is the alignment of the ligands under investigation, *i.e.*, the relative positioning of the ligands in the fixed lattice, prior to the generation of the 3D descriptors. Even when the ligands possess a large degree of conformational flexibility, a single conformation has to be selected for each ligand. In order to be able to extrapolate the results of a 3D QSAR analysis in terms of receptor residues surrounding the ligands, we considered it essential to align the ligands in their pharmacologically relevant, *i.e.*, receptor binding conformations. Therefore, we employed the active analogue approach to determine the presumed pharmacologically active conformations of the ligands.[22] This approach is based on the assumption that ligands binding to the same binding site share the same pharmacophore pattern, *i.e.*, the three-dimensional arrangement of structural features essential for recognition by the receptor. The pharmacophore pattern of a flexible ligand can be determined by comparing it with that of a rigid analogue (template), in which the activity is retained. We chose (–)-piquindone as the template molecule. Although belonging to different chemical classes, (–)-piquindone and *N*-[(1-ethyl-2-pyrrolidinyl)methyl]-derived benzamides share several structural and pharmacological characteristics, a prerequisite when employing the active analogue approach. Thus, both classes of compounds contain an aromatic ring capable of π-π stacking with receptor residues, a carbonyl functionality, and a basic nitrogen atom at a certain distance from the aromatic ring, capable of forming hydrogen bonds with receptor residues. In addition, the binding to dopamine D_2 receptors is highly stereoselective and sodium-dependent for both classes, suggesting that they may share the same binding site and binding mode. This makes (–)-piquindone a suitable template for *N*-[(1-ethyl-2-pyrrolidinyl)methyl]-derived benzamides, as has been shown in the past.[13,23,24]

A thorough conformational analysis was performed on *N*-[(1-ethyl-2-pyrrolidinyl)methyl]benzamide (**59**), which constitutes the basic skeleton of the series. Although all compounds contain a 6-OMe substituent, this functionality was omitted from the basic skeleton, since its orientation was dependent on the adjacent substituent and thus not identical for all compounds. Conformation 6 (ΔE = 5.80 kJ/mol) of **59** was identified by APOLLO as fitting best on conformation 2 of (–)-piquindone with respect to the defined fitting points (Figure 6.2). In this conformation of **59** the *N*-[(1-ethyl-2-pyrrolidinyl)methyl side-chain adopts a half-folded conformation. This finding is consistent with previous reports,[13,23] although the conformation we identified is not exactly identical to previously reported conformations of the side-chain. This is probably the result of differences in the fitting procedures.

It is clear from Figures 6.5(a)–6.5(d), that positions corresponding to grid points, predominantly in the periphery of the grid, have low weights, and by omitting these variables, a reduced model (Model 2), with 2940 variables was obtained. Model 2 described slightly more of the variance in **y** (R^2 = 87 %) as compared to Model 1 (R^2 = 81 %), and the crossvalidated Q^2 (LOO) was slightly increased from 57 % to 63 % after six N-PLS components in Model 1 and 2, respectively. The stability of Model 2 was confirmed since the crossvalidated Q^2 (62 %) did not change, *i.e.*, it did not decrease when larger groups of compounds were left out each time (Table 6.3). Important, additional components did not significantly increase (nor decrease) the crossvalidated Q^2. Therefore, in order to keep the number of components at a minimum and avoid insignificant variation, we used six N-PLS components for Model 2.

The role of the crossvalidated Q^2 is to act as a descriptor for the predicted Q^2, which it often does too optimistically.[8] In this model, however, the crossvalidated Q^2 (62 %) is in good agreement with the predicted Q^2 (62 %). Knowing that the model is stable and possesses high predictability, we may interpret the iso-contour plots in Figures 6.7(a)–6.7(c).

From this model we can conclude that the conformation of the 6-OMe group has a great influence on the pIC_{50} value. In a planar conformation, the 6-OMe group protrudes in a region (I) with negative (dark grey) C3 PLS coefficients (Figure 6.7(a)) and, consequently, grid points within this region will have repulsive (positive) interactions with the 6-OMe group. Thus, 6-OMe group in a planar conformation affects the pIC_{50} value negatively. For compounds with the 6-OMe group in a perpendicular conformation the 6-OMe group protrudes in a positive (light grey) C3 PLS coefficient region (II), which favours a high pIC_{50} value (*i.e.*, high affinity). At the same time, hydrogen bonding (Figure 6.7(b)) with the 6-OMe oxygen becomes favourable in region III, which promotes a high pIC_{50} value. From the CA+2 fields (Figure 6.7(c)) it is also clear that the substituents in positions 5 and 6 must possess negative electrostatic potential in order to promote a high pIC_{50} value. These are the most obvious conclusions that can be drawn from Figures 6.7(a)–6.7(c).

The substituents at the 2- and 3-position have less influence on the pIC_{50} value as compared to substituents at the 5- and 6-position.[25,26] However, a substituent with positive electrostatic potential in the 3-position is favourable. The size of the substituent is of minor importance. Also, important hydrogen bonding sites are, obviously, not present in the vicinity of positions 2 and 3. This is consistent with results reported by Norinder *et al.*[12]

The importance of the basic nitrogen on the pyrrolidinyl moiety has already been established elsewhere,[24] but is not recognised in our model since the variance in the GRID-descriptors in this region is very low.

6.5 Conclusions

We conclude that 3D QSAR modelling with multilinear PLS works very well and definitely is an alternative to the traditional bilinear PLS method.

We have demonstrated the importance of a proper strategy for the conformational analyses and the alignment procedure in order to succeed in 3D QSAR modelling. We found that the conformations of the ligands, to a great extent, explained the difference in pIC_{50} values of these ligands. Further, interpretations of our model are in line with what others have reported and confirm the reliability of our model.[11,12]

We used multilinear PLS to reduce the number of variables by omitting variables with low weights, *i.e.*, only variables with variation correlated with the pIC_{50} values were considered in the final model (Model 2). The crossvalidated Q^2 for Model 2 was 62 %, independent of how many groups were left out each time. This result confirms that our data set is homogenous and that our model is stable. Normally,[10] leave-one-out crossvalidation produces higher Q^2, as compared to when larger groups of compounds are left out each time. The crossvalidated Q^2 (62 %) from this model is a perfect estimation of the predicted Q^2, which also was found to be 62 %. This is most certainly an effect originating from the multilinear PLS[9] method, which has been discussed by Nilsson *et al.*[8] and Smilde.[21]

6.6 References

1. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analyses (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
2. SYBYL- Molecular Modeling Software, *6.3*, Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA,
3. APOLLO - Automated PharmacOphore Location through Ligand Overlap, Koehler, K.F. The APOLLO program available upon request from Konrad F. Koehler (koehler@irbm.it)
4. Jones, G.; Willet, P.; Glen, R.C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput. -Aided Mol. Design* **1995**, *9*, 532-549.
5. GOLPE, *3.0*, Clementi, S. Multivariate Infometric Analyses(MIA), Perugia, Italy.
6. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct. -Act. Relat.* **1993**, *12*, 9-20.
7. Norinder, U. Single Domain Mode Variable Selection in 3D QSAR Applications. *J. of Chemometrics* **1996**, *10*, 95-105.
8. Nilsson, J.; De Jong, S.; Smilde, A.K. Multiway Calibration in 3D QSAR. *J. of Chemometrics* **1997**, *11*, 511-524.
9. Bro, R. Multiway Calibration. Multilinear PLS. *J. of Chemometrics* **1996**, *10*, 47-61.
10. Nilsson, J.; Wikström, H.; Smilde, A.K.; Glase, S.; Pugsley, T.A.; Cruciani, G.; Pastor, M.; Clementi, S. A GRID/GOLPE 3D-QSAR Study on a Set of Benzamides and Naphthamides, with Affinity for the Dopamine D_3 Receptor Subtype. *J. Med. Chem.* **1997**, *40*, 833-840.
11. Norinder, U. A PLS QSAR Analysis Using 3D Generated Aromatic Descriptors of Principal Property Type: Application to Some Dopamine D_2 Benzamide Antagonists. *J. Comput. -Aided Mol. Design* **1993**, *7*, 671-682.
12. Norinder, U. and Högberg, T. A Quantitative Structure-Activity Relationship for Some Dopamine D_2 Antagonists of Benzamide Type. *Acta Pharm. Nord.* **1992**, *4*, 73-78.
13. Högberg, T. Novel Substituted Salicylamides and Benzamides as Selective D_2-Receptor Antagonists. *Drugs Fut.* **1991**, *16*, 333-357.
14. Marshall, G.R.; Barry, C.D.; Bosshard, H.E.; Dammkoehler, R.A.; Dunn, D.A. The Conformational Parameter in Drug Design: the Active Analogue Approach. *ACS Symp. Ser.* **1979**, *112*, 205-226.
15. Jansen, J.M.; Copinga, S.; Gruppen, G.; Molinari, E.J.; Dubocovich, M.L.; Grol, C.J. The High Affinty Melatonin Binding Site Probed with Conformationally Restricted Ligands-I. Pharmacophore and Minireceptor Models. *Bioorg. &Med. Chem.* **1996**, *4*, 1321-1332.
16. Mohamadi, F.; Richards, N.G.J.; Guida, W.C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W.C. MacroModel-An integrated Software System for Modelling Organic and Bioorganic Molecules Using Molecular Mechanics. *J. Comp. Chem.* **1990**, *11*, 440-467.
17. Olson, G.L.; Cheung, H.-C.; Morgan, K.D.; Blount, J.F.; Todaro, L.; Berger, L.; Davidson, A.B.; Boff, B. A Dopamine Receptor Model and Its Applications in the Design of a New Class of Rigid Pyrrolo[2,3-g]isoquinoline Antipsychotics. *J. Med. Chem* **1981**, *24*, 1026-1034.
18. GRID, Goodford, P.J. Molecular Discovery Ltd, University of Oxford, England.
19. Cruciani, G.; Baroni, M.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part 1: Standard Deviation of Prediction Errors (SDEP). *J. of Chemometrics* **1992**, *6*, 335-346.
20. Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model. *J. of Chemometrics* **1992**, *6*, 347-356.
21. Smilde, A.K. Comments on Multilinear PLS. *J. of Chemometrics* **1997**, *11*, 367-377.
22. Jansen, J.M. PhD Thesis. Three Dimensions in Drug Design. Probing the Melatonin Receptor Dept. of Medicinal Chemistry, University of Groningen, The Netherlands. **1995**
23. Högberg, T.; Rämsby, S.; Ögren, S.0.; Norinder, U. New Selective Dopamine D_2 Antagonists as Antipsychotic Agents. *Acta Pharm. Suecia* **1987**, *24*, 289-328.

24. Rognan, D.; Sokoloff, P.; Mann, A.; Martres, M.P.; Schwartz, J.C.; Costentin, J.; Wermuth, C.G. Optically Active Benzamides as Predictive Tools for Mapping the Dopamine D_2 Receptor. *Eur. J. Pharmacol* **1990**, *189*, 59-70.
25. De Paulis, T.; Tayar, N.A.; Carrupt, P.-A.; Testa, B.; Van de Waterbeemd, H. Quantitative Structure-Affinity Relationships of Dopamine D_2 Receptor Antagonists: A Comparison between Orthopramides and 6-Methoxysalicylamides. *Helv. Chim. Acta* **1991**, *74*, 241-254.
26. De Paulis, T.; Kumar, Y.; Johansson, L.; Rämsby, S.; Hall, H.; Sallemark, M.; Angeby Möller, K.; Ögren, S.O. Potential Neuroleptic Agents. 4. Chemistry, Behavioral Pharmacology, and Inhibition of [^{3}H]-Spiperone Binding of 3,5-Disubstituted *N*-[(1-ethyl-2-pyrrolidinyl)methyl]-6-methoxysalicylamides. *J. Med. Chem.* **1986**, *29*, 61-69.

7

Multiway Simultaneous Two-Block Analysis with Applications to 3D QSAR

This chapter is based mainly on the article: Nilsson, J; Kiers, H.A.L.; Smilde, A. K. Multiway Simultaneous Two-Block Analysis with Applications to 3D QSAR. *In preparation.*

Summary

In this chapter, the algorithms and the procedures for the analyses with the multiway simultaneous two-block methods, PCovR/Tucker5 and PCovR/PARAFAC, have been scrutinized and applied to the 3D QSAR data set analyzed in Chapters 4 and 5.

Since the methods are based on alternating least squares algorithms the stability of the obtained best models were estimated by means of repeated calculations using different starting parameters. The predictability of the models were estimated with predictions of external test sets.

The most predictive PCovR/Tucker5 model (Q^2 = 31 %) was obtained using s_r = 100, a = 0.4 with (6, 7, 4, 3 and 2) number of components in the different modes. This model was found stable with high reproducibility which is in sharp contrast to the best PCovR/PARAFAC model which was unstable with poor reproducibility, obtained with s_r = 1000, a = 0.7 and seven components.

The a value and the number of components have direct effects on the predictability of the models while the size of s_r only controls at which a value the most predictive model will occur.

The 3D QSAR data set from Chapters 4 and 5 has been analyzed with several different methods and the best predictabilities were obtained using the PCovR/Tucker5 and the N-PLS methods (Q^2 = 31 %).

7.1 Introduction

In QSAR, the typical problem is to create models from theoretically generated molecular descriptors in order to make predictions of biological responses possible. In the previous chapters, QSAR and 3D QSAR data sets have been dealt with and they all comprise two blocks of descriptors, *i.e.*, one independent and one dependent descriptor block. This situation is pertinent to most problems in regression analysis. Until recently, it was assumed that the variable blocks were organized in two-way matrices, simply due to that no algorithms existed that could handle data sets of higher orders than two. Thus, in order to analyze a data set of higher order, the multiway matrix must first be unfolded[1] into a two-way matrix, normally leaving the object mode intact. MLR,[2] PCR[3] and PLS[3] are examples of regression methods designed for the analysis of two-way data sets.

Wold *et al.*[4] analyzed multiway data sets by combining the Lohmöller-Wold decomposition[4] with the non-linear partial least squares (NIPALS)[5,6] algorithm. This method was called multiway

PLS, which solution is the two-way PLS solution of the unfolded matrix. Later, Ståhle[7] developed an algorithm for linear three-way decomposition (LTD) of three-way data sets, composed of one independent and one dependent variable block. Ståhle's algorithm is a multiway generalization of the two-block PLS algorithm (see Chapter 2).

Recently, Bro developed the multilinear PLS algorithm[8] (N-PLS) which combines PLS and the PARAFAC[9,10] decomposition method (see Section 2.8). The N-PLS algorithm generates *partial* least squares solutions, with score vectors that have maximum covariance with **y**.

All the models obtained with the above mentioned algorithms are calculated in a component-wise fashion, *i.e.*, parameters from one component are calculated from the residuals of the previous component. Consequently, the variation accounted for by an early component is more significant for the description of **y**, as compared with subsequently extracted components. However, alternative approaches have been developed, *i.e.*, algorithms that are not nested and where all components are calculated simultaneously.[11,12] In a paper by De Jong and Kiers,[12] principal covariate regression (PCovR) was presented as a method that simultaneously minimizes the **X** and **Y** residuals by means of an alternating least squares (ALS) algorithm. The original PCovR algorithm was defined for the case when **X** and **Y** were two-way matrices. Recently, Smilde[13] postulated frameworks for the possibility of extending the algorithm also for multiway $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$, *e.g.*, PCovR/PARAFAC and PCovR/Tucker. In the present chapter the multiway simultaneous two-block regression algorithms will be discussed and applied to 3D QSAR data. This investigation is mainly focused on the calibration procedure, *e.g.*, the data pretreatment, the optimization and the validation of the most successful models. Finally, the performance of the different regression methods investigated in this thesis, *i.e.*, PLS (Chapter 5), N-PLS (Chapter 5), PCovR/PARAFAC and PCovR/Tucker5 (this chapter), will be compared and discussed.

7.2 Theory and Methods

The PCovR Algorithms

The theory of principal covariate regression has already been discussed in Section 2.9 and, therefore, only the most important features of the five-way methods are repeated here. In PCovR, the data are fitted to the following model:

$$\mathbf{T} = \mathbf{XW} \tag{7.1}$$

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} + \mathbf{E}_{\mathrm{X}} \tag{7.2}$$

$$\mathbf{y} = \mathbf{Tb} + \mathbf{e}_{\mathrm{y}} \tag{7.3}$$

$$\min\left[\alpha\left\|\mathbf{X} - \mathbf{XWP}^{\mathrm{T}}\right\|^{2} + (1-\alpha)\left\|\mathbf{y} - \mathbf{XWb}\right\|^{2}\right] \tag{7.4}$$

where **T** ($I \times A$) contains the A score vectors (principal covariates $\mathbf{t}_{\mathrm{i}}$) and **W** ($JKLM \times A$) the weights. **P** ($JKLM \times A$) and **b** ($A \times 1$) are the regression parameters relating **X** ($I \times JKLM$) and **y** ($I \times 1$), respectively, with the scores in **T** (see Figure 7.1). The algorithm is balanced between reconstructing $\underline{\mathbf{X}}$ ($\alpha\to 1$) and fitting $\mathbf{y}$ to the data ($\alpha\to 0$). Obviously, an important step in PCovR modeling is to find the α that maximizes, *e.g.*, the predictability.

Smilde[13] defined a number of different models by simply imposing different structures on $\mathbf{P}$, *i.e.*, constraints. The five-way simultaneous two-block PARAFAC model (PCovR/PARAFAC) is obtained when

$$\mathbf{P}^{\mathrm{T}} = \left(\mathbf{W}_{\mathrm{M}}^{\mathrm{T}} \otimes \mathbf{W}_{\mathrm{L}}^{\mathrm{T}} \otimes \mathbf{W}_{\mathrm{K}}^{\mathrm{T}} \otimes \mathbf{W}_{\mathrm{J}}^{\mathrm{T}}\right) \tag{7.5}$$

where $\mathbf{W}_{\mathrm{J}}$ ($J \times A$), $\mathbf{W}_{\mathrm{K}}$ ($K \times A$), $\mathbf{W}_{\mathrm{L}}$ ($L \times A$) and $\mathbf{W}_{\mathrm{M}}$ ($M \times A$) are the loading matrices. The same number of components (A) are used in all five modes. (The five modes are defined graphically in Figure 5.3.)

In analogy, the simultaneous two-block Tucker5 model (PCovR/Tucker5) is obtained when

$$\mathbf{P}^{\mathrm{T}} = \mathbf{G}\left(\mathbf{W}_{\mathrm{M}}^{\mathrm{T}} \otimes \mathbf{W}_{\mathrm{L}}^{\mathrm{T}} \otimes \mathbf{W}_{\mathrm{K}}^{\mathrm{T}} \otimes \mathbf{W}_{\mathrm{J}}^{\mathrm{T}}\right) \tag{7.6}$$

where $\mathbf{G}$ ($N \times PQRS$) is the properly concatenated core matrix and $\mathbf{W}_{\mathrm{J}}$ ($J \times P$), $\mathbf{W}_{\mathrm{K}}$ ($K \times Q$), $\mathbf{W}_{\mathrm{L}}$ ($L \times R$) and $\mathbf{W}_{\mathrm{M}}$ ($M \times S$) are the loading matrices from each mode. In contrast to the PARAFAC model,[9,10,14] the corresponding Tucker model[15] may have different number of components in each mode. The number of components in the molecular, x, y, z, and probe modes for the Tucker5 model are given as N, P, Q, R and S, respectively. The PARAFAC model uses A components in each mode.

In the PARAFAC model, the core matrix $\underline{\mathbf{G}}$ is a matrix where all off-superdiagonal elements are zero, since no interactions between modes from different components are allowed. In a Tucker model, however, this type of interactions are accounted for by the off-superdiagonal elements in $\underline{\mathbf{G}}$. Hence, the PARAFAC model is a special case of the Tucker model (*e.g.*, more constrained). Additionally, each rotation of the core matrix or of the loading matrices, from a Tucker model, may give solutions with the same sum of squared residuals as the model found. Hence, the Tucker model does not have a unique solution. The general PCovR model is presented graphically in Figure 7.1.

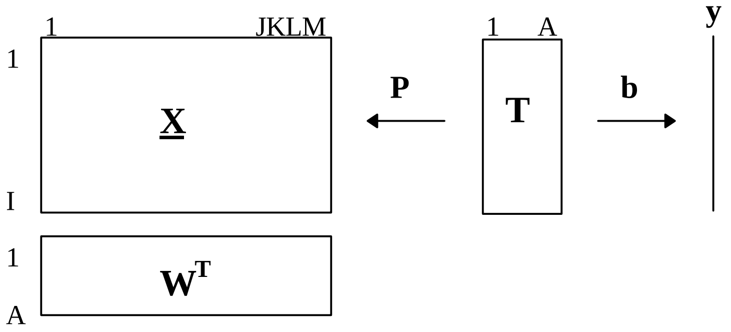


Figure 7.1 A graphical representation of PCovR, where $\underline{\mathbf{X}}$ ($I \times J \times K \times L \times M$) is a five-way matrix and $\mathbf{y}$ ($I \times 1$) is univariate.

Since the PCovR algorithm is solved by an alternating least squares (ALS) algorithm, the initial starting parameters have to be selected *a priori*. Accordingly, the parameters, *e.g.*, loadings, scores, $\mathbf{P}$ and $\mathbf{b}$ are updated alternately until convergence. It is a well known fact that in non-linear modeling an algorithm may occasionally converge into a local minimum, since the result depend on the starting parameters used. Unfortunately, there is no easy way to determine whether the minimum value found is a local or the global minimum function value. However, the chance of finding the global minimum value is increased if several calculations, with different starting parameters, are attempted.

Predictions of the external test sets (see below) were carried out as described in Section 2.9.

: Definition of the Data Set

In order to enable comparison of the present analysis methods with, *e.g.*, PLS and N-PLS, the data set from Chapter 5 was analyzed also in the present chapter. The data set comprises 30 training and 21 test molecules with affinity for the dopamine D_3 receptor subtype. The molecules were all modeled, aligned and characterized as described in Chapters 4 and 5.

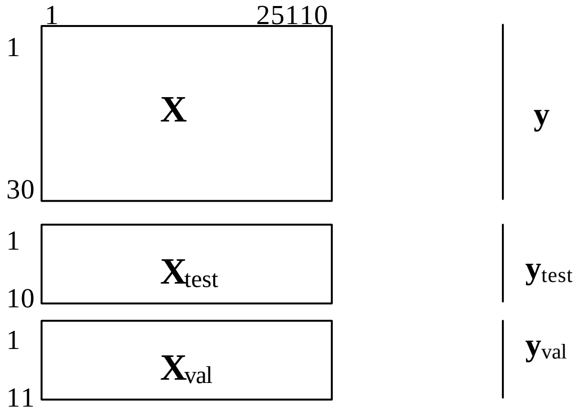


Figure 7.2 The unfolded data set used in the present application comprises a training set $\mathbf{X}$ (30 × 25110), a test set $\mathbf{X}_{test}$ (10 × 25110) and a validation set $\mathbf{X}_{val}$ (11 × 25110).

Traditionally, crossvalidation[16] is used to estimate the predictability of a model. In the present application, however, the test set from Chapter 5 was divided into a test set ($\underline{\mathbf{X}}_{test}$) and a validation set ($\underline{\mathbf{X}}_{val}$) as described in Figure 7.2. In order to make sure that both the test set and the validation set spanned the same descriptor space as the training set ($\underline{\mathbf{X}}$), a PCA of the 30 training compounds merged with the 21 test compounds (51 × 25110) was performed. From the first two score vectors, accounting for 38 % of the variation, ten compounds **(t1**, **t4**, **t6**, **t8**, **t12–t16** and **t21** in Tables 5.1–5.3) were, arbitrarily, selected for the test set denoted $\underline{\mathbf{X}}_{test}$ (10 × 25110). The remaining eleven compounds (**t2**, **t3**, **t5**, **t7**, **t9–t11**, **t17–t20** in Tables 5.1–5.3) comprised the validation set denoted $\underline{\mathbf{X}}_{val}$ (11 × 25110). Accordingly, the ability of a model to predict $\underline{\mathbf{X}}_{test}$ was used during the calibration procedure for the selection of the best model, while the prediction of $\underline{\mathbf{X}}_{val}$ was used to estimate the predictability of the final model. This procedure was preferred, since crossvalidation would be a far too time consuming procedure considering the large number of models that need to be validated.

: Data Pretreatment

Previously, in Chapter 5 where the present data set was analyzed with PLS and N-PLS, the only pretreatment applied was column mean-centering in the direction of the molecular mode. As soon will become evident, the α value with which the best predictability is obtained can be adjusted with the ratio of the sum of squares of $\underline{\mathbf{X}}$ and $\mathbf{y}$ (s_r):

$$s_r = \mathrm{SSQ}(\underline{\mathbf{X}})/\mathrm{SSQ}(\mathbf{y}) \tag{7.8}$$

For reasons of convenience, the s_r was adjusted such that the best predictability was obtained using an α value as close as possible to 0.5.

: Determination of the Number of Components

In contrast to the PARAFAC model, where the same number of components are used in all modes, a Tucker model may have different number of components in all modes. With PCovR/PARAFAC modeling, the number of components can be estimated with crossvalidation[16] or by predictions of an external test set (this chapter). With PCovR/Tucker modeling, the estimation of the optimal number of components is not that straightforward. There is one approach available, however, based on singular value decompositions (SVD) of the unfolded five-way matrix $\underline{\mathbf{X}}$,[17] that can be unfolded in five different ways[18] each time leaving one mode intact. (Note that $\underline{\mathbf{X}}$ was mean-centered, as described in the previous section, prior to the unfolding procedure.) Accordingly, the number of singular values from each unfolded matrix that accounts for, *e.g.*, 30–70 % of the variation, determines the number of components to be considered in the different modes. The number of singular values obtained are presented in Table 7.1. In the fifth mode, 50 % of the variation was explained already by the first singular value.

Table 7.1 The number of singular values, obtained from SVDs of $\underline{\mathbf{X}}$ unfolded in five different directions explaining 30–70 % of the variation in $\underline{\mathbf{X}}$.

mode	30 %	40 %	50 %	60 %	70 %
molecular	3	4	6	8	11
x	4	5	7	10	13
y	2	3	3	4	5
z	2	3	4	5	7
probe			1	1	2

In the following procedure where the optimal Tucker5 model is searched for, the columns in Table 7.1 will represent the combination of components to be used in the different calculations. Just as the number of components in PARAFAC models may be chosen between, *e.g.*, three to eight, the number of components with the Tucker5 models will be chosen between 30 % and 70 % according to Table 7.1, in the optimization procedures below.

7.3 PCovR/Tucker5 Analysis

The analysis with the PCovR/Tucker5 method comprised the following sequence: First, the sum of squares of the descriptor blocks, *i.e.*, the s_r, that produce the best predictability with an α value near 0.5, was searched for. Second, using this s_r the combination of α and number of components that produced the highest predictive Q^2 (predictions of $\underline{\mathbf{X}}_{\text{test}}$) was determined. Finally, the stability of the optimized model was investigated with repeated calculations using different starting parameters and the predictability was estimated by prediction of the validation set ($\underline{\mathbf{X}}_{\text{val}}$).

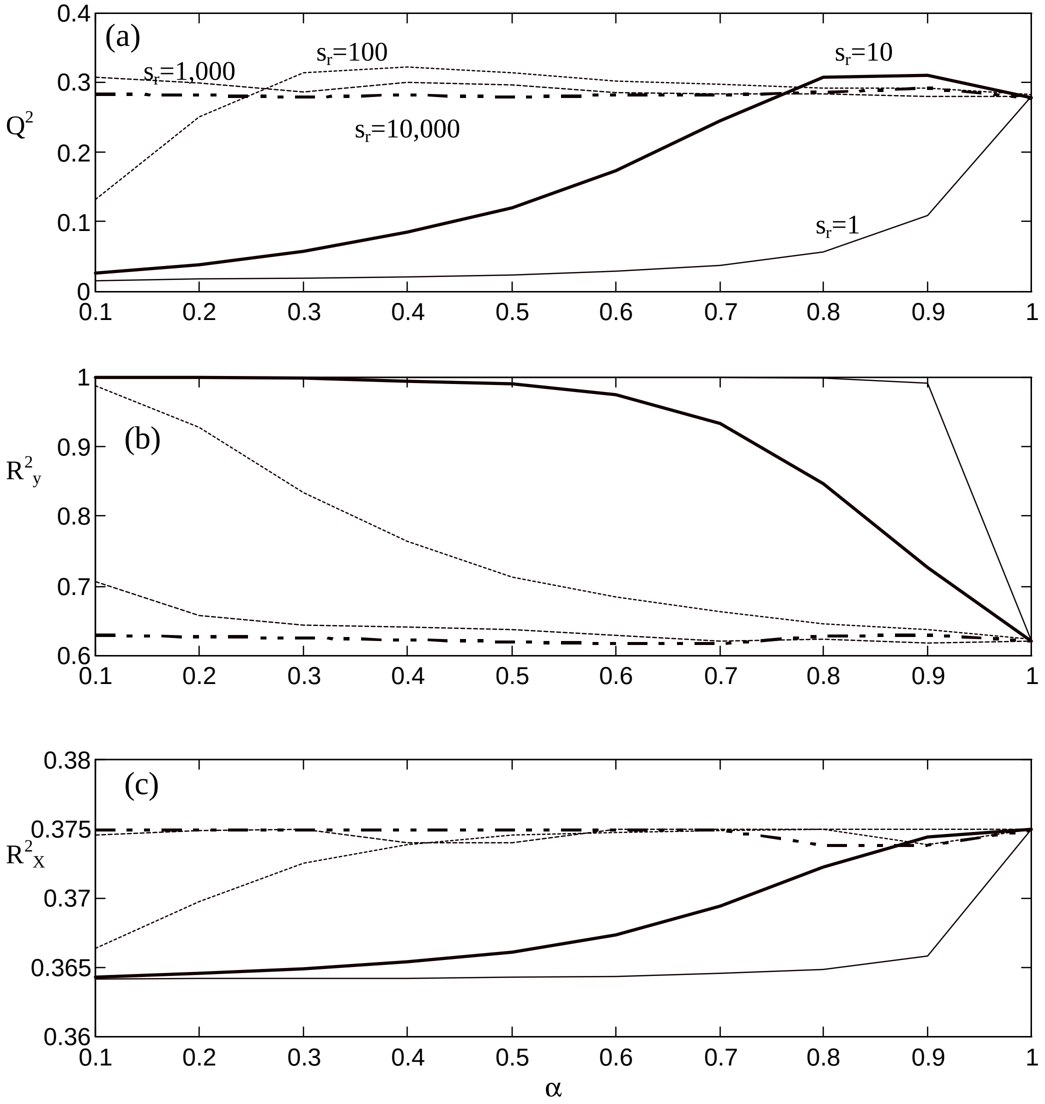


Figure 7.3 (a) The effect of different ratios (s_r) of the sum of squares of $\underline{\mathbf{X}}$ and $\mathbf{y}$ on the predictability. In (b) and (c) the R^2_y and R^2_X, respectively, from the corresponding models are presented. The number of components corresponds to the 50 % column in Table 7.1, using two components in the probe mode. (thin full line: s_r = 1; thick full line: s_r = 10; dotted line: s_r = 100; broken line: s_r = 1,000; thick broken/dotted line: s_r = 10,000)

: Determination of the Scaling Ratio (s_r)

It became evident during the preparation of this work that the sum of squares of $\underline{\mathbf{X}}$ and $\mathbf{y}$ have a significant impact on the size of α that produces the best predictability. Accordingly, in a number of calculations an optimal s_r was searched for by evaluating all combinations of α (0.1–1.0) and s_r (1; 10 ;100; 1,000 and 10,000). In each calculation, the number of components corresponded to the 50 % column in Table 7.1, using two components in the probe mode. Since the number of components will be optimized first in the next step, the number of components used corresponds to the 50 % column in Table 7.1. The performance of the obtained models are summarized in Figure 7.3. Clearly, when the s_r was increased (SSQ($\underline{\mathbf{X}}$) > SSQ($\mathbf{y}$)) the optimal predictability was obtained at a decreasing size of α. The best predicted Q^2s (predictions of $\underline{\mathbf{X}}_{test}$) obtained with s_r adjusted to 1, 10 ,100 were found at the α values 0.99, 0.8 and 0.4, respectively. (The calculations with $\alpha = 0.99$ are not presented here.) The question is whether the best models obtained with different s_r really are different models since the predicted Q^2 were identical in all models? Nevertheless, it was decided to proceed and optimize α and the number of components using $s_r = 100$.

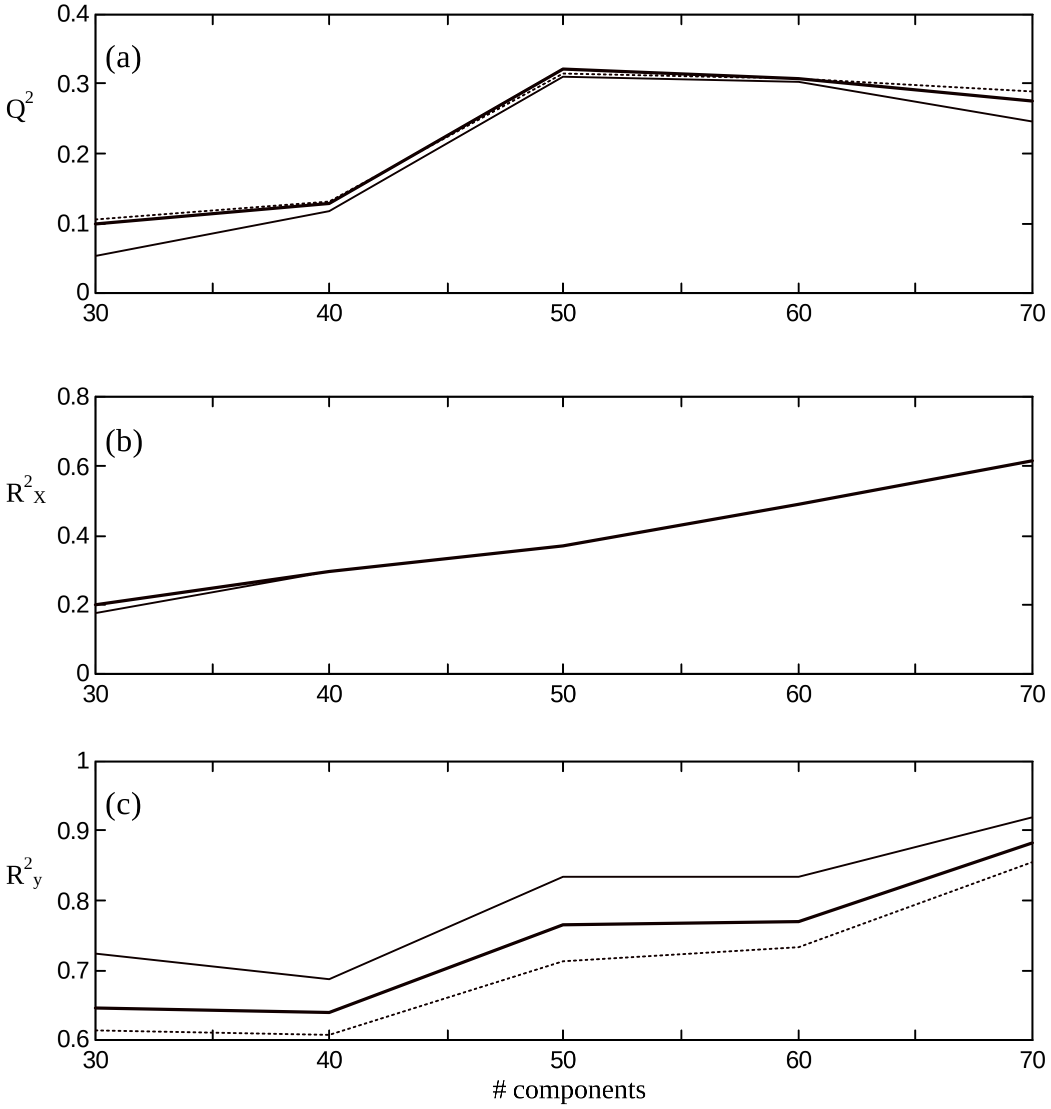


Figure 7.4 Determination of the optimal α using $s_r = 100$. For each of the α values 0.3, 0.4 and 0.5 models were calculated using the number of components corresponding to the 30, 40, 50, 60 and 70 % columns in Table 7.1 with two components in the probe mode. (thin full line: $\alpha = 0.3$; thick full line: $\alpha = 0.4$; dotted line: $\alpha = 0.5$)

: Optimization of α

The combination of α and number of components was optimized using $s_r = 100$ (Equation 7.5), as was decided in the previous section. Since the best predictability, in Figure 7.3(a), was found when α = 0.4 it seemed reasonable to assume that the optimal model must be close to this α value. Accordingly, in a number of calculations all the combinations of the α values, 0.3, 0.4 and 0.5, and the number of components corresponding to the 30, 40, 50, 60 and 70 % columns in Table 7.1 were evaluated. The results are summarized in Figure 7.4 revealing that a model calculated with α = 0.4 and the number of components corresponding to the 50 % column in Table 7.1 produced the highest Q^2 (0.32; Figure 7.4(a)). Although the predictabilities, obtained with the different α values in Figure 7.4(a), probably are not significantly different, the model with α = 0.4 was selected as the final model.

: Reproducibility and Predictability

It is always a risk that iterative methods, like the PCovR methods, converge into local minima with sometimes spurious solutions as the result. If several models with different starting parameters are calculated, however, the chance of finding the global minimum value increases. Accordingly, in order to investigate the reproducibility of the final PCovR/Tucker5 model, it was recalculated 100 times with different starting parameters. In Figure 7.5, the R_y^2, R_X^2, Q^2 and the total loss (**f**; see Equation 7.4) values from the 100 models are compared. The f values are scaled with 0.01 in order to fit into the scale used in Figure 7.5.

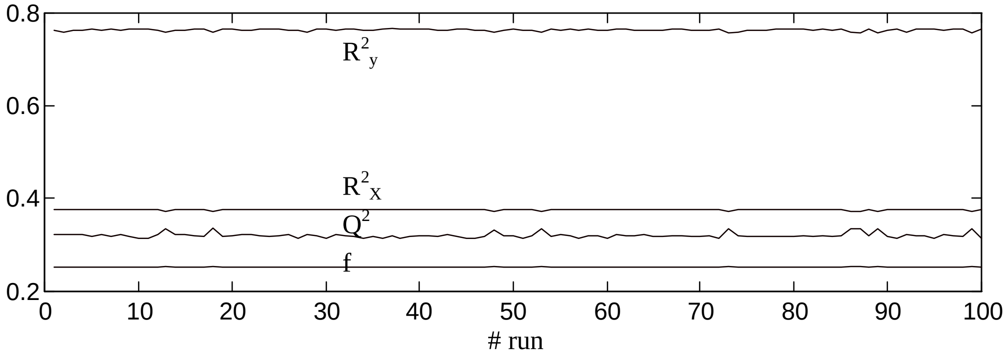


Figure 7.5 The final model (α = 0.4; $s_r = 100$; # lv = 50 % with two components in the probe mode) calculated 100 times with different starting parameters. In order to plot the total function loss values (**f**) in the same plot it was necessary to scale **f** with 0.01.

The predictability of the final model was estimated by means of prediction of the validation set ($\underline{\mathbf{X}}_{val}$). Five models, calculated with different starting parameters, all predicted $\underline{\mathbf{X}}_{val}$ identically with $Q^2s = 0.31$. The test set $\underline{\mathbf{X}}_{test}$ was predicted with $Q^2s = 0.32$. Obviously, the chance that the PCovR/Tucker5 model converges into a local minimum is low, although probably not negligible. In conclusion, the PCovR/Tucker5 model is stable, reproducible and possesses a relative high predictability.

7.4 PCovR/PARAFAC Analysis

Since the analysis with the PCovR/Tucker method worked well a similar protocol was followed also for the analysis with the PCovR/PARAFAC method (see beginning of Section 7.3).

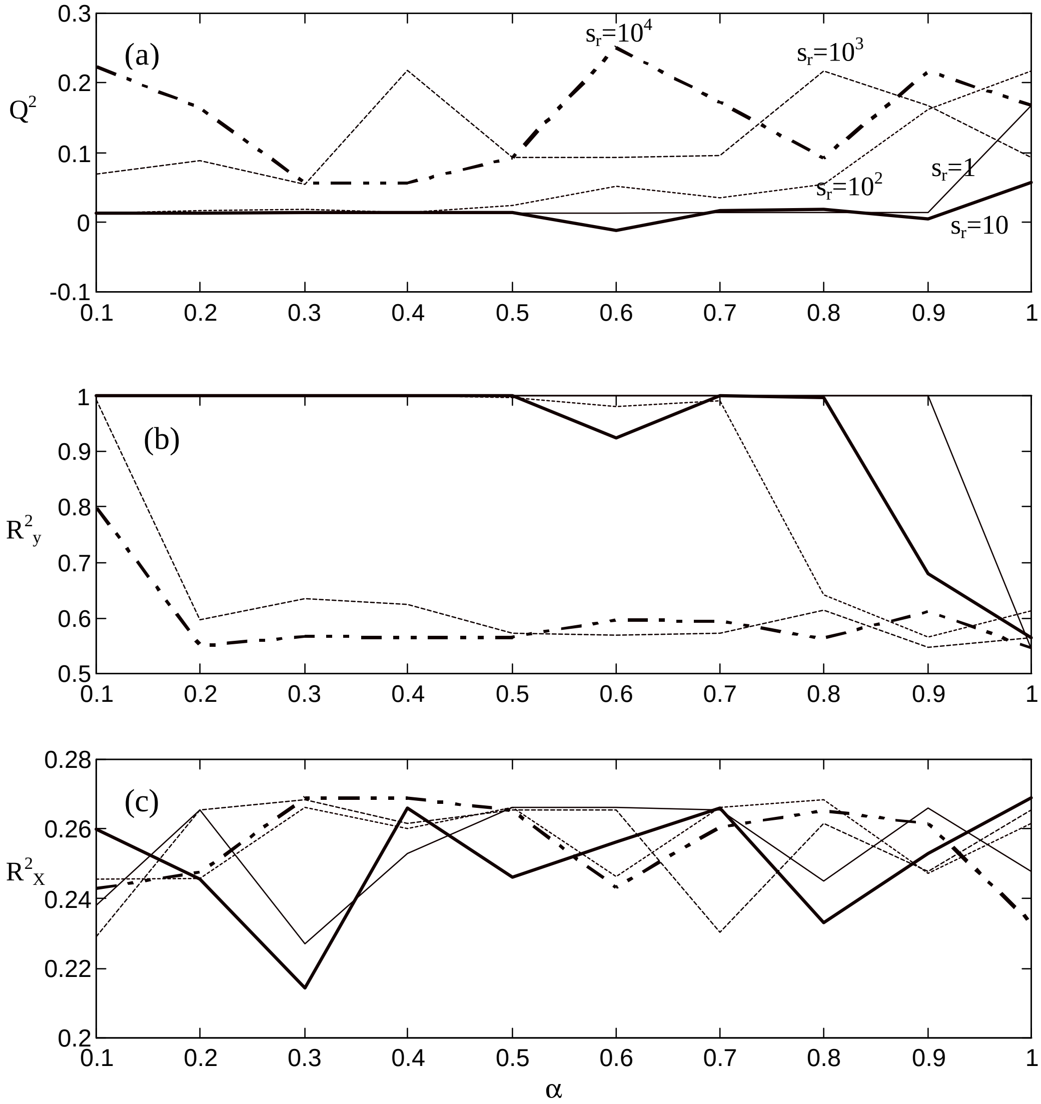


Figure 7.6 (a) The effect of different s_r on the predictability. In (b) and (c) the R^2_y and R^2_X, respectively, from the corresponding models are presented. Five components were used in all calculations. (thin full line: s_r = 1; thick full line: s_r = 10; dotted line: s_r = 100; broken line: s_r = 1,000; thick broken/dotted line: s_r = 10,000)

Determination of the Scaling Ratio (s_r)

In a series of calculations the optimal s_r was searched for by evaluating all combinations of s_r (1; 10; 100; 1,000 and 10,000) and α (0.1–1.0). In all calculations five components were considered. The objective with these calculations was to find a s_r such that the optimal predictability was found with an α value near 0.5. As was pointed out, this will probably not produce models with higher predictability but may help to improve the stability of the models. In Figure 7.6(a), each curve represents different s_r and in contrast to the corresponding plot from the Pcovr/Tucker5 method (Figure 7.3(a)), a clear optimum in predictability is hard to find. It is clear from Figure 7.6, however, that when s_r is small (*e.g.*, 1, 10 and 100) the predictability remains low as long as the R_y^2 is 100 %. However, when α approaches unity the R_y^2 starts to decrease with increased predictability as the result.

For models with larger s_r (*e.g.*, 1,000 and 10,000) no trend in the predictability (Q^2 simply fluctuates between 0.1 and 0.25) is observed and the R_y^2 remains stable at approximately 0.6 for models with $\alpha > 0.2$. Interestingly, when α is < 0.2 and approaches zero the R_y^2 increases but no change in the predictability is observed. Preliminary results with unscaled data (corresponding to a $s_r \approx 26{,}000$) showed that optimal predictability was obtained when α was ≈ 0.06 (six components).

Taken together, high predictability is not observed when R_y^2 is close to unity; predictive models with large s_r (> 10,000) are observed at low α while predictive models with small s_r (< 100) are obtained when α is large. Consequently, it was decided to proceed and optimize α and the number of components with $s_r = 1{,}000$ with the intention of finding the most predictive model with an α value somewhere in the range between 0.4 and 0.8.

The predictability of the models with s_r = 1,000 and 10,000 in Figure 7.6(a) displayed large variation which may be an indication of instability caused by the different starting parameters used. This issue will be dealt with later in this chapter.

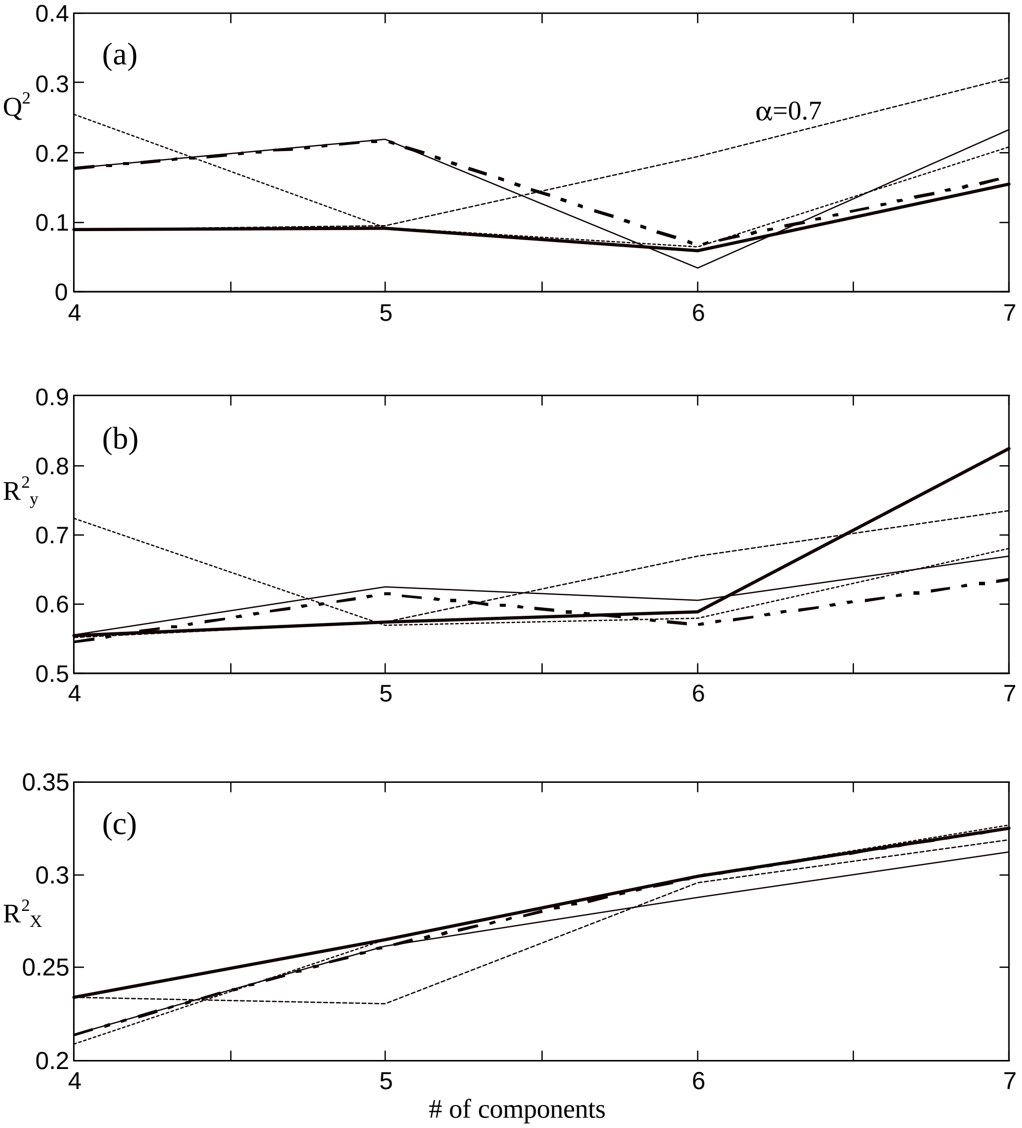


Figure 7.7 Simultaneous optimization of α and number of components using s_r = 1,000. For each of the αs 0.4, 0.5, 0.6, 0.7 and 0.8 models were calculated using 4, 5, 6 and 7 components. (thin full line: α = 0.4; thick full line: α = 0.5; dotted line: α = 0.6; broken line: α = 0.7; thick broken/dotted line: α = 0.8)

: Optimization of α

The most predictive model was searched for by evaluating all the combinations of α values (*i.e.*, 0.4, 0.5, 0.6, 0.7 and 0.8) and number of components (*i.e.*, 4, 5, 6 and 7) using s_r = 1,000. The results from the calculations are summarized in Figure 7.7, where each curve represents different α values. The selection of one best model from Figure 7.7(a) is difficult. However, the model obtained with α = 0.7 and seven components produced the highest predicted Q^2 and was, consequently, selected and considered the best PCovR/PARAFAC model.

Reproducibility and Predictability

The reproducibility of the PCovR/PARAFAC method was elucidated by repeating the calculation 70 times, using $\alpha = 0.7$, $s_r = 1{,}000$ and six components, each time with different starting parameters. Since the computational time necessary for each calculation increases significantly with each additional component, only six components were considered here. In order make extrapolation of these result to the model with seven components possible, it was assumed that the variation in the predicted Q^2 was independent of the number of components considered. In Figure 7.8, the R_y^2, R_X^2, Q^2 and the total loss values (f; see Equation 7.4) from the 70 models are plotted. It was necessary to scale the f values with 0.001 in order to fit into the scale used in Figure 7.8.

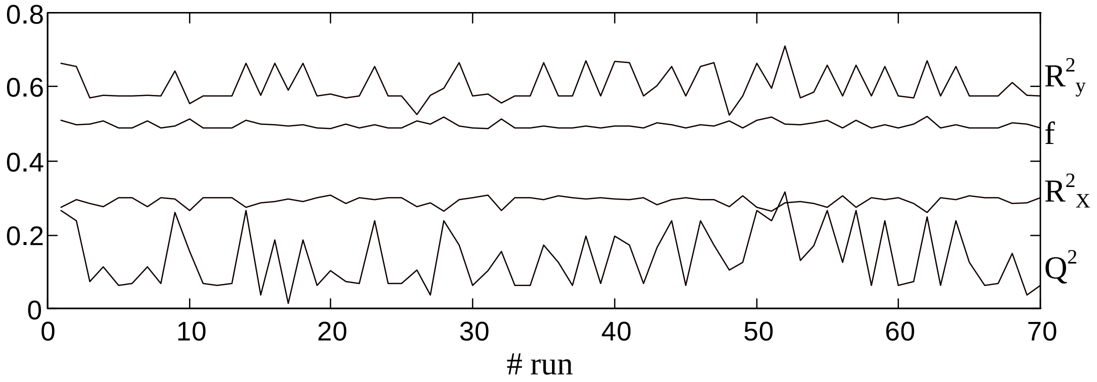


Figure 7.8 The model with $\alpha = 0.7$, $s_r = 1{,}000$ and six components was recalculated 70 times in order to investigate the reproducibility of the PCovR/PARAFAC method. (Note **f** is scaled by 0.001 in order to fit into the scale used)

From Figure 7.8 it is clear that the PCovR/PARAFAC method displays large variation in the predicted Q^2 and it can be concluded that the method is unstable with low reproducibility.

The predictability of the selected best model was estimated by means of prediction of the validation set ($\underline{\mathbf{X}}_{val}$). The calculation was repeated five times with different starting parameters and the obtained results are presented in Table 7.2.

Table 7.2 The predictability of the best PCovR/PARAFAC model ($\alpha = 0.7$; $s_r = 1{,}000$ and # lv = 7). The model was recalculated five times with different starting parameters and Q^2 was obtained from predictions of the validation set ($\underline{\mathbf{X}}_{val}$).

# run	R_X^2	R_y^2	f[a]	Q^2
1	0.33	0.67	472	0.17
2	0.31	0.73	485	0.27
3	0.32	0.67	480	0.30
4	0.31	0.58	481	0.12
5	0.34	0.58	464	0.16

[a] the total loss function value (Equation 7.4)

These five calculations confirm the instability of the method since all models converged to different minima. It may be argued that some of the models in Table 7.2 were not allowed to converge properly, but calculations with a more narrow criteria for convergence just increased the computational time while no improvements in the models were observed. Another disturbing observation is that the model with the best predictability ($Q^2 = 0.30$; run number three in Table 7.2) was not the global minimum model which may influence the stability negatively and decrease the reliability of the model.

7.5 Discussion

The PCovR/Tucker5 and the PCovR/PARAFAC Methods

The simultaneous multiway regression methods discussed in the present chapter, comprise three parameters, *i.e.*, s_r, α and the number of components, that need to be optimized in order to obtain models with high predictability. Initially, $\underline{\mathbf{X}}$ and $\mathbf{y}$ were normalized, *i.e.*, $\mathrm{SSQ}(\underline{\mathbf{X}}) = \mathrm{SSQ}(\mathbf{y}) = 1$, resulting in that most predictive models were obtained using α values close to unity. In addition, the calculations with PCovR/PARAFAC often converged into local minima with spurious solutions as the result. Simply, the obtained models were unstable and unreliable. The PCovR/Tucker5 models, on the contrary, were found stable and the most predictive model was found at a slightly lower α (0.99) than with the PCovR/PARAFAC method (0.999).

In calculations when α is chosen close to unity, the emphasis lies more on reconstructing $\underline{\mathbf{X}}$ rather than explaining $\mathbf{y}$ and since one objective in QSAR modeling is to explain $\mathbf{y}$, lower α values are desired. One way of obtaining high Q^2s at lower α values is by balancing the sum of squares of $\underline{\mathbf{X}}$ and $\mathbf{y}$, as was performed in Figures 7.3 and 7.6. The consequence of increasing s_r was clear, at least with the PCovR/Tucker5 method: the optimal predictability was found at lower α values. However, this was probably due to an algebraic effect rather than that different models were obtained. For instance, the model obtained with $\alpha = 0.8$ and $s_r = 10$, in Figure 7.3(a), most certainly corresponds to the model obtained when $\alpha = 0.3$ and $s_r = 100$. Consequently, s_r was not considered as a parameter that needed to be optimized simultaneously with the α and the number of components.

Apparently, the problem with unstable PCovR/PARAFAC models could not be solved by increasing the s_r, as can be concluded from the calculations in Figure 7.8. The 75 models calculated with different starting parameters display high variation in the predicted Q^2. However, it may be that the superior stability of the PCovR/Tucker5 method is due to this particular data set, and not an artifact of the PCovR/PARAFAC method. It has been suggested,[17] that a multiway data set that displays different numbers of singular values in the different modes when treated as described in Table 7.1, has a Tucker structure. According to this definition, the present data set has a Tucker structure and may explain the superior performance of the PCovR/Tucker5 method over the PCovR/PARAFAC method.

Although it is unclear whether the size of s_r improves the stability of the PCovR models, it is convenient to adjust the size in order to make the analysis more lucid. For clarity, it appears not to be necessary nor essential to optimize the size of s_r in order to achieve maximal predictability, at least as far as this investigation shows.

Due to the low stability of the PCovR/PARAFAC models, selection of one optimal model was difficult. Nevertheless, the model with $s_r = 1000$, $\alpha = 0.7$ and seven components (Figure 7.7(a)) was selected and recalculated five times with different starting parameters. None of the five calculations (Table 7.2) converged to the same function value (f). This is in sharp contrast to the best PCovR/Tucker5 model obtained when $s_r = 100$, $\alpha = 0.4$ and the number of components corresponding to the 50 % column in Table 7.1 (*i.e.*, 6, 7, 3, 4 and 2 components in the respective modes), which is very stable and portrays high predictability ($Q^2 = 0.31$ from predictions of $\underline{\mathbf{X}}_{\mathrm{val}}$).

The calculations performed in this chapter reveal the overall superior performance of the PCovR/Tucker5 method as compared with the PCovR/PARAFAC method. At this point it is important to stress that this conclusion applies to the present investigation and with another data set the performance of the PCovR/PARAFAC method may be better.

Multiway Analysis in 3D QSAR

The PCovR methods scrutinized in this chapter are the last methods introduced and enable comparison of four different regression analysis methods, *i.e.*, PLS (Chapter 5), N-PLS (Chapter 5), PCovR/PARAFAC (Chapter 7) and PCovR/Tucker5 (Chapter 7). These methods originate in two different classes of methods, *i.e.*, component-wise and simultaneous regression methods, where PLS and N-PLS represent the former class, PCovR/PARAFAC and PCovR/Tucker5 the latter class. Additionally, PLS is a two-way regression method while the remaining three methods are designed for the analysis of multiway data sets. It is difficult to determine, without any *a priori* knowledge of a data set, which analysis method is the better choice. In Table 7.3, the best models obtained with the different methods are compared.

The predictability of models obtained using multiway methods, *e.g.*, N-PLS and PCovR/Tucker5 ($Q^2 = 0.31$), are better than the two-way method, *e.g.*, PLS ($Q^2 = 0.26$). As was speculated in Chapter 5, an explanation for the better predictability of the N-PLS model, as compared to the PLS model, is the lower number of parameters that need to be estimated (see Table 7.3). This would also explain the lower fit ($R_X^2 = 0.17$) of the N-PLS model after four components as compared with a one component PLS model ($R_X^2 = 0.22$). For comparison, after four components PLS accounts for 53 % of the variation in **X** (Table 5.5). The same reasoning applies also to the simultaneous methods although they portray better fits than both the N-PLS and the PLS models. The PCovR/Tucker5 model is better fitted than the PCovR/PARAFAC model which could be due to the larger number of parameters estimated. If the better fits of the PCovR models are due to the larger number of components used or if they are effects of the methods, is not clear. According to De Jong and

Kiers,[12] it is to be expected that, occasionally, both R_X^2 and R_y^2 will be higher when using PCovR since these are the parameters that the algorithm actually optimizes.

Table 7.3 Comparison of the best models obtained with PLS (Chapter 5), N-PLS (Chapter 5), PCovR/Tucker5 and PCovR/PARAFAC (this chapter).

method	s_r[a]	# lv[b]	α	R_X^2 [c]	R_y^2 [d]	Q^2 [e]	# par.[f]
PLS	–	1	–	0.22	0.62	0.26	25140
N-PLS	–	4	–	0.17	0.73	0.31	388
PCovR/PARAFAC	1000	7	0.7	0.32	0.67	0.30	679
PCovR/Tucker5	100	(6,7,3,4,2)[g]	0.4	0.37	0.76	0.31	1528

[a] $s_r = SSQ(\underline{\mathbf{X}})/SSQ(\mathbf{y})$; [b] Number of latent variables; [c] Percentage variation accounted for in $\underline{\mathbf{X}}$; [d] Explained percentage of **y**; [e] Predictions of external validation sets: for PLS and N-PLS 21 compounds, for PCovR/PARAFAC and PCovR/Tucker5 eleven compounds; [f] the total number of parameters estimated after lv components; [g] see 50 % column in Table 7.1.

Interestingly, the performance of the PARAFAC related methods, *i.e.*, N-PLS and PCovR/PARAFAC, is very different. The PCovR/PARAFAC models are unstable with low reproducibility while the N-PLS models appears to be very stable, have high predictability and are easy to interpret. It was speculated above that the performance of the PCovR/PARAFAC method was poor since the present data set possible has a Tucker structure. To date, no further explanation is available.

7.6 Conclusions

The procedure for the analysis with the PCovR methods are more complex than the analysis with PLS and N-PLS. More parameters, *i.e.*, s_r, α and the number of components, need to be optimized in order to obtain optimal predictability. By balancing the sum of squares of the descriptor blocks (s_r) the size of α that produces the most predictive models can be tuned. Although the size of s_r has no clear influence on the predictability it is convenient not to use s_r close to unity or zero.

The size of α and the number of components directly affects the predictability and the quality of the obtained models. Therefore, these parameters were optimized simultaneously using an appropriate size of s_r. The best PCovR/Tucker5 model was obtained using $s_r = 100$, $\alpha = 0.4$ and (6, 7, 4, 3 and 2) components in the five modes. It predicted the validation set with a Q^2 of 0.31. The corresponding PCovR/PARAFAC model was obtained with $s_r = 1000$ and $\alpha = 0.7$ using seven components, but this model was unstable with poor reproducibility.

The structure of the data appears to influence the performance of different regression methods. The present data set possibly has a Tucker structure, which may contributed to the poor performance of the PCovR/PARAFAC method. The PCovR/Tucker5 method performed excellently with stable, reproducible and predictive models as the result. These conclusions can not explain the apparent excellent performance of the N-PLS method which also is a PARAFAC method.

The predictability of the best PCovR/Tucker5 model is comparable with the best N-PLS model in Chapter 5. The interpretation of the N-PLS models are straightforward while the interpretation of the PCovR models needs to be scrutinized further.

7.5 References

1. Esbenssen, K. and Wold, S. Proc. Symp. Applied Statistics. **1983**; pp.11-36.
2. Myers, R.H. Classical and Modern Regression with Applications. PWS-KENT Publishing Company: **1997**
3. Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons: New York, **1989**
4. Wold, S.; Geladi, P.; Esbenssen, K.; Öhman, J. Multi-Way Principal Components- and PLS-analyses. *J. of Chemometrics* **1987**, *1,* 41-56.
5. Wold, H. Research Papers in Statistics. David, F. Ed.; Wiley: New York, **1966**; pp. 411-444.
6. Wold, H. Perspectives in Probability and Statistics. Soft Modeling by Latent Variable, the Non-Linear Iterative Partial Least Squares (NIPALS) Algorithm. Gani, J. Ed. Academic Press.: London, **1975**; pp. 117-142.
7. Staehle, L. Aspects of the Analyses of Three-Way Data. *Chemom. and Intell. Lab. Syst.* **1989**, *7,* 95-100.
8. Bro, R. Multiway Calibration. Multilinear PLS. *J. of Chemometrics* **1996**, *10,* 47-61.
9. Carroll, J. and Chang, J.J. Analysis of Individual Differences in Multidimensional Scaling with an N-way Generalisation of the Eckart-Young Decomposition. *Psycometrika* **1970**, *35,* 283-319.
10. Harshman, R.A. Foundations of the PARAFAC Procedure: Models and Conditions for an Exploratory Multimodal Factor Analysis. *UCLA Working Papers in Phonetics* **1970**, *16,* 1-84.
11. Brooks, R. and Stone, M. Joint Continuum Regression for Multiple Predictands. *J. Am. Stat. Assoc.* **1994**, *89,* 1374-1377.
12. De Jong, S. and Kiers, H.A.L. Principal Covariates Regression Part I: Theory. *Chemom. and Intell. Lab. Syst.* **1992**, *14,* 155-164.
13. Smilde, A.K. Comments on Multilinear PLS. *J. of Chemometrics* **1997**, *11,* 367-377.
14. Smilde, A.K. Three-way Analyses. Problems and Prospects. *Chemom. and Intell. Lab. Syst.* **1992**, *15,* 143-157.
15. Tucker, L. Problems of Measuring Change. Implications of Factor Analysis of Three-Way Matrices for Measurement of Change. Harris, C. Ed. University of Wisconsin Press.: Madison, **1963**; pp. 122-137.
16. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185,* 1-17.
17. Tucker, L. Some Mathematical Notes on Three-Mode Factor Analysis. *Psycometrika* **1966**, *31,* 279-311.
18. Smilde, A.K.; Wang, Y.; Kowalski, B.R. Theory of Medium-Rank Second-Order Calibration with Restricted-Tucker Models. *J. of Chemometrics* **1994**, *8,* 21-36.

Future Perspectives
Multiway Analysis in Medicinal Chemistry

8

8.1 Introduction

It is difficult, or maybe not even possible, to *a priori* predict which regression method is best suited for the analysis of a specific data set. A researcher simply has to rely on experience and proceed by means of a trial and error protocol. Multiway analysis methods have been used for quite some time in the field of analytical chemistry[1] and psychology. Not since 1988, when Cramer *et al.* introduced the CoMFA[2] method in 3D QSAR, no alternative for PLS has been reported. In this thesis, however, PLS[3,4] has successfully been replaced by multilinear PLS,[5] for the analysis of two 3D QSAR data sets[6,7] (see Chapters 5 and 6).

In the field of pharmacology and medicinal chemistry, data from *in vivo* (*e.g.*, microdialysis) and *in vitro* (*e.g.*, receptor binding) experiments are generated. The utilization of multivariate methods, *e.g.*, PLS and PCA for the analyses of the data has been very sparse[8] and are, as yet, not really accepted. A further extension of the statistical boundaries in medicinal chemistry, is the introduction of multiway analysis.[9,10] In the following, two real examples (Sections 8.2 and 8.4) taken from neuropharmacology and one hypothetical example (Section 8.3) from combinatorial chemistry will be presented. The results are presented as they originally were reported, together with suggestions of how two-way and/or multiway methods could be used as an alternative.

The objective of this chapter is to demonstrate the abundance of data, in medicinal chemistry, that can be arranged in multiway matrices. Consequently, no calculations are carried out here.

8.2 Example One: Neuropharmacology with Microdialysis

In the following example, two series of experiments were performed in rats,[11] with the objective to find out whether the citalopram (a selective serotonin reuptake inhibitor) induced increase in 5-HT levels, had an effect on the release of acetylcholine in the *ventral Hippocampus* area. A number of drugs administrated at different dosages were injected *sub cutaneously*, and for the duration of several hours, starting 60 minutes before the injection, samples were collected each 15 minutes, by means of microdialysis.[12] In the first and second series of experiments, the levels of serotonin (5-HT) and acetylcholine (Ach) were monitored, respectively. It was concluded from these experiments that no significant change in acetylcholine levels were observed, as the result of the increased serotonin levels in the *ventral Hippocampus* area.

Data of this type, usually are presented in two dimensional plots where the concentration levels of each drug are plotted as a function of time (Figure 8.1). For clarity, in Figure 8.1 the 15 acetylcholine samples are appended after the 15 serotonin samples, the drug concentrations (percentage of basal levels) are plotted as logarithmic values and the zero level corresponds to the basal concentration level.

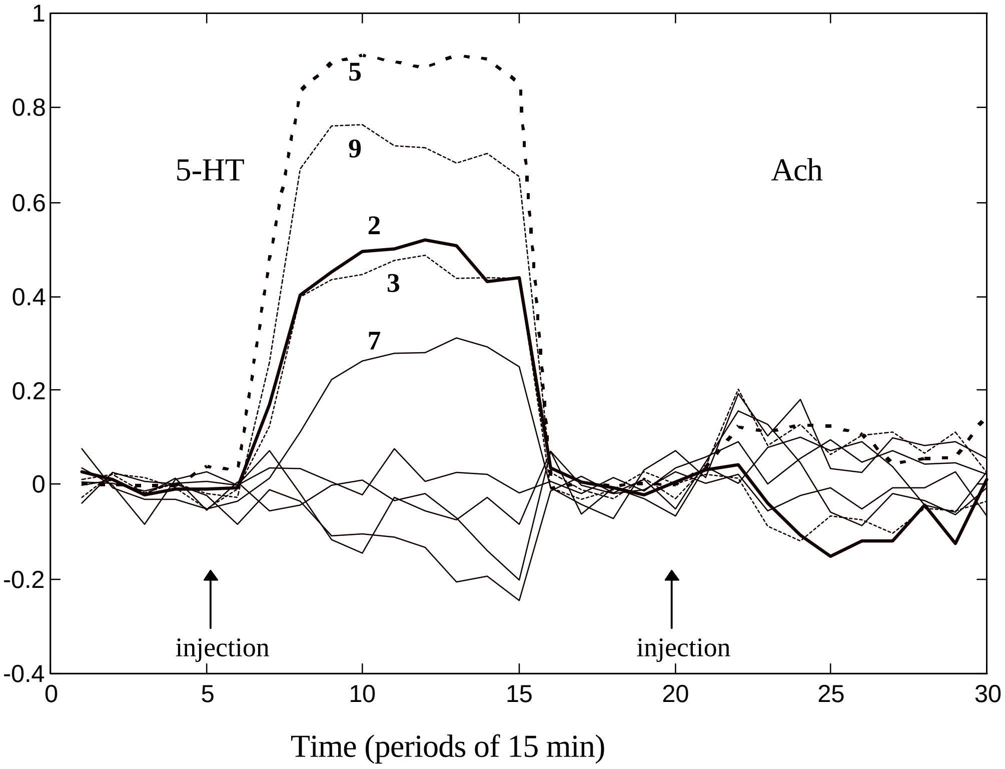


Figure 8.1 The levels of serotonin (5-HT; 1–15) and acetylcholine (Ach; 16–30) measured in the ventral Hippocampus area. The values were originally reported as percentage of basal levels, but here the logarithm of the same levels are presented. After centering of the data, the zero level corresponds to the basal level. The ID-numbers of the drugs with the most significant increase in 5-HT concentration are reported.

Figure 8.1, is a graphical representation of the two-way matrix in Figure 8.2(b), which may be decomposed by means of a Principal Component decomposition, as in Figure 8.3(b), into a score vector, $\mathbf{t}$ ($I \times 1$), and a loading vector, $\mathbf{p}$ ($JK \times 1$).

As an alternative, this data set could also be assembled in a three-way array, $\underline{\mathbf{X}}$ ($I \times J \times K$), like in Figure 8.2(a) and accordingly decomposed by means of a three-way PARAFAC decomposition (Figure 8.3(a)) into a score vector, $\mathbf{t}$ ($I \times 1$), and two loading vectors, $\mathbf{w}_J$ ($J \times 1$) and $\mathbf{w}_K$ ($K \times 1$), corresponding to the drug, the time and the response modes, respectively.

Independent of the decomposition method used, it is to be expected that most of the insignificant variation, especially in the acetylcholine mode, will be filtered off in the first few components and, consequently, not detrimentally affect the interpretation of the models. Furthermore, when the number of drugs is large, and with several responses, it is likely that PCA or PARAFAC models become more advantageous than traditional methods, *e.g.*, Figure 8.1, since, among other things, all experiments can be analyzed simultaneously.

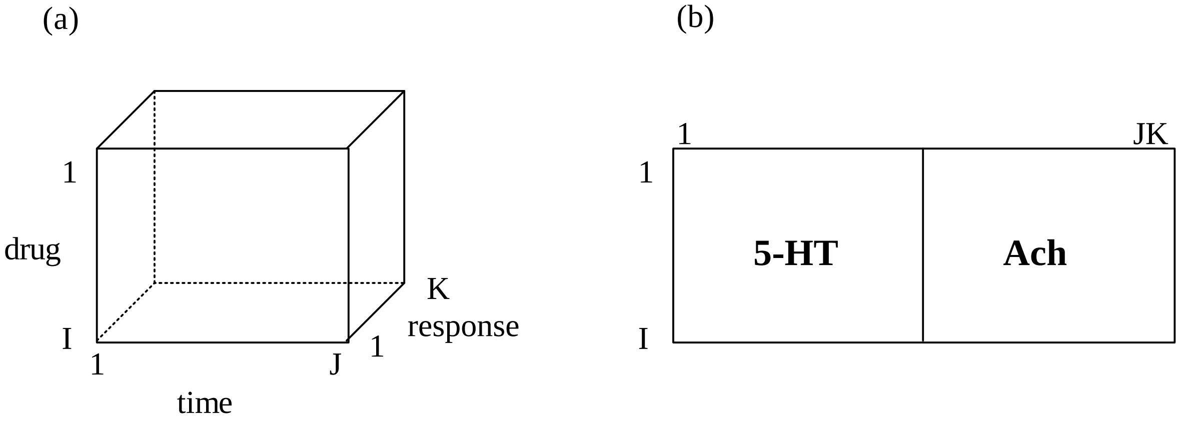


Figure 8.2 (a) Graphical representation of the three-way data set, $\underline{\mathbf{X}}$ ($I \times J \times K$), where nine different drugs or doses (I = 9) where injected and monitored for 5-HT and Ach, in the *ventral Hippocampus* area, for the duration of 210 minutes, with samples taken every 15 minutes (J = 15). (b) The unfolded three-way matrix, *i.e.*, the K slices ($I \times J$) of $\underline{\mathbf{X}}$ concatenated to form $\mathbf{X}$ ($I \times JK$).

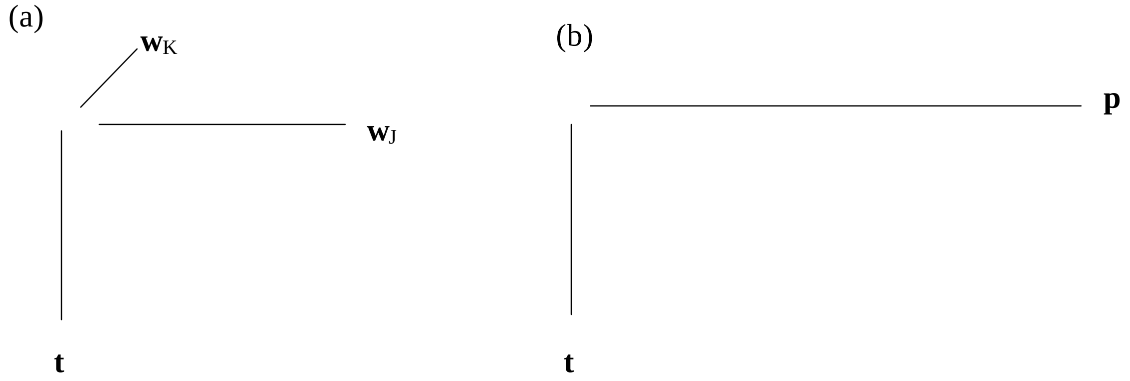


Figure 8.3 In (a), the one component PARAFAC decomposition of $\underline{\mathbf{X}}$ ($I \times J \times K$) is presented, where $\mathbf{w}_J$ ($J \times 1$) and $\mathbf{w}_K$ ($K \times 1$) are the loading vectors, corresponding to the time and the response modes in Figure 8.2(a), respectively. In (b), the one component Principal Component decomposition of $\mathbf{X}$ is presented, where $\mathbf{w}$ ($JK \times 1$) is the loading vector. The score vector $\mathbf{t}$ ($I \times 1$), corresponds in both (a) and (b) to the mode representing the drugs.

9.3 Example two: Combinatorial Chemistry

In the field of combinatorial chemistry the following situation may occur. Imagine a number of compounds are to be synthesized by permuting all the possible combinations of three different types of building blocks (A, B and C). Typically, A could be I different aromatic skeletons, B J different substituents on position R_1 and C K different substituents on position R_2. Since A, B and C consist of I, J and K different building blocks, respectively, IJK number of compounds must be synthesized. The complete design can be comprised in the form of a cube, as in Figure 8.4. The compounds are synthesized with the help from robots and, subsequently, subjected to screening in a number of different receptor binding assays,[13,14] *e.g.*, the dopamine D_2, D_3 and D_4 receptors. In order to evaluate the results, the receptor affinities from the synthesized compounds are collected in a two-way matrix with IJK rows and L columns (assuming L different receptors were considered).

Accordingly, the data may be analyzed by means of a PCA or with any other method able to handle two-way matrices. Alternatively, the cube structure of the designed compounds, in Figure 8.4, can be maintained. That is, for each receptor binding assay that the compounds are tested in, a new cube with binding results is obtained. In Figure 8.5, the data is collected in a four-way matrix $\underline{\mathbf{X}}$ ($I \times J \times K \times L$), which can be decomposed by means of a four-way PARAFAC[15,16] or Tucker4[10,17,18] model. It is likely that compounds from the same region in the cube, are structurally associated, having affinity for about the same receptors. If the number of screened receptors is large, multivariate analysis methods (*e.g.*, PCA) is recommended, and whether multiway methods, *e.g.*, PARAFAC or Tucker, will improve the interpretation of the data is still to be found out.

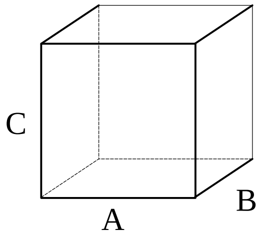


Figure 8.4 A possible combinatorial design $\underline{\mathbf{X}}$ ($I \times J \times K$) in three different modes, *i.e.*, A, B and C.

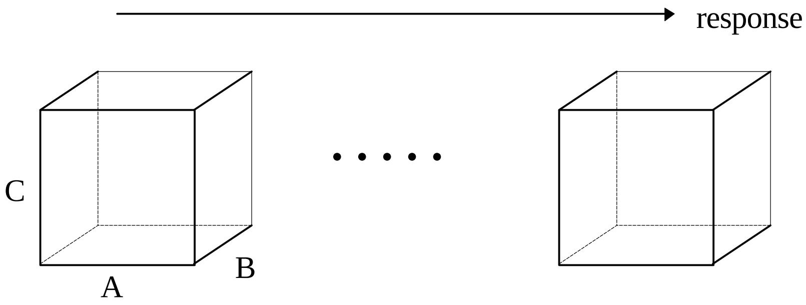
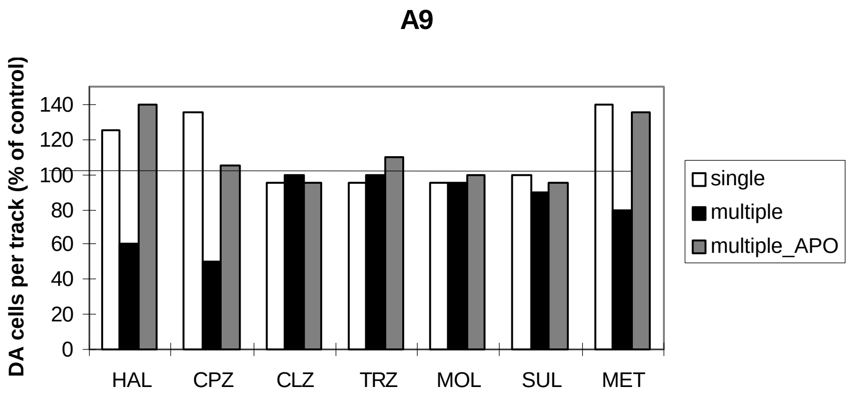


Figure 8.5 The matrix $\underline{\mathbf{X}}$ ($I \times J \times K \times L$), where the modes A, B, and C consist of I, J and K different building blocks and the response mode represents the L different receptor types.

8.4 Example Three: Neuropharmacology

In 1983, White and Wang[19] reported on the effect of prolonged treatment with classical and atypical antipsychotic drugs on the number of spontaneously active dopamine neurons, in both the *substantia nigra* (A9) and the *ventral tegmental* area (A10). It was found that atypical antipsychotic drugs selectively decreased the number of A10 cells, while drugs with typical antipsychotic efficacy, failed to decrease the dopaminergic activity. The results from the investigation were comprised in two figures (Figures 8.6(a) and 8.6(b)). For each drug, the effect of one single injection (white bars) was compared with the effect from repeated treatments (black bars). After each experiment with repeated treatments, the effect of a single injection of apomorphine (0.063 mg/kg) was investigated (gray bars).

(a)



(b)

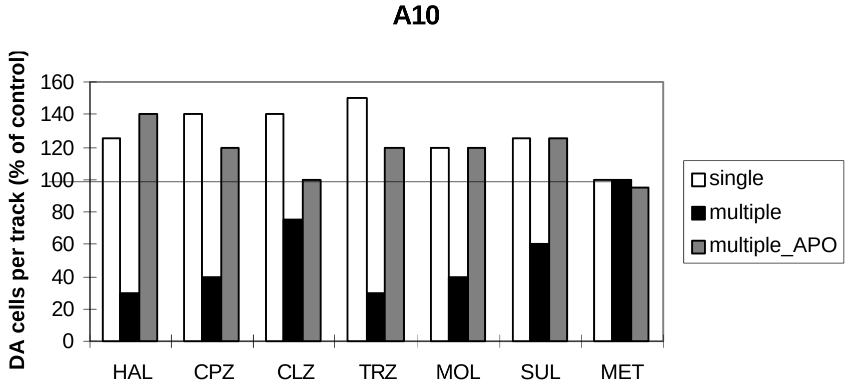


Figure 8.6 The effects of short and long-term treatment with various typical and atypical antipsychotics, on the number of spontaneously active A9 and A10 DA cells. Abbrevations: HAL (haloperidol), CPZ (chlorpromazine), CLZ (clozapine), TRZ (thioridazine), MOL (molindone), SUL (sulpiride) and MET (metoclapramide).

Also in this example a multiway approach can be employed by assembling the data in Figure 8.6, in a three-way array, like in Figure 8.7(a). The decomposition of $\underline{\mathbf{X}}$ is performed as in Figure 8.3(a) but here the score vector (**t**) corresponds to the drugs and the weight vectors $\mathbf{w}_J$ and $\mathbf{w}_K$ correspond to the type of injection and brain area, respectively. Since two modes in $\underline{\mathbf{X}}$ are very small (J = 3 and K = 2), a PCA of the unfolded three-way array, in Figure 8.7(b), is also likely to be successful.

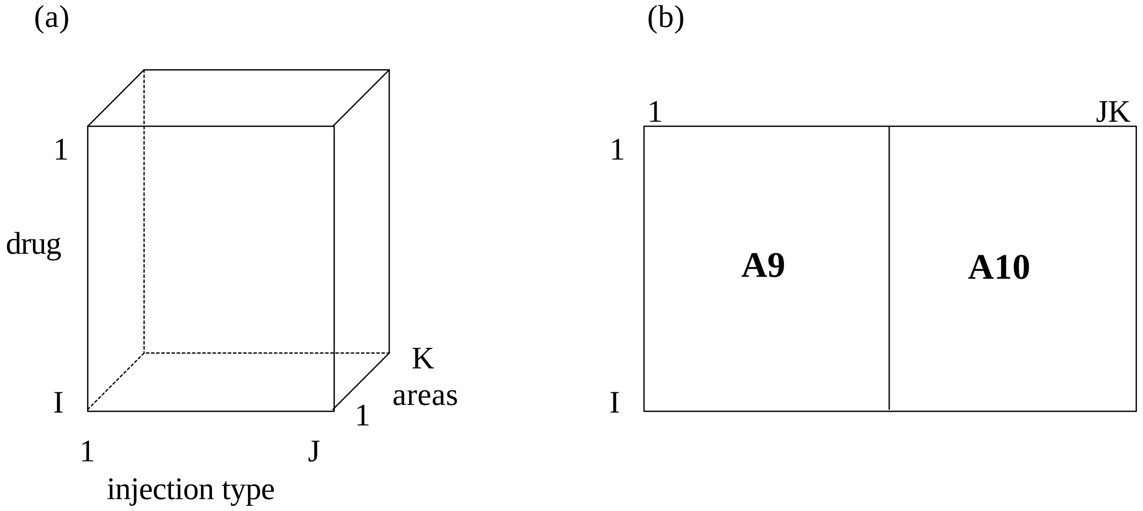


Figure 8.7 (a) The data in Figure 8.6 assembled in a cube $\underline{\mathbf{X}}$ ($I \times J \times K$), *i.e.*, J different types of treatments were performed with I different drugs and DA neuron firing were measured in K different brain areas. (b) The cube in (a), $\underline{\mathbf{X}}$, is unfolded to form a two-way matrix, $\mathbf{X}$ ($I \times JK$).

8.5 Conclusions

In this chapter, three different examples where selected on rather arbitrary grounds, in order to demonstrate the abundance of data sets that can be arranged in multiway arrays. The objective was not to favor multiway methods before any of the present analysis methods, but to introduce the multiway methods as potential and effective analysis methods. Furthermore, there exist no clear solutions or suggestions of how to handle, for example, multiway data obtained from microdialysis where each experiment has been repeated several times. What kind of data scaling method is most effective? Is block scaling, as was used in Chapter 4 for 3D QSAR data, a suitable option for the example in Section 8.2? Clearly, before multivariate and multiway analysis methods, *e.g.*, PCA, PLS, PARAFAC and Tucker, can be introduced and fully accepted in pharmacology several important questions need to be answered.

8.6 References

1. Smilde, A.K. PhD Thesis. Multivariate Calibration of Reversed-Phase Chromatography Systems Research Group of Chemometrics, University of Groningen, The Netherlands. **1990**
2. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analyses (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959-5967.
3. SYBYL- Molecular Modeling Software, *6.3*, Tripos Incorporated, 1699 S. Hanley Rd, St. Louis, Missouri 63144-2913, USA,
4. Geladi, P. and Kowalski, B.R. Partial Least Squares: A Tutorial. *Anal. Chem. Acta* **1986**, *185,* 1-17.
5. Bro, R. Multiway Calibration. Multilinear PLS. *J. of Chemometrics* **1996**, *10*, 47-61.
6. Nilsson, J.; De Jong, S.; Smilde, A.K. Multiway Calibration in 3D QSAR. *J. of Chemometrics* **1997**, *11,* 511-524.
7. Nilsson, J.; Homan, E.J.; Smilde, A.K.; Grol, C.J.; Wikström, H. A Multiway 3D QSAR Analysis of a Series of (*S*)-*N*-[(1-Ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides. *J. Comput. -Aided Mol. Design* **1997** in press
8. Hansson, L.; Waters, N.; Winblad, B.; Gottfries, C.-G.; Carlsson, A. Evidence for Biochemical Heterogenity in Schizophrenia: A Multivariate Study of Monoaminergic Indices in Human Post-Mortal Brain Tissue. *J. Neural Transm.* **1994**, *98,* 217-235.
9. Geladi, P. Analyses of Multi-way (Multi-Mode) Data. *Chemom. and Intell. Lab. Syst.* **1989**, *7,* 11-30.

10. Smilde, A.K. Three-way Analyses. Problems and Prospects. *Chemom. and Intell. Lab. Syst.* **1992**, *15,* 143-157.
11. Cremers T.[Unpublished data]; Department of Medicinal Chemistry; University of Groningen, The Netherlands **1997**
12. Westerink, B.H.C. Brain Microdialysis and its Application for the Study of Animal Behaviour. *Behav. Brain Res.* **1995**, *70,* 103-124.
13. Broach, J.R. and Thorner, J. High-Throughput Screening for Drug Discovery. *Nature.* **1996**, *384*, 14-16.
14. Chapman, D. The Measurement of Molecular Diversity: a Three-Dimensional Approach. *J. Comput. -Aided Mol. Design* **1996**, *10*, 501-512.
15. Carroll, J. and Chang, J.J. Analysis of Individual Differences in Multidimensional Scaling with an N-way Generalisation of the Eckart-Young Decomposition. *Psycometrika* **1970**, *35*, 283-319.
16. Harshman, R.A. Foundations of the PARAFAC Procedure: Models and Conditions for an Exploratory Multimodal Factor Analysis. *UCLA Working Papers in Phonetics* **1970**, *16*, 1-84.
17. Tucker, L. Problems of Measuring Change. Implications of Factor Analysis of Three-Way Matrices for Measurement of Change. Harris, C. Ed. University of Wisconsin Press.: Madison, **1963**; pp. 122-137.
18. Tucker, L. Some Mathematical Notes on Three-Mode Factor Analysis. *Psycometrika* **1966**, *31*, 279-311.
19. White, F.J. and Wang, R.Y. Differential Effects of Classical and Atypical Antipsychotic Drugs on A9 and A10 Dopamine Neurons. *Science* **1983**, *221,* 1054-1057.

Summary

There are many opinions about how medicinal chemistry should be practiced. The procedure described in this thesis for the design of new drugs comprises several steps including the selection of a lead compound, experimental design, syntheses of new compounds, pharmacological *in vitro* testing, molecular modeling and multivariate statistical evaluation. Each step has achieved more or less attention but the main emphasis has been put on the multivariate statistical evaluation.

In the pursuit of potent and selective dopamine D_3 antagonists the *trans*-*N*-(n-propyl)-7-[[(trifluoromethyl)sulfonyl]oxy]-OHB[*f*]Q (**1** in Table 3.3) was initially considered as a lead compound for this investigation. The compound displayed presynaptic DA receptor antagonistic properties in rats (see Table 1.4), although the 7-hydroxy analogue was a potent agonist. The racemic **1** showed a 10-fold selectivity for the DA D_3 over the D_2 receptor. The influence of the 7-triflate group on these effects was of special interest.

Several OHB[*f*]Qs were designed and described with theoretical physicochemical descriptors using compound **1** as the template. Only a fraction of the most different compounds was selected by means of an experimental design in the descriptor space (see **Chapter 3**). Accordingly, the 16 compounds were synthesized and tested for *in vitro* affinities for the DA D_2, D_3 and D_4 receptor subtypes. None of the compounds were real selective for any of the receptors and, in general, compounds with a hydroxy group at the seven position displayed significant high affinities for all three dopamine receptors, while the compounds with a sulfon ester group were less potent. In addition, the sulfon ester group suppressed the affinity for the DA D_4 receptor. The nitrogen substituent may be as large as a phenylethyl group without detrimentally affecting the affinity for the DA receptors. Finally, a compound with a 7-OH group and an *N*-propargyl group lacks affinity for the DA D_4 receptor. The somewhat rigid *N*-propargyl group and the low pK_a value (6.1) may be contributing factors to the low D_4 affinity.

At this point, a 3D QSAR model may provide information about how to proceed further. Which compound should be synthesized next? However, due to lack of time further investigation of this data set was not pursued. Instead, a data set was retrieved from Dr. Shelly Glase at Parke-Davis in the USA with which the multivariate statistical analyses were investigated. This theoretical part of the thesis is focused mainly on the optimization of multivariate and multiway regression analysis methods in 3D QSAR.

In **Chapter 4**, conformational analyses and alignments of mutual and potential interaction points with the DA D_3 receptor of Dr. Glase's flexible compounds were carried out. The absolute configuration of the compound (**1** in Table 4.1) used as the template to fit the remaining 29 compounds on, was determined with help from X-ray crystallographic structures. It is important to stress that the emphasis and aim of a 3D QSAR study is to measure the differences (*e.g.* steric and electrostatic fields) between the compounds under investigation and, consequently, the aligned compounds (Figure 5.1) do not necessarily fit into the active site of the receptor. This is in particular the case when the 3D structure of the target receptor is not known, which is the case for all dopamine receptors.

The Grid program was used to generate molecular fields from ten different probe atoms for all the aligned compounds. Accordingly, the three probes that generated the most different molecular fields were selected by means of a Principal Component Analysis (see Figure 4.2). The OH2, C3 and CA+2 probes were selected and reflect the hydrogen bonding, steric and electrostatic interactions, respectively, between the target receptor protein and the ligands. Each molecular field, in form of a 3D grid (see Figure 1.9), is unfolded to form a row vector with as many elements as there are grid points in the grid. The complete data set, **X** in Figure 4.4, comprises 30 molecules described by 25110 molecular descriptors and one dependent variable, **y**, *i.e.* the affinity for the DA D_3 receptor subtype. The very essence of 3D QSAR modeling is to establish a regression model between **X** and **y**.

In Chapter 4, the GOLPE program was used for the variable selection, data pretreatment, and subsequently the PLS regression analysis. The necessity of eliminating grid points with more or less insignificant variation, *e.g.* grid points at large distances from the ligands placed in the periphery of the grid, was investigated. It turned out that the crossvalidated Q^2 increased from 0.45 to 0.65 when the number of variables was reduced from 19180 (after pretreatment) to 784. The variable reduction was carried out in two steps: first by means of D-optimal preselection in the loading space from an initial PLS model followed by a fractional factorial design selection procedure. The GOLPE analysis is based on the two-way PLS method and, as a consequence, the descriptor grids must be unfolded into a two-way matrix prior to the analysis. Actually, the raw data set looks like in Figure 5.2 where five different directions or modes are defined; one mode represents the 30 molecules, the x, y, and z modes correspond to the axes of the three dimensional grids and the fifth mode represents the three probes. For the analysis of a five-way data set the Multilinear PLS method (N-PLS) intuitively is a better choice than the two-way PLS method, since the unfolding procedure is unnecessary. Instead, the data set is directly decomposed into a score vector and four loading vectors (see Figure 5.3) corresponding to the five modes as defined above.

In **Chapter 5**, it was shown that the N-PLS models are easier to interpret due to the loading vectors obtained in each mode, and that they possess higher predictability than the PLS models. However, the fit was worse with N-PLS, as compared with PLS, possible due to the smaller number of parameters that needed to be estimated. PLS probably overfits and therefore N-PLS reflects better the relationship between $\underline{\mathbf{X}}$ and **y**. At this point it was important to verify the excellent performance of the N-PLS method by analyzing yet another data set. In **Chapter 6**, a very well known data set, consisting of 58 benzamides (raclopride analogues) with affinity for the DA D_2 receptor subtype, was re-analyzed using GRID descriptors and N-PLS as the regression method. The result was convincing and the final model had a predicted Q^2 of 0.62. It was concluded that the N-PLS method certainly is an alternative to PLS for the analysis of 3D QSAR data sets.

Both PLS and N-PLS are so component wise regression methods since each component is calculated from the residuals of **X**, after removing the variation accounted for by the previous component. In **Chapter 7**, two methods that calculate all components simultaneously in an alternating least squares algorithm have been evaluated. The two methods, PCovR/PARAFAC and PCovR/Tucker, are combinations of the Principal Covariate Regression (PCovR) method with PARAFAC and Tucker decompositions, respectively. These algorithms balance between reconstructing $\underline{\mathbf{X}}$ and explaining $\mathbf{y}$ by adjusting the α value (see Equation 7.4). Therefore, in order to find the models with optimal predictability the α value and the number of components were optimized simultaneously. When the sum of squares of $\underline{\mathbf{X}}$ and $\mathbf{y}$ were normalized, the most predictive models had α values close to unity. Interestingly, the size of α that produced the most predictive models could be tuned with the ratio of the sum of squares of $\underline{\mathbf{X}}$ and $\mathbf{y}$ (s_r). The best PCovR/Tucker5 model had a predictive $Q^2 = 0.31$ and was found when $\alpha = 0.4$, $s_r = 100$ using (6,7,3,4,2) number of components in the five modes. This model was considered stable, reproducible and it had a high predictability. The corresponding best PCovR/PARAFAC model was unstable, possibly since the data set has a Tucker structure. It was shown that the convergence of this latter algorithm strongly depended on the starting parameters used and frequently converged into local minima. This observation reduced significantly the reliability of the method, at least for the analysis of the present 3D QSAR data set. The question is how reliable is the N-PLS method since it too, is a PARAFAC derived method? It is clear, however, that the predictability of the N-PLS and the PCovR/Tucker5 models are comparable but N-PLS models are easier to interpret and provides simpler solutions.

Finally, two of the three multiway regression methods, evaluated in this thesis, definitely are alternatives to consider along with the traditional PLS method when 3D QSAR data sets need to be analyzed in the future.

In **Chapter 8**, three additional fields of research, where possible multiway data are generated, are proposed. Suggestions for the analysis of these types of data and a future prospective for multiway methods in medicinal chemistry are given.

Samenvatting

Er bestaan verschillende meningen over de manier waarop farmacochemie in de praktijk uitgevoerd zou moeten worden. De in dit proefschrift beschreven procedure voor het ontwerpen van nieuwe geneesmiddelen bestaat uit verschillende fases, t.w. de selectie van een *template* verbinding, de experimentele proefopzet, de synthese van nieuwe verbindingen, de *in vitro* farmacologische evaluatie, het molecular modelen en de multivariate statistische evaluatie. Op elk van deze fases is reeds in meerdere of mindere mate de aandacht gevestigd geweest, maar in dit proefschrift wordt de meeste nadruk gelegd op de multivariate statistische evaluatie.

In de zoektocht naar potente en selectieve D_3 antagonisten werd in eerste instantie de *trans*-*N*-(n-Propyl)-7-[[trifluoromethyl)sulfonyl]oxy]-OHB[*f*]Q (zie **1** in Tabel 3.3) verondersteld een *template* verbinding te zijn voor dit onderzoek. De verbinding vertoonde presynaptisch DA receptor antagonistische eigenschappen in ratten (zie Tabel 1.4), ondanks het feit dat de 7-hydroxy-analoog een potente agonist is. Het racemaat **1** vertoonde een tienvoudige selectiviteit voor de DA D_3 receptor boven de D_2 receptor. De invloed van de 7-triflaatgroep op deze effecten was van speciaal belang.

Verschillende OHB[*f*]Qs werden ontworpen en beschreven met theoretische fysisch-chemische variabelen, terwijl verbinding **1** werd gebruikt als *template*. Slechts een fractie van de meest verschillende verbindingen werd geselecteerd door middel van een experimentele proefopzet in the *descriptor space* (zie **Hoofdstuk 3**). Derhalve werden de 16 verbindingen gesynthetiseerd en getest op *in vitro* affiniteit voor de DA D_2, D_3 en D_4 receptor subtypes. Geen van de verbindingen was echt selectief voor één van deze receptoren; in het algemeen vertoonden verbindingen met een hydroxy-groep op de 7-positie een significante hoge affiniteit voor alle drie dopamine receptoren, terwijl de verbindingen met de sulfon ester groep minder potent waren. Bovendien onderdrukte de sulfon ester groep de affiniteit voor de DA D_4 receptor. De stikstof substituent mag zo groot zijn als een fenylethyl groep zonder dat de affiniteit voor de DA receptors nadelig wordt beïnvloed. Uiteindelijk ontbreekt bij een verbinding met een 7-OH groep en een *N*-propargyl groep de affiniteit voor de DA D_4 receptor. De enigszins starre *N*-propargyl groep en de lage pK_a-waarde (6.1) zouden bijdragende factoren kunnen zijn voor de lage D_4 affiniteit.

Op dit punt zou een 3D QSAR model informatie kunnen verschaffen over de verder te volgen procedure; welke volgende verbinding dient te worden gesynthetiseerd? Echter, door gebrek aan tijd werd verder onderzoek naar deze data set niet voortgezet. In plaats daarvan werd een data set ter beschikking gesteld door Dr. Shelly Glase (Parke-Davis, VS) waarmee de multivariate statistische analyses werden onderzocht. Dit theoretische deel van het proefschrift is hoofdzakelijk gericht op de optimalisering van multivariate en multiweg regressie-analyse methoden in 3D QSAR.

Hoofdstuk 4 beschrijft hoe de conformatie-analyses en de *alignments* van wederzijdse en mogelijke interactie-punten met de DA D_3 receptor van de flexibele verbindingen van Dr. Glase werden uitgevoerd. De absolute configuratie van de verbinding (**1** in Tabel 4.1), die werd gebruikt als *template* voor de resterende 29 verbindingen, werd bepaald met behulp van röntgen kristallografische structuren. Het is van belang te benadrukken dat het doel van een 3D QSAR studie is om de verschillen (b.v. sterische en electrostatische velden) te meten tussen de verbindingen die worden onderzocht. Hierdoor hoeven de uitgelijnde verbindingen (Figuur 5.1) niet noodzakelijkerwijs te passen in de *active site* van de receptor. Dit is in het bijzonder het geval wanneer de 3D structuur van de beoogde receptor onbekend is, hetgeen het geval is voor alle dopamine receptoren.

Het programma Grid werd gebruikt om moleculaire velden van tien verschillende *probe* atomen te genereren voor alle uitgelijnde verbindingen. Vervolgens werden de drie *probes*, die de meest verschillende moleculaire velden genereerden, geselecteerd met behulp van een Principal Component Analyse (zie Figuur 4.2). De OH2, C3 en CA+2 probes werden geselecteerd en reflecteren achtereenvolgens de hydrogene binding en de sterische en electrostatische interacties tussen de liganden en het bedoelde receptoreiwit. Elk moleculair veld, in de vorm van een 3D rooster (zie Figuur 1.9) wordt uitgevouwen om een rij-vector te vormen met evenveel elementen als het aantal roosterpunten in het rooster. De complete data set, **X** in Figuur 4.4, bestaat uit 30 moleculen beschreven door 25110 moleculaire variabelen en één afhankelijke variabele **y**, te weten de affiniteit voor het DA D_3 receptorsubtype. Het grote belang van 3D QSAR modeling is om een regressiemodel vast te stellen tussen **X** en **y**.

In Hoofdstuk 4 werd het programma GOLPE gebruikt voor de variabele selectie, voor de voorbehandeling van de data én voor de daarop volgende PLS regressie analyse. De noodzaak van het elimineren van roosterpunten met min of meer onbeduidende variatie, bijvoorbeeld roosterpunten geplaatst op grote afstand van de liganden in de periferie van het rooster, werd onderzocht. Gebleken is dat de gekruis-validerede Q^2 toenam van 0.45 naar 0.65 als het aantal variabelen werd beperkt van 19180 (na voorbehandeling) tot 784. De variabele reductie werd in twee stappen uitgevoerd: eerst door middel van een D-optimale voorselectie in de ladingen ruimte van een initieel PLS model, gevolgd door een selectieprocedure voor *Fractional Factorial Design*. De GOLPE analyse is gebaseerd op de tweeweg PLS methode, en als gevolg daarvan moeten de descriptor roosters, voorafgaand aan de analyse, worden uitgevouwen in een tweeweg matrix.

Feitelijk lijkt de ruwe data set op Figuur 5.2, waar vijf verschillende richtingen of modes zijn gedefinieerd; een mode vertegenwoordigt de 30 moleculen, the x, y en z modes komen overeen met de assen van de driedimensionale roosters en de vijfde mode stelt de 3 probes voor. Voor de analyse van een vijf-weg data set is de multi-lineaire PLS methode (N-PLS) intuïtief gezien een betere keuze dan de twee-weg PLS-methode, aangezien de uitvouwingsprocedure niet nodig is. In plaats daarvan wordt de data set direct ontleed in een score-vector en vier lading-vectoren (zie Figuur 5.3), die overeenkomen met de hiervoor beschreven 5 modes.

In **Hoofdstuk 5** werd aangetoond dat de N-PLS modellen gemakkelijker te interpreteren zijn dankzij de lading-vectoren die in elke mode verkregen zijn, en dat ze een hogere voorspelbaarheidswaarde bezitten dan de PLS modellen. Echter, de fit met de N-PLS was, vergeleken met de PLS, slechter, hetgeen waarschijnlijk te wijten is aan een kleiner aantal parameters dat geschat moest worden. De PLS modellen waren waarschijnlijk *overfitted* en zou de betere voorspelbaarheid van de N-PLS modellen kunnen verklaren, die werd geschat met een externe test set van 21 verbindingen. Op dit punt was het van belang het excellente optreden van de N-PLS methode te verifiëren door nog een andere data set te analyseren. In **Hoofdstuk 6** werd een zeer bekende data set, bestaande uit 58 benzamides (raclopride analoga) met affiniteit voor het DA D_2 receptorsubtype, opnieuw geanalyseerd door gebruik te maken van GRID descriptors, met N-PLS als de regressie-methode. Het resultaat was overtuigend en het uiteindelijke model had een voorspelde Q^2 van 0.62. De conclusie is dat de N-PLS methode wel degelijk een alternatief is voor de PLS voor de analyse van 3D QSAR data sets.

Zowel PLS en N-PLS zijn componentgewijze regressie methodes omdat elke component wordt berekend vanuit het residu van **X**, nadat de voor de vorige component verantwoordelijke variatie is verwijderd. In **Hoofdstuk 7** zijn twee methoden die alle componenten gelijktijdig berekenen in een alternerende kleinste kwadraten algorithmen geëvalueerd. De twee mehoden, PCovR/PARAFAC en PcovR/Tucker zijn combinaties van de Principal Covariate Regression (PcovR) methode met achtereenvolgens PARAFAC en Tucker decomposities. Deze algorithmen balanceren tussen het reconstrueren van **X** en het verklaren van **y** door de α-waarde toe te voegen (zie vergelijking 7.4). Om de modellen te vinden met een optimale voorspelbaarheidswaarde werden de α-waarde en het aantal componenten gelijktijdig geoptimaliseerd. Als de som van de kwadraten van **X** en **y** werden genormaliseerd, hadden de best voorspellende modellen α-waarden dichtbij één. Interessant was dat de grootte van de α-waarde, die de best voorspellende modellen produceerde, kon worden afgestemd met de verhouding van de som van de kwadraten van **X** en **y** (s_r). Het beste PcovR/Tucker5 model had een $Q^2 = 0.31$ en werd gevonden bij een $\alpha = 0.4$, $s_r = 100$, gebruik makend van een aantal (6,7,3,4,2) componenten in de 5 modes. Dit model was stabiel, reproduceerbaar en beschikken over een hoge mate van voorspellende vermogen. Het overeenkomstige beste PcovR/PARAFAC model was instabiel, waarschijnlijk omdat de data set een Tucker-structuur heeft. Aangetoond was dat de convergentie van dit laatste algorithme erg afhankelijk was van de gebruikte startparameters en dat ze vaak convergeren in locale minima. Deze waarneming verminderde de betrouwbaarheid van de methode, in ieder geval met betrekking tot de analyse van de huidige 3D QSAR data set. De vraag is hoe betrouwbaar de N-PLS methode is aangezien het ook een van PARAFAC afgeleide methode is. Het is echter duidelijk dat de voorspelbaarheid van de N-PLS en de PcovR/Tucker5 modellen vergelijkbaar zijn maar dat N-PLS modellen gemakkelijker te interpreteren zijn.

Uiteindelijk zijn twee van de drie in dit proefschrift geëvalueerde multiweg regressie methoden beslist alternatieven die het overwegen waard zijn tezamen met de traditionele PLS methode als 3D QSAR data sets geanalyseerd zouden moeten worden in de toekomst.

In **Hoofdstuk 8** zijn drie additionele onderzoeksgebieden voorgesteld, waar mogelijk multiweg data worden gegenereerd. Verder worden suggesties voor de analyse van dit soort data en de toekomstperspectieven voor multiweg methoden in de farmacochemie gegeven.

Curriculum Vitae

Personal Data

Name : Jonas Nilsson
Nationality : Swedish
Sex : Male
Date of Birth : 1967-06-09
Place of Birth : Kalmar, Sweden

List of Publications

[1] Stjärnlöv, P.; Elebring, T.; Nilsson, J.; Andersson, B.; Lagerquist, S.; Svensson, K.; Ekman, A.; Carlsson, A.; Wikström, H. 6,7,8,9-Tetrahydro-*N*,*N*-di-n-propyl-3*H*-benzindol-8-amines. Derivatives as Potent and Orally Active Serotonin 5-HT_{1A} Receptor Agonists. *J. Med.Chem.* **1994**,37,3263-3273

[2] Sonesson, C.; Barf, T.; Nilsson, J.; Dijkstra, D.; Carlsson, A.; Svensson, K.; W. Smith, M.; J.Martin, I.; Neil Duncan, J; J. King, L.; Wikström, H. Synthesis and Evaluation of Pharmacokinetic Properties of Monopropyl Analogs of 5-, 7-, and 8-[[(Triflouromethyl)sulfonyl]oxy]-2-aminotetralins: Central Dopamine and Serotonine Receptor Activity. *J. Med. Chem*., **1995**, 38, 1319-1329

[3] Nilsson, J.; Glase, S; Pugsley, T; Pastor, M.; Cruciani, G.; Clementi, S.; Smilde, A.; Wikström, H., A GRID/GOLPE 3D QSAR study on a Set of Benz- and Naphthamides with Affinity for the Dopamine D_3 Receptor Subtype., *J. Med. Chem.*, **1997**, 40, 833-840.

[4] Nilsson, J.; De Jong, S.; Smilde, A. Multiway Calibration in 3D QSAR, *J. of Chemometrics,* **1997**, 11,511-524

[5] Nilsson, J.; Homan, E. J.; Smilde, A.K.; Grol, C.J.; Wikström, H., A Multiway 3D QSAR Analysis of a Series of (*S*)-*N*-[(Ethyl-2-pyrrolidinyl)methyl]-6-methoxybenzamides, *Accepted in J. of. Comput. Aided Mol. Design*

[6] Nilsson, J.; Selditz, U.; Sundell, S.; Lundmark, M.; Pugsley, T.; Smilde, A.K.;Wikström, H., Design, Syntheses and QSAR of a Series *trans*-1,2,3,4,4a,5,6,10b-Octahydrobenzo[*f*]quinolines with Dopaminergic Affinity, *submitted*

[7] Van der Graaf, P. H.; Nilsson, J.; Danhof, M; P. Ijzerman, A. P. A New Approach to Quantitative Structure-Pharmacokinetic Relationships (QSPKR) Analysis: Application to Adenosine A_1 Receptor Agonists, *in preparation*

[8] Nilsson, J; Kiers, H.A.L.; Smilde, A. K., Multiway Simultaneous Two-Block Analysis with Applications to 3D QSAR. *In preparation*

Posters and Presentations

[9] A Tentative 5-HT_{1A} Receptor Model Based on Geometrical Fits and Electrostatic Potential Calculations. *XIIIth International Symposium on Medicinal Chemistry*, Paris, **1994**.

[10] Work-shop: Multi-way PLS en GOLPE in QSAR, *KNCV Workshop Multivariate Data Analyse*, Zeist, The Netherlands, **1995**

[11] Comfa: A Statistical Evaluation. *SON Discussiegroup Farmacochemie*, Luntheren, The Netherlands, **1995**

[12] A Novel Approach for Optimal Selection of Regions Correlated with Receptor Binding in a CoMFA grid. *10th Camerino-Noordwijkerhout Symposium*, Camerino, Italy, **1995**

[13] A Comparison between Multi-way PLS and GOLPE utilised as Variable Selection Tools, Applied on GRID-parameters from a Set of Compounds with Affinity for the Dopamine D_3 Receptor Subtype, *11thEuropean Symposium on Quantitative Structure-Activity Relationships: Computer-Assisted Lead Finding and Optimization,* Lausanne, Switzerland, **1996**

[14] Lecture, Partial Least Squares and CoMFA, *Course in Drug Design and Computational Chemistry (GUIDE-Groningen Utrecht Institute for Drug Exploration*), Utrecht, **1996**

[15] Oral Presentation, Multiway Calibration in 3D QSAR, *Werkgroep Chemometrie van de Sectie Analytische Chemie (KNCV),* Nijmegen, The Netherlands, **1996**

[16] Oral presentation, Multiway Calibration in 3D QSAR, *Department of Medicinal Chemistry, Astra Arcus,* Södertalje, Sweden, **1996**

[17] Oral presentation, Multiway Calibration in 3D QSAR, *TRICAP '97 (Three Way Methods in Chemistry and Psychology)*, Lake Chelan, Seattle, Washington, USA, **1997**

All these persons, colleagues, friends, family and events…

Håkan and Age, I want to start with thanking you for giving me the opportunity to become an AiO in Groningen, five years ago. You have introduced me to the medicinal chemistry and multiway data analysis and allowed me to work independently. I have appreciated the excellent supervision that you have provided and I hope that we will keep in contact also in the future. Age, I especially want to thank you for 'keeping' me as your AiO even after that you traded Groningen for Amsterdam.

Thanks Janita and Janneke for helping me out whenever I got lost in the 'RUG bureaucracy', in the jungle of Dutch impossible expressions and other administrative difficulties. Janita, I am very grateful for your help with the translation of the Summary of this thesis. Thanks, I owe you one!

I want to thank the following persons who have contributed to my research: Sijmen de Jong, Henk Kiers and Rasmus Bro for providing Matlab codes for several of the algorithms used in this thesis; Steven van Helden at Organon in Oss, for helping me with the generation of molecular descriptors; Shelly Glase at Parke Davis in USA, for providing me with the ligands that comprise the data set introduced in Chapter 4; Tom Pugsley also at Parke Davis, for the *in vitro* receptor binding screening of my compounds in Chapter 3; Wia Timmerman and Thomas Cremers for providing two of the data sets discussed in Chapter 8; Jim Bristol and Larry Wise at Parke Davis and GUIDE for financial support covering the printing and reproduction costs of this thesis.

I want to express my appreciation for the members in my 'leescommissie' Jos ten Berge, Sergio Clementi and Tommy Liljefors who struggled with my manuscript, pointed out errors and made suggestions that improved the quality of the manuscript. Thank you very much.

Sergio, Gabry, Manolo and the rest of the 'Perugian' group, I want to thank you all for making my stay in Perugia such a pleasant experience. I really appreciated the excellent supervision, the good advice and the friendly working atmosphere, that you provided.

Tack går till alla mina goda vänner i Lund som fått stå ut med mina mellanlandningar på vägen hem i samband med jul och sommar semestrar. Då framför allt Johann och Christel, Putte och Annette, Johan och Amanda vars gästfrihet jag profiterat på oftast. Tack allesammans, nu är det er tur! Johann och Christel, jag kommer alltid att sakna soffan! Trots att jag har haft en otrolig tid här i Holland har det varit tillfällen då jag önskat att jag var på hemma plan istället. Jag tänker då framför allt på alla N+U sammankomster, midsommar-aftnar, Lucia, Valborg, kräftskivor och alla andra traditioner som bara, på riktigt, kan firas med lands-män/kvinnor hemma i Sverige. För att kompensera för allt jag missat hoppas jag få se så många av er som möjligt i Groningen den 27 mars. Johann och Göran, ni visste inte vad ni gav er in på när ni lovade ställa upp som mina paranimfen, men det kommer att bli kul!

Ola och Maria, ni vet inte hur tacksam för att ni ville hjälpa mig att organisera flytten hem till Sverige. Det hade blivit litet tjurigt ensam.

The light at the end of most weeks has been “De Toeter” and I want to thank all ‘the regulars’, Cor, Evert, Pieter, Jack, Gert, Kees, Marguerite and others, with whom I have spend a countless number of ‘gezellige’ Friday nights.

Pieter, I already miss the long hours that we spend maintaining our endorphin addiction at ‘het ACLO sportcentrum’. An as dedicated training partner and good friend, all in one, is hard to find. I know for sure that we will keep in contact and I hope to see you and Petra in Uppsala, sometime soon.

Tjeerd and Sander, as the first of my friends in Groningen you helped me feel at home already from the start and I hope to see you regularly also in the future. Evert, your care of our good friends IRIS, INDIGO and INDY has always been impeccable. I have appreciated our several ‘modeling’ discussions and your input and comments on my work has been valuable. Before you know it, we will go ‘rundjes schaatsen’ and ‘biertjes drinken’ together again. Uli and Tjeerd, now we just need to convince Evert that Sweden is the place to be. Our “✂-kip” has been waiting long enough. I’ll see you in Uppsala. Yi and Marguerite, my room mates, it was a pleasure and now it seems that I will be replaced by another Swede. You just can’t have to many of us, can you! Pieter, without you taking care of the ‘synthese-zaal’ and keeping an eye on the chemists, synthesis would be very difficult. Cor and Durk, although we have not worked together directly I have always appreciated your advice and friendship.

The last five years in Holland has been a fantastic experience and I have appreciated the always relaxed and ‘gezellige’ atmosphere of Groningen. The several schaatstochten along the canals of Groningen and Friesland, the ‘Elfstedentocht’ of 1997, Sinter-Klaus and Zwarte-Piet, zeevissen, diving trips to Zeeland with Ad Fundum, Connie borrels, ‘terassjes pakken’, Schiermonnikoog and Waadloopen are just a few of events and places that for me have become, Holland. I will keep coming back and I will always feel at home.

Slutligen, mamma och pappa, det är omöjligt att i ord uttrycka hur tacksam jag är för ert förtroende och den frihet som ni alltid har givit mig. Ni är utan tvivel de bästa föräldrar man kan önska sig och utan er hade denna avhandling aldrig blivit skriven. Därför, som ett tecken på min tacksamhet, tillägnar jag er min avhandling.

…made it all worthwhile!

STELLINGEN

1. When your about to make the last move, to reach the top of a mountain, everything else is redundant.
   (Akkavare, Sarek, May 5th 1996)

2. To worry about tomorrow, is like paying interest on a loan you have not got yet.
   (an American proverb)

3. The very foundation of a successful climbing team, or any team, is indefinite trust.
   (Toulpagorni, Kebnekaise, June 1993)

4. "The dopamine hypothesis of schizophrenia" is too vague a hypothesis and must be considered as the pharmacological equivalence to the chemometric term "OVAT" (One Variable At Time).

5. The quality of a lecturer's slides/sheets is inversely proportional to his/hers professional experience (*i.e.*, age).

6. Providing 'mooie plaatjes' by means of sophisticated molecular modeling programs is easier said than done.

7. When you feel at home, away from home, it is not just the local people you have accepted, it is also their traditions.

8. Genuine "Kroppkakor" are prepared from raw pork and unboiled potatoes!

9. Neurophysiologist Louis Monti-Bloch has found a connection between the limbic system in the brain and the *vomeronasala* organ (VNO), located in the nose and thought to be sensitive for pheromones. If he is right, it is possible that doctors of the future will prescribe social intercourse with certain individuals as the treatment for, *e.g.*, schizophrenia and depression.

10. The "Elfstedentocht" is a diligent Dutch employer's worst nightmare.

11. QSAR is an iterative process that functions as *rational* drug design only when it is incorporated in a project at the same level as molecular modeling, synthesis and pharmacology.

Jonas Nilsson

Groningen, Netherlands, March 27th 1998

TESER

1. När du tar det sista steget, för att nå toppen av ett berg, då saknar allt annat betydelse.
   (Akkavare, Sarek, May 5th 1996)

2. Att oro sig inför morgondagen, det är som att betala räntan på det lån man inte fått.
   (Amerikanskt ordspråk)

3. En absolut förutsättningen för ett framgångsrikt klätter team, eller vilket team som helst, är ömsesidigt förtroende.
   (Toulpagorni, Kebnekaise, June 1993)

4. “The dopamine hypothesis of schizophrenia” är en alltför vag teori och kan anses som den farmakologiska motsvarigheten till den kemometriska termen EVIT (En Variable i Taget).

5. Kvaliten på en föreläsares dia-bilder/sheets är omvänt proportionell mot hans/hennes professionella erfarenhet (läs: ålder).

6. Att framställa snygga figurer med hjälp av sofistikerade molekyl modelering program är lättare sagt än gjort.

7. Där du känner dig hemma, fast du är borta, är det inte människorna där du har lärt dig acceptera, det är deras vanor.

8. Äkta Kroppkakor tillagas med rått fläsk och okokt potatis!

9. Neurofysiologen Louis Monti-Bloch har hittat en koppling mellan det limbiska systemet i hjärnan och det *vomeronasala* organet (VNO), som finns i näsan och anses vara känsligt för signalsubstanser eller feromoner. Om han har rätt är det möjligt att framtidens läkare kommer att föreskriva socialt umgänge, med vissa individer, som behandling mot t.ex. schizofreni och depression.

10. “Elfstedentocht” är en nitisk Holländsk arbetsgivares värsta mardröm.

11. QSAR är en iterativ process vilken fungerar som ‘*rational* drug design’ bara om den integreras i ett projekt på samma nivå som molekyl modulering, syntes och farmakologi.

Jonas Nilsson Groningen, Nederländerna, 27 Mars 1998