



3D-QSAR is a method based on the analysis and the comparison of the 3D molecular fields (steric, electrostatic etc..) produced in the surrounding of different compounds for the establishment of a correlation between the biological activities and the fields. We present with some details CoMFA, the first approach developed and validated in the steroid area. The discipline has become mature and led to the development of new methods such as HASL, CoMSIA, CoMMA, MS-WHIM, SOMFA, HQSAR, GRIND, Quasar, CoMASA, Wep.

Author(s): Claude Cohen (Synergix), Elie Cohen (Synergix), Hanoch Senderowitz (Predix Pharmaceutical)

Prerequisites: QSAR Principles and Methods

Number of Pages: 134 (134 Screens)

Last updated: January 2005

 Voice: available

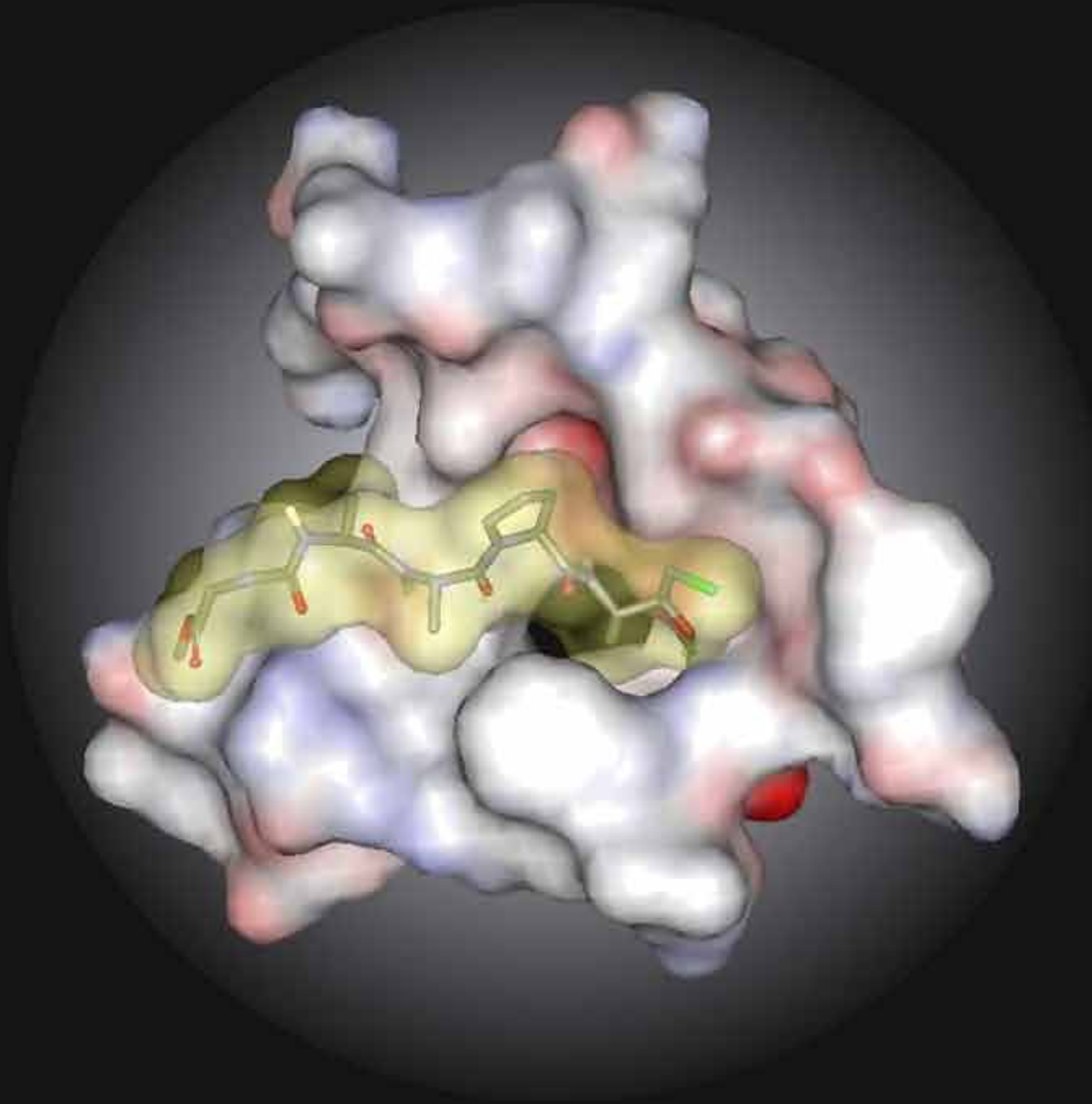


The topic Introduction contains the following 8 pages:

- Molecular Binding Occurs in 3D
- How Does the Receptor Perceives the Ligand?
- What is 3D-QSAR?
- Principle of 3D-QSAR Approach
- Intermolecular Forces
 - Electrostatic Field
 - Steric Field
- Difference between 2D-QSAR and 3D-QSAR

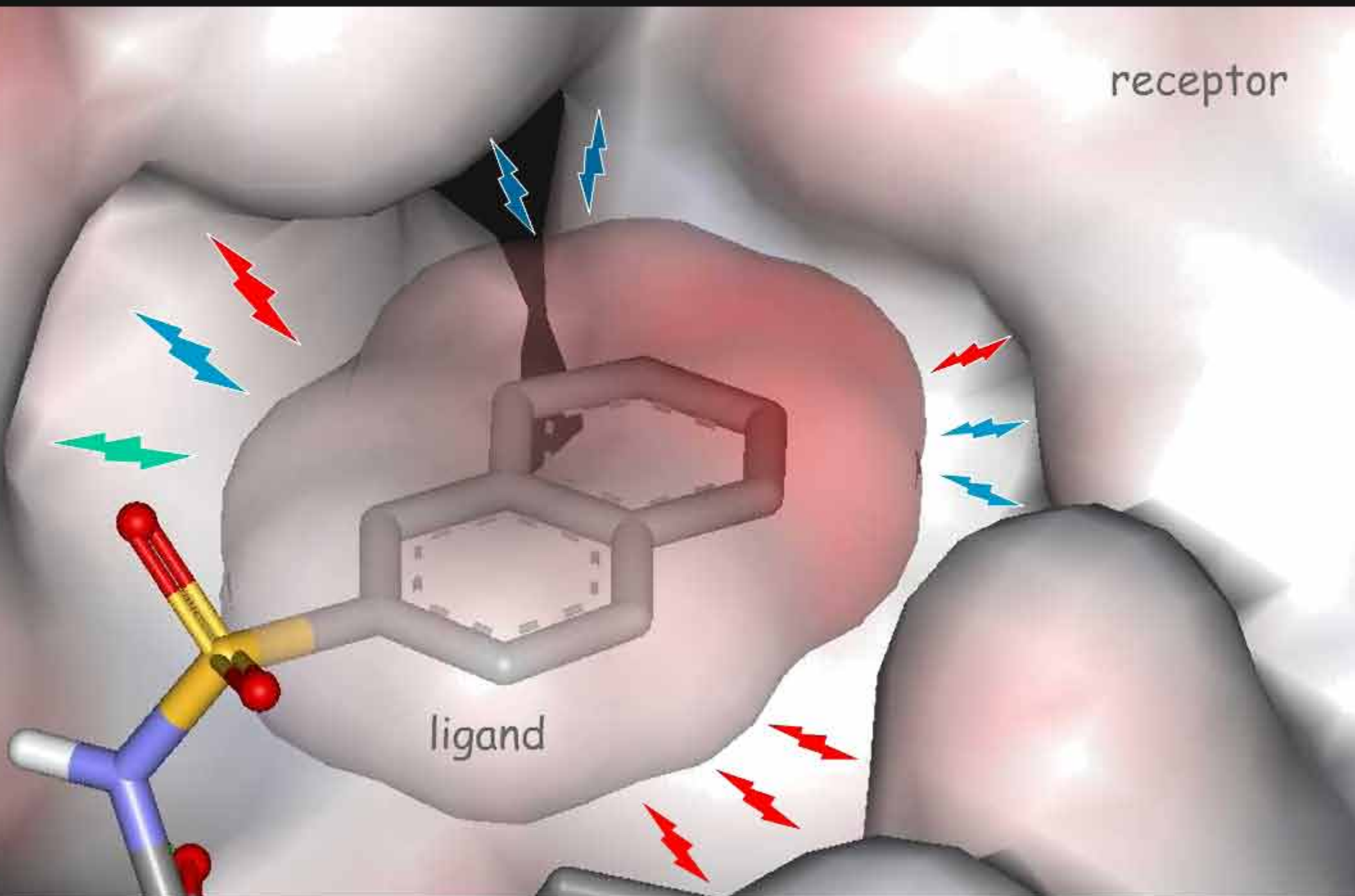
F2.1.1 Molecular Binding Occurs in 3D

The biological activity of a ligand depends of its affinity with its receptor; this can be understood in molecular detail by identifying the interactions and forces involved in the binding process. Molecular binding occurs in 3D.



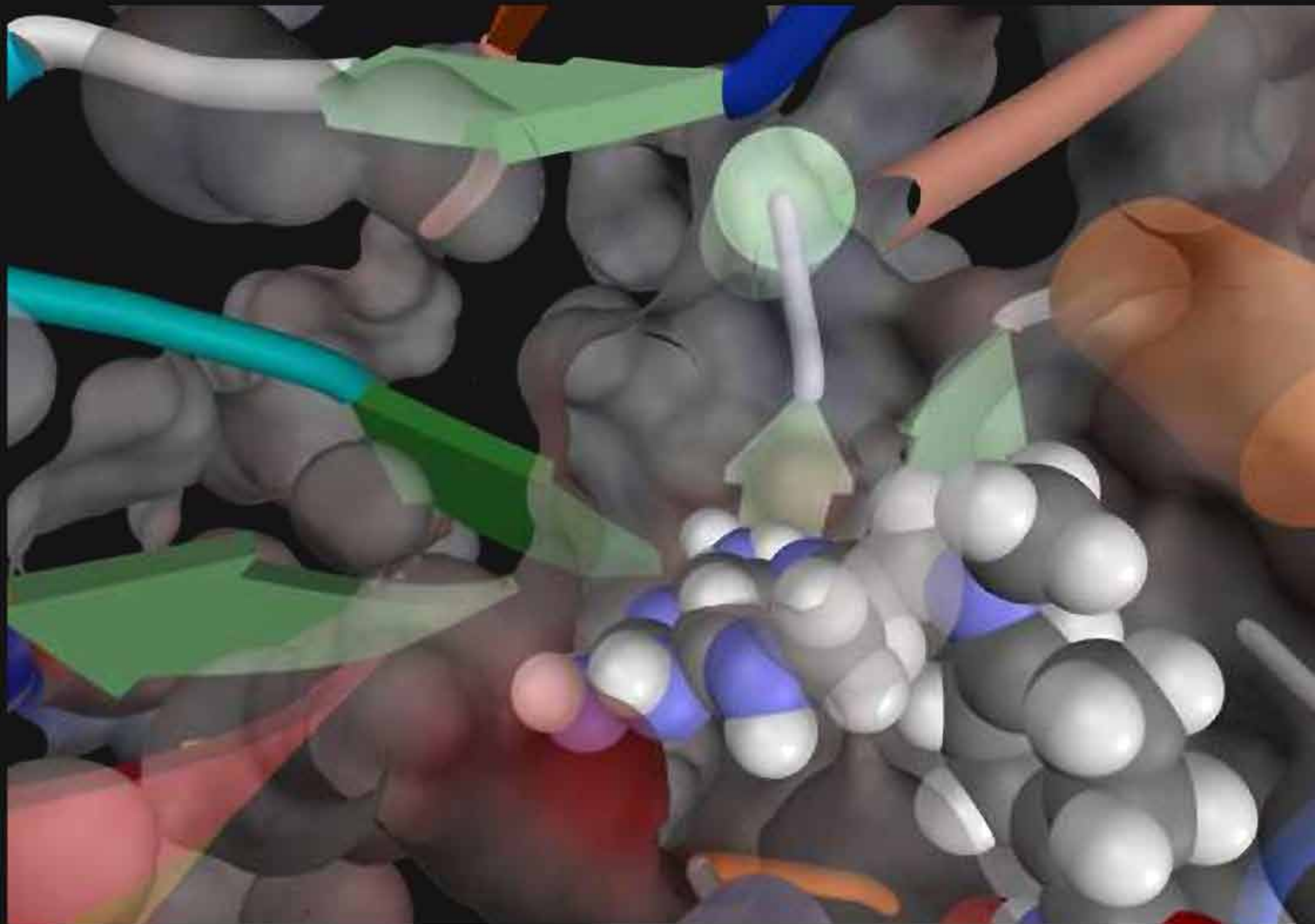
F2.1.2 How Does the Receptor Perceives the Ligand?

The biological receptor does not see a ligand as a set of atoms and bonds, rather it perceives a shape that carries complex forces. In the 3D-QSAR framework these forces are considered to be determined predominantly by electrostatic and steric (van der Waals) interactions.



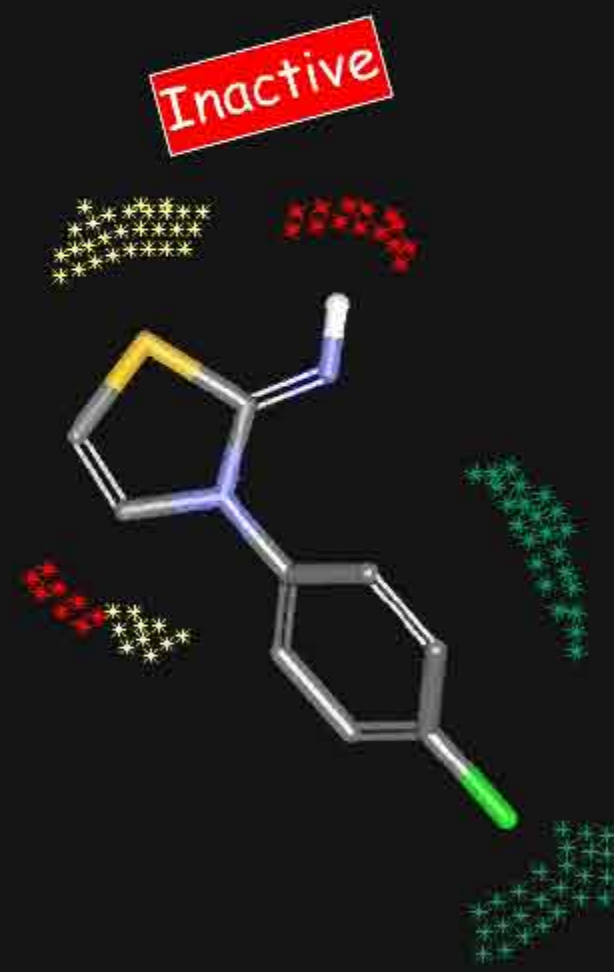
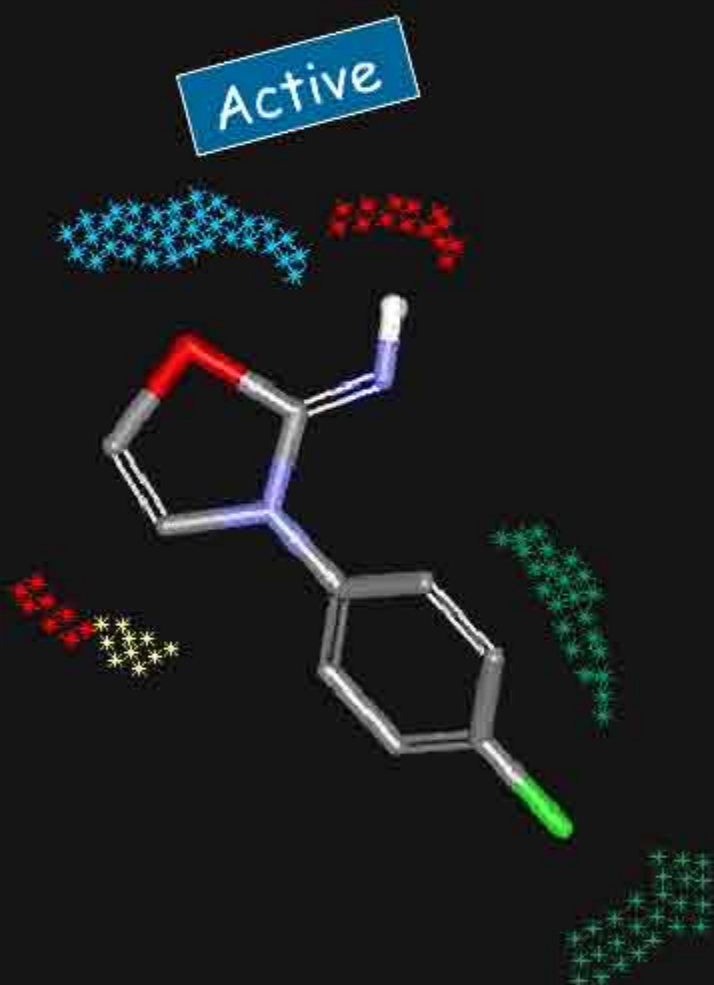
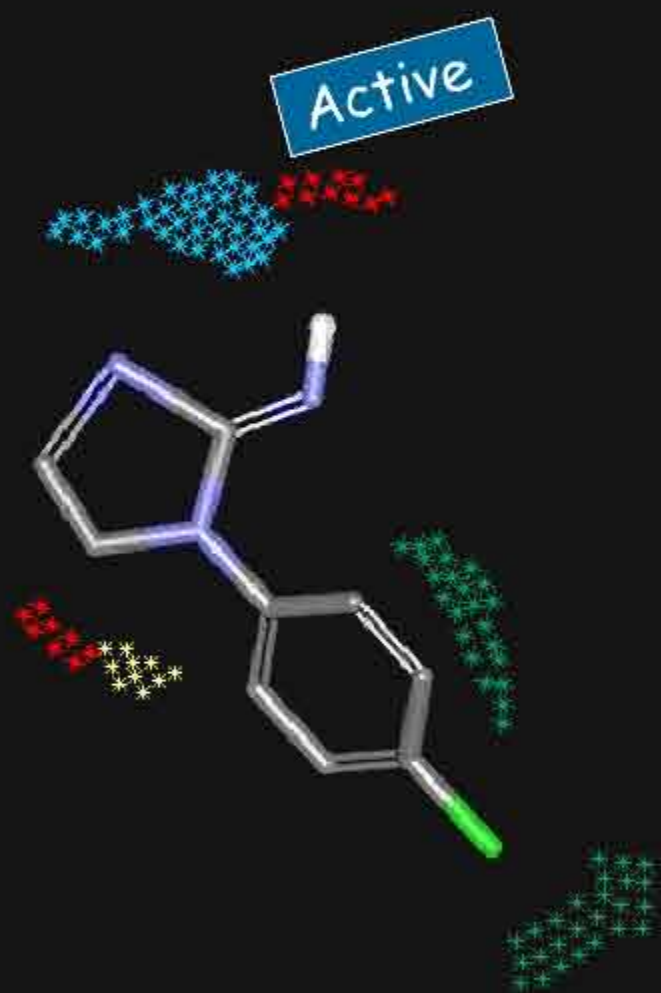
F2.1.3 What is 3D-QSAR?

3D-QSAR is a method based on statistical correlation analyses enabling the comparison of 3D molecular forces ("fields") produced in the vicinity of different compounds to find correlations between biological activities and fields. This method generally applies in situations where the structure of the receptor is not known.



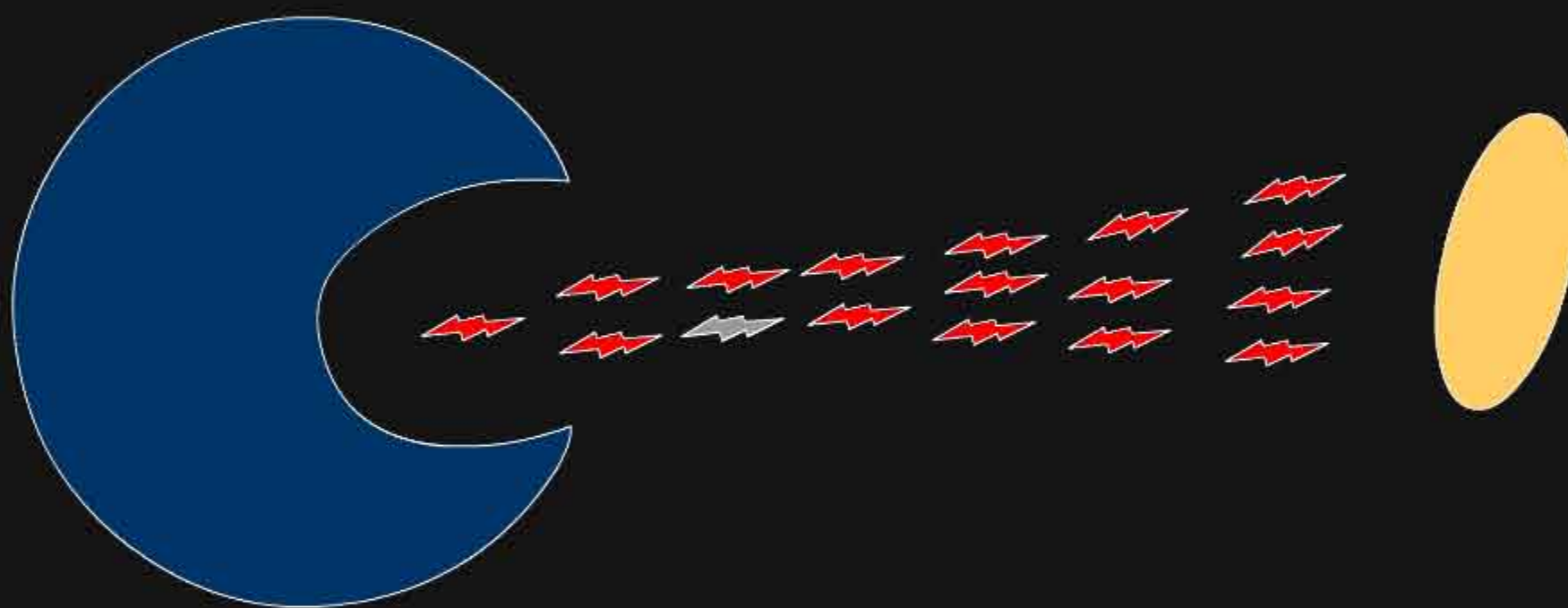
F2.1.4 Principle of 3D-QSAR Approach

3D-QSAR is based on the mapping and the comparison of the steric and electrostatic fields around a set of ligands, to establish a 3D quantitative structure-activity relationship (3D-QSAR).



F2.1.5 Intermolecular Forces

Electrostatic energy can be expressed as the inverse of the distance of the interacting atoms, whereas the steric energy depends on its inverse twelfth power. Thus, at long distances the electrostatic field drives the approach of the ligand to the active site, whereas at short ranges the steric forces become more important and control the final step of binding.



electrostatic potential - Long range

$$\sim \frac{1}{r}$$



VdW potential - Short range

$$\sim \frac{1}{r^{12}}$$

F2.1.6 Electrostatic Field

Electrostatic interactions occur between polar or charged groups. The electrostatic interactions between two molecules A and B are calculated as the sum of the interactions between point charges using Coulomb's law; they can be either attractive or repulsive. Since the electrostatic term can be expressed as the inverse of the distance, the electrostatic field is far from being negligible even when the groups involved are rather far apart (e.g. 10 angstroms or more).

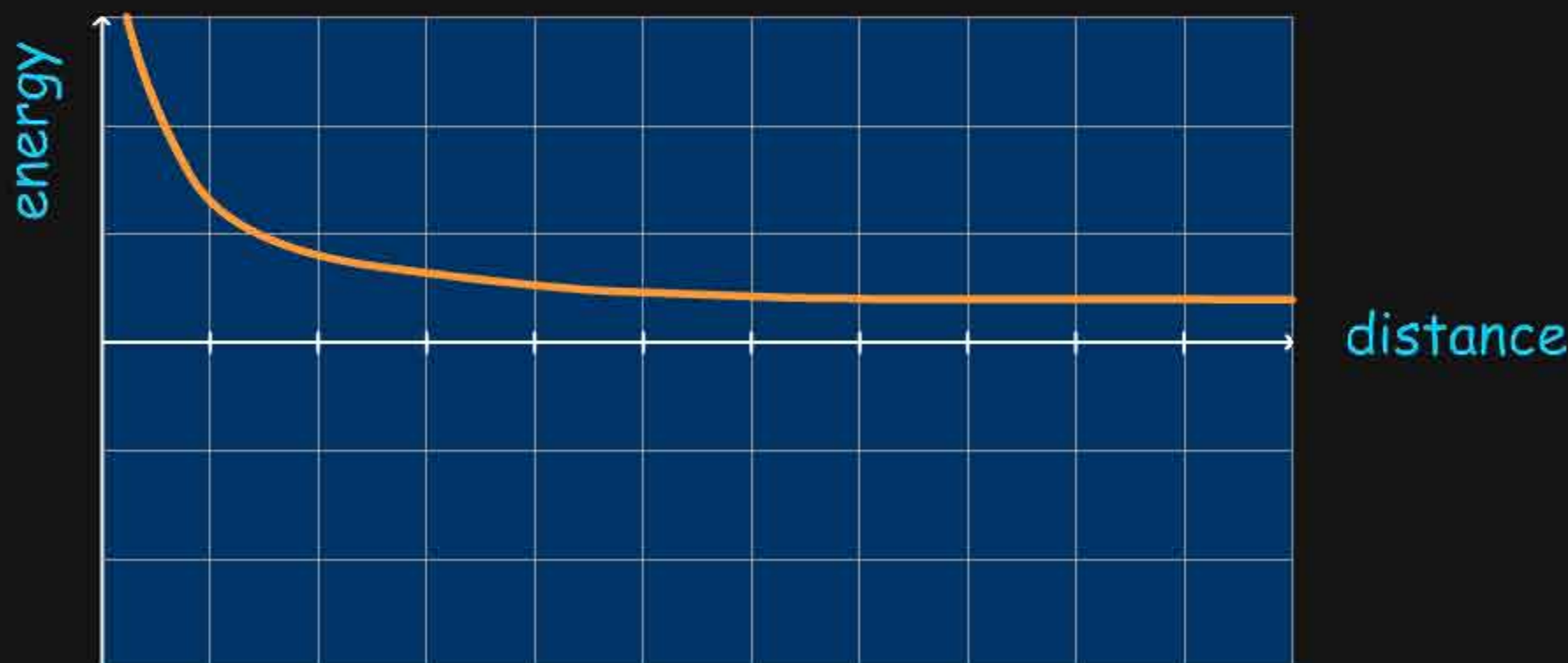
● Same Charges

● Opposite Charges



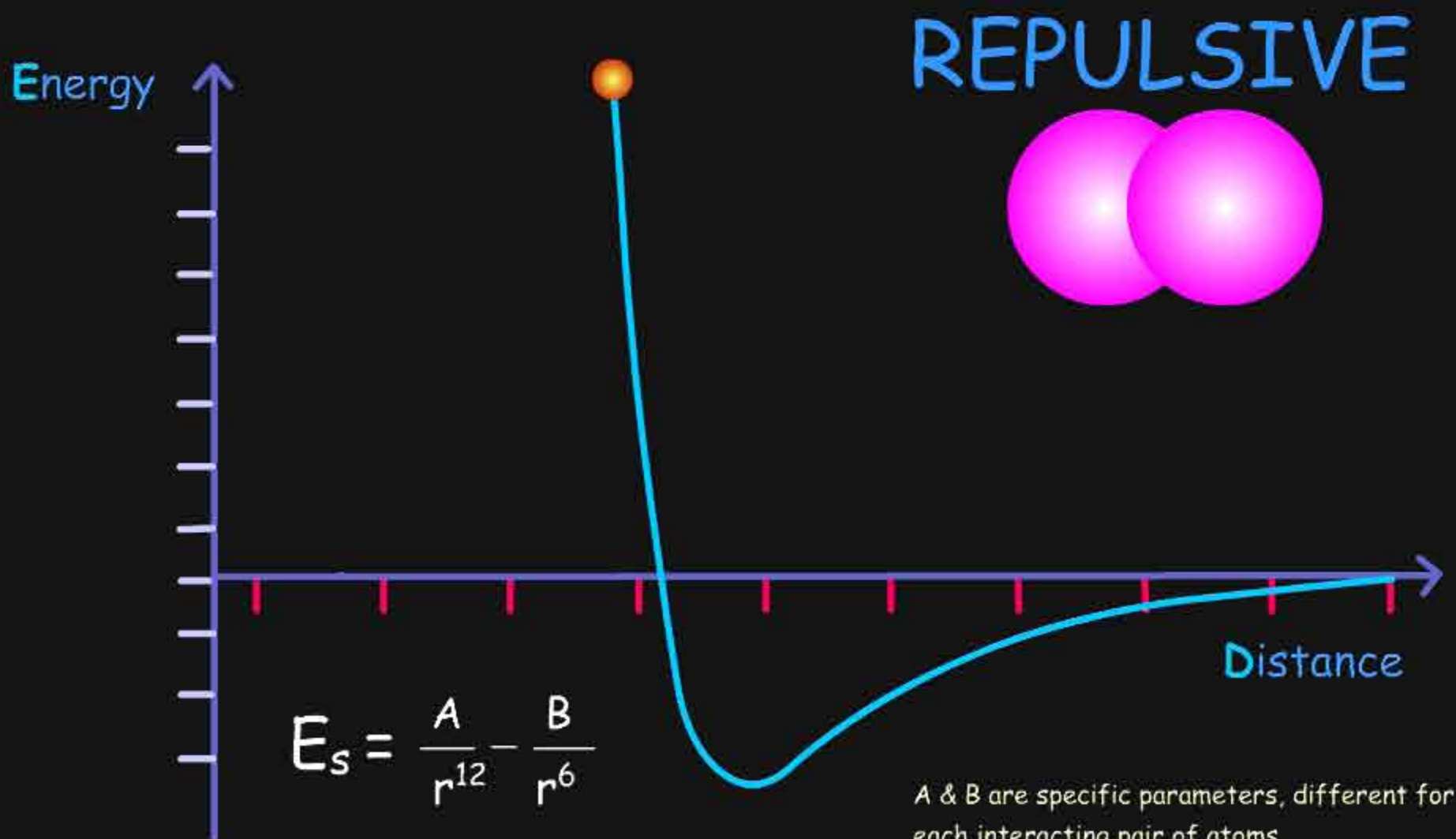
$$E_{es} = \frac{q_a q_b}{r_{ab}} > 0$$

repulsive potential



F2.1.7 Steric Field

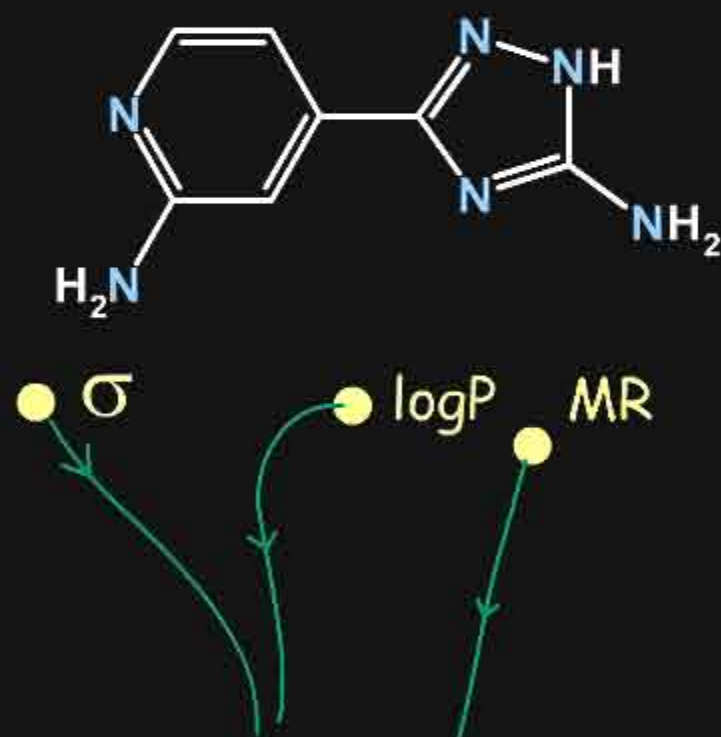
The steric potential describes the non-electrostatic interactions between non-bonded atoms. The associated forces (called 'van der Waals' forces) can be either repulsive or attractive, depending on the distance between the atoms involved. At short distances, there is repulsion between atoms (due to the interpenetration of their electronic clouds), and at long distances, there is a small attraction (dispersion forces).



F2.1.8 Difference between 2D-QSAR and 3D-QSAR

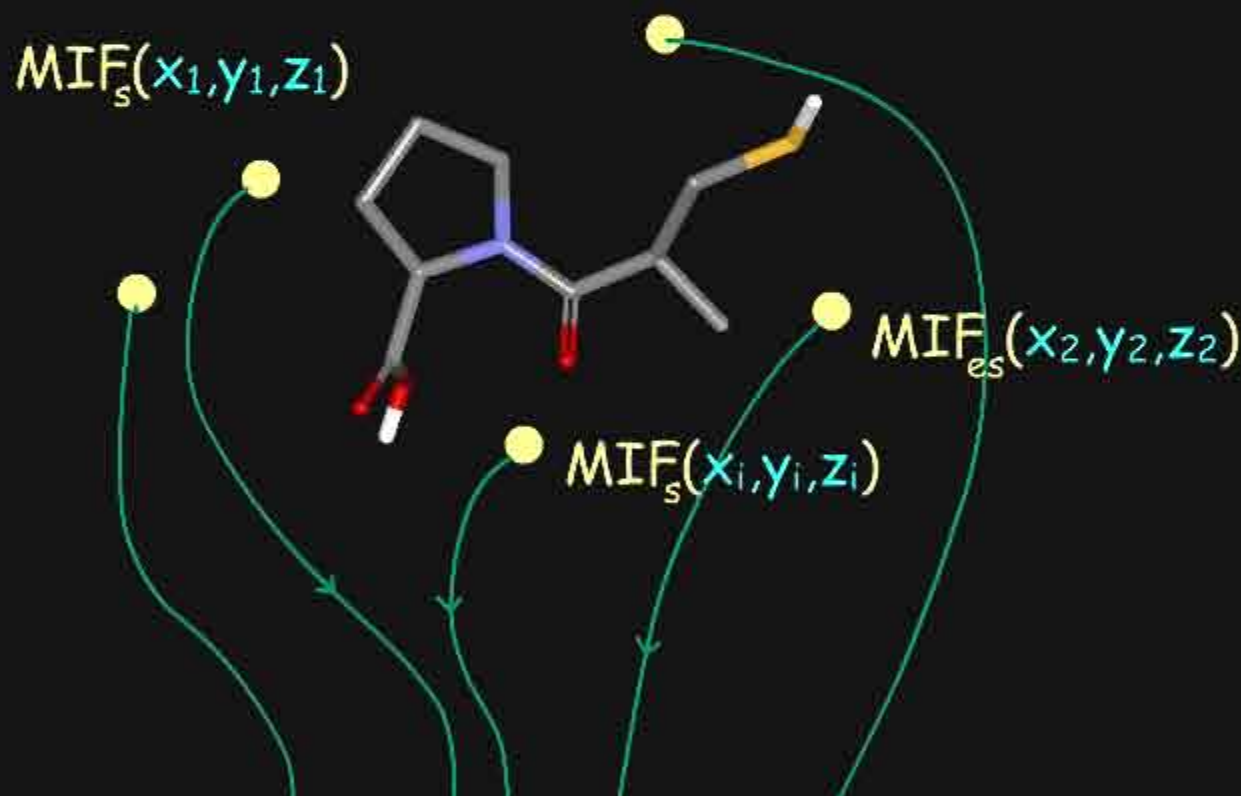
QSAR approaches aim to establish relationships between biological activities and chemical structure; however in classical QSAR molecular properties are described by parameters that are NOT x,y,z dependent (e.g. logP, MR, E_s , σ , π etc...), whereas in 3D-QSAR they are represented by a set of values of (x,y,z) functions, measured at many different locations in the space around the molecules. One consequence of this difference is that there are many more descriptors in 3D-QSAR than in classical QSAR.

Classical QSAR



descriptors are
x,y,z - INDEPENDENT

3D-QSAR



descriptors are
x,y,z - DEPENDENT



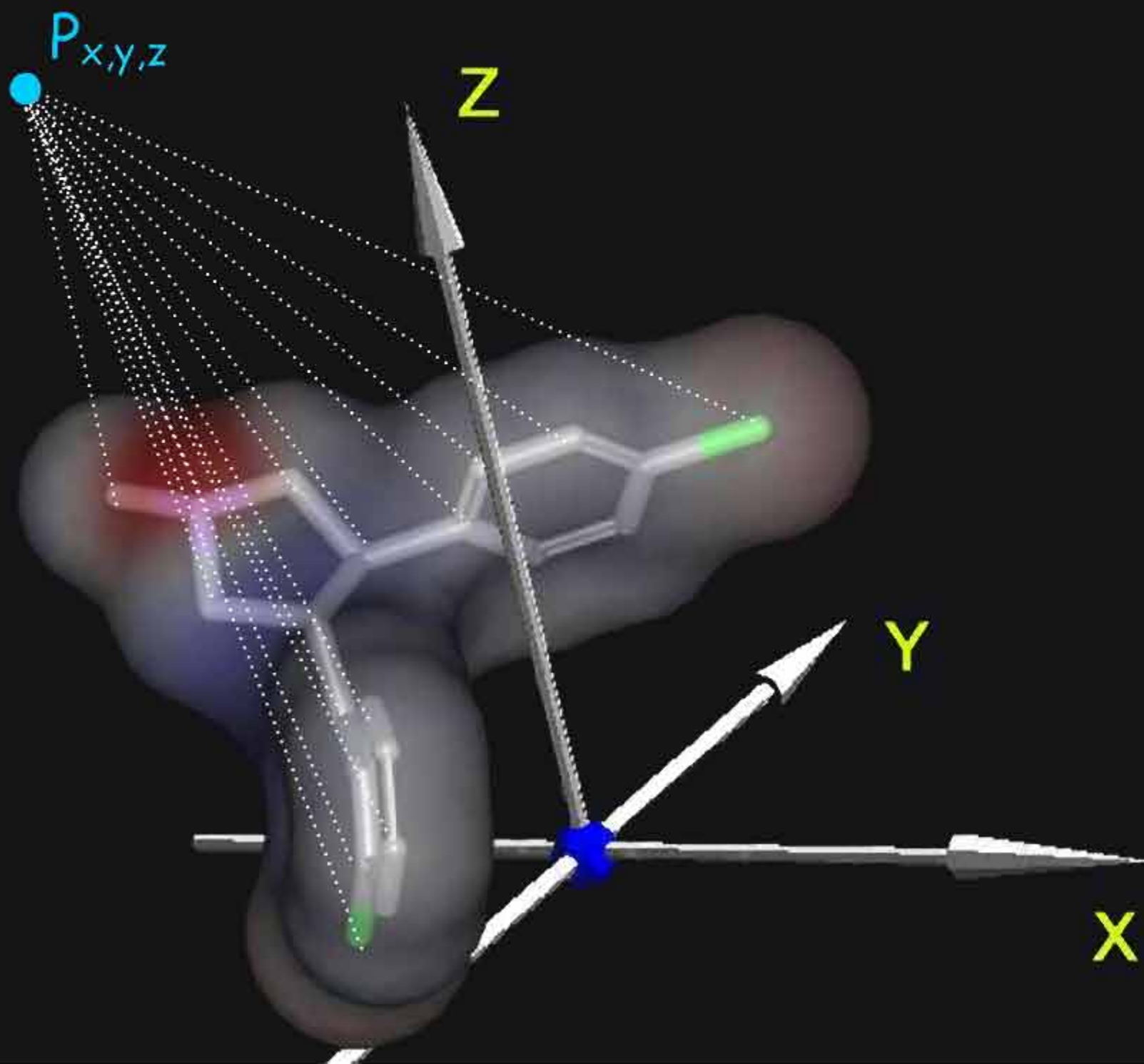
F2.2 Molecular Interaction Fields (MIF)

The topic Molecular Interaction Fields (MIF) contains the following 11 pages:

- Interaction Field Surrounding a Molecule
- Perception of Interaction Fields
- The Probe Concept
 - Probing Steric Field with Single Atom Probe
 - Probing Electrostatic Field with Single Atom Probe
 - Multi-Atom Probes
- 3D Lattice and Grid Points to Capture the MIFs
- Calculating the Electrostatic Field
- Calculating the Steric Field
- Visualization of MIFs with Iso-Potential Surfaces
- Other Molecular Interaction Fields

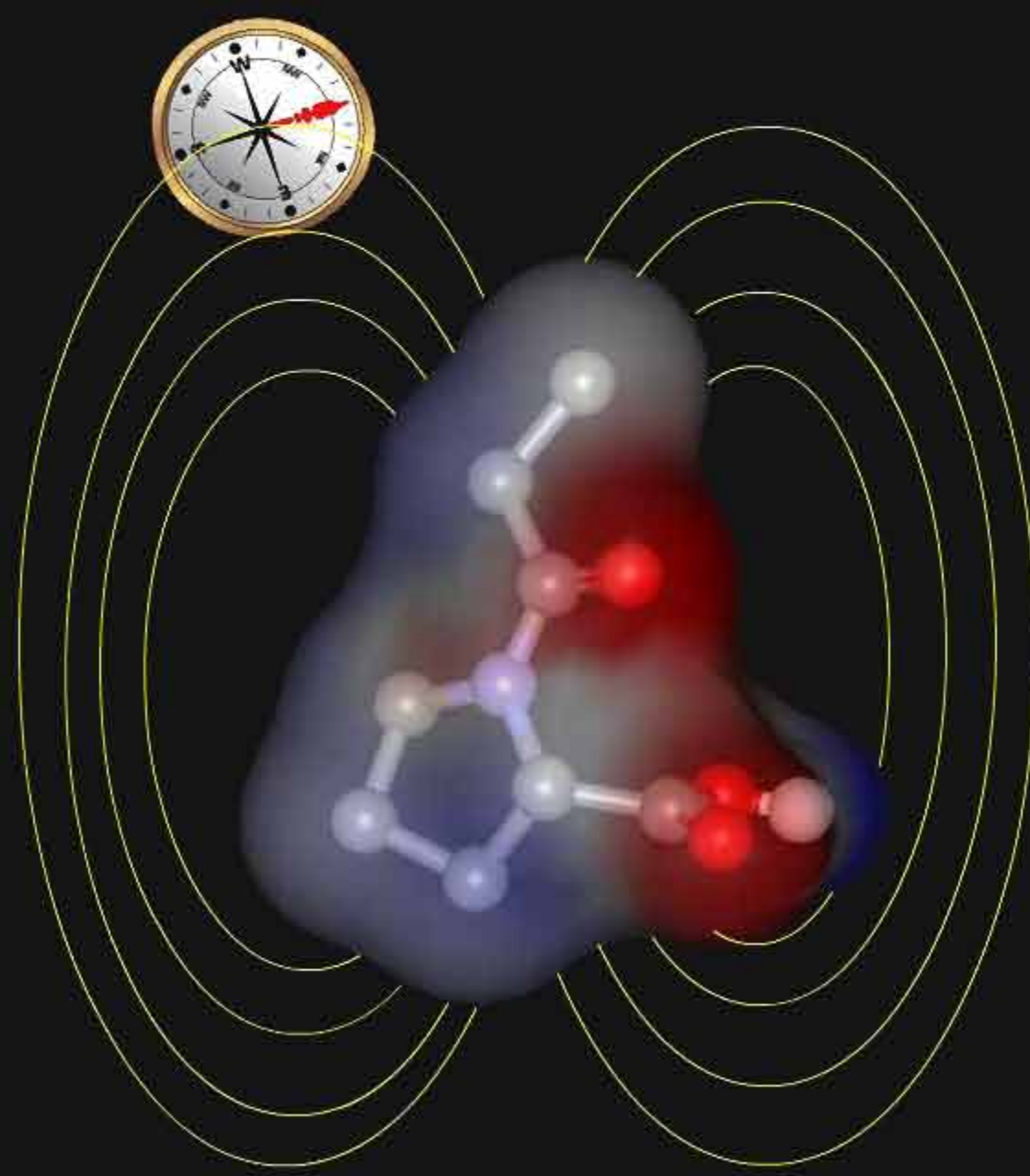
F2.2.1 Interaction Field Surrounding a Molecule

Suppose that you have a molecule and a molecular property, for example the electrostatic field that can be calculated at any point from your molecule. The 3D distribution of the interaction field can provide relevant information concerning the properties of the molecule.



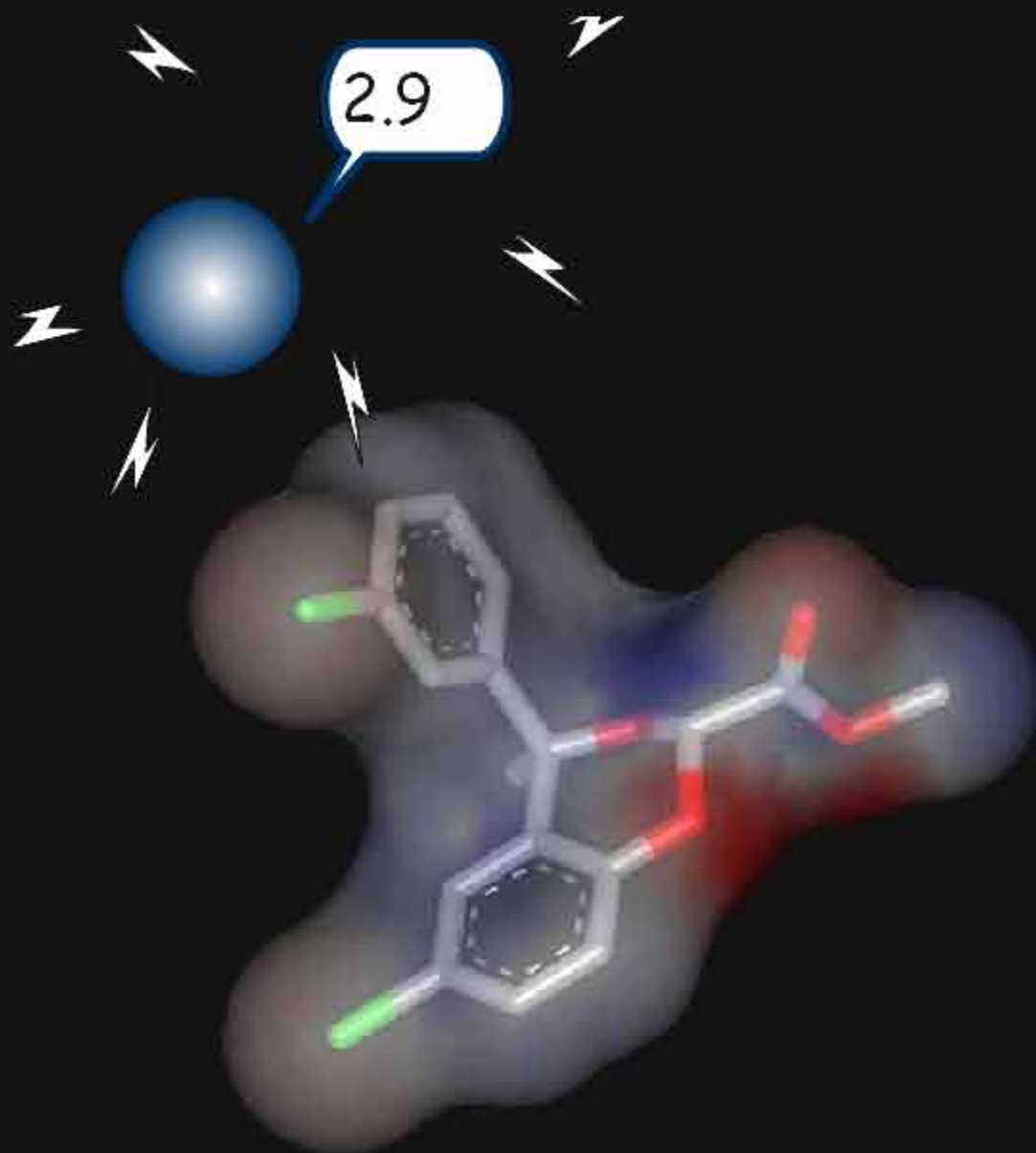
F2.2.2 Perception of Interaction Fields

A field can be perceived only if there is a proper receiver able to interact with it. Take the example of the earth's magnetic field: you feel its existence if you have a compass, otherwise you cannot know if it exists. The situation is the same for molecular interaction fields: they can only be measured with the use of appropriate "probes"; this is discussed in the following pages.



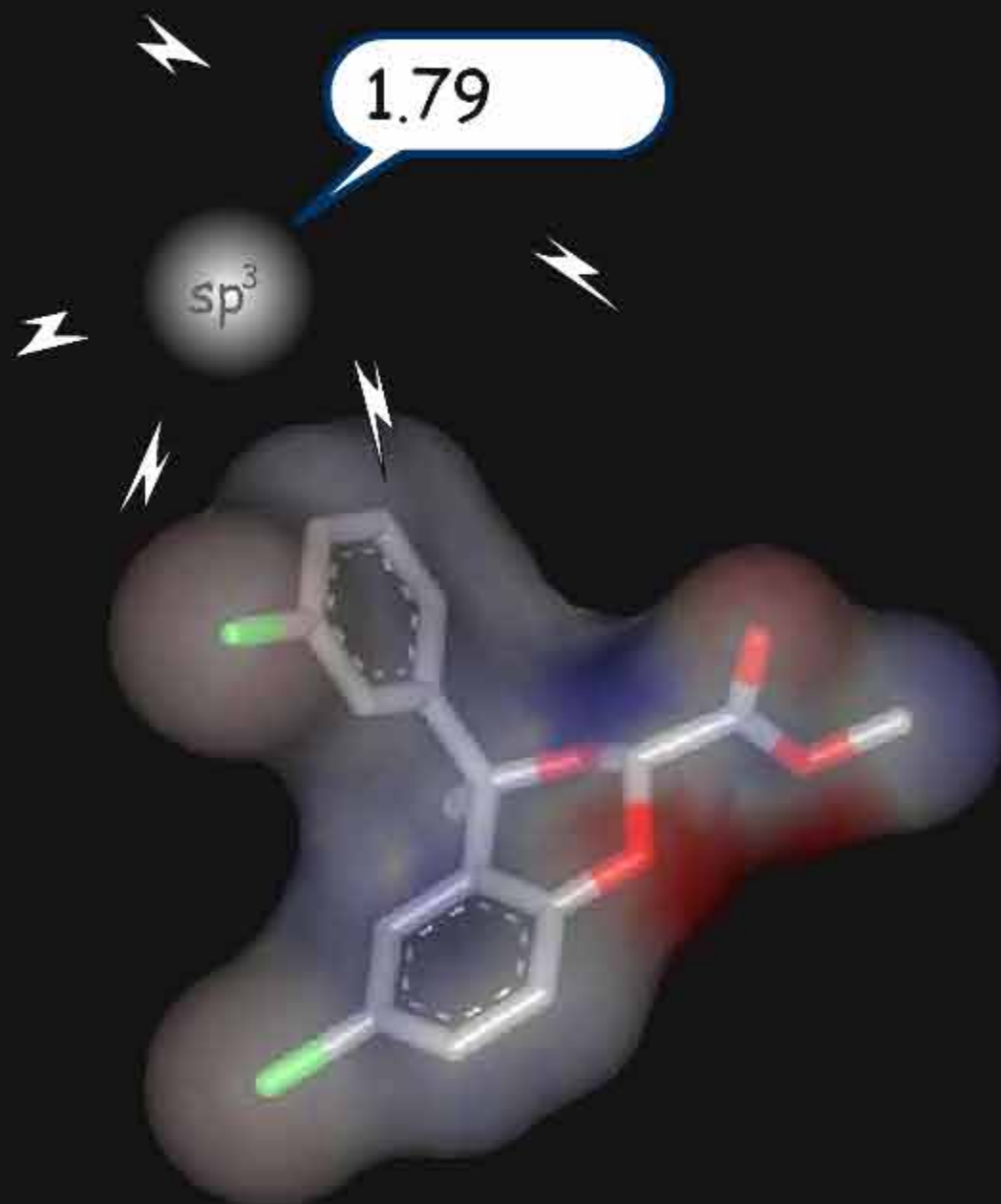
F2.2.3 The Probe Concept

To test for the presence of a field and to measure it requires the use of probes with associated energy functions. Usually a probe atom is employed, which is placed at well selected points in the space, to quantitatively measure the value of the field created by the molecule at the point considered. The probe must be of the same type of the field to be measured (e.g. van der Waals probes for steric fields, charged probes for electrostatic fields).



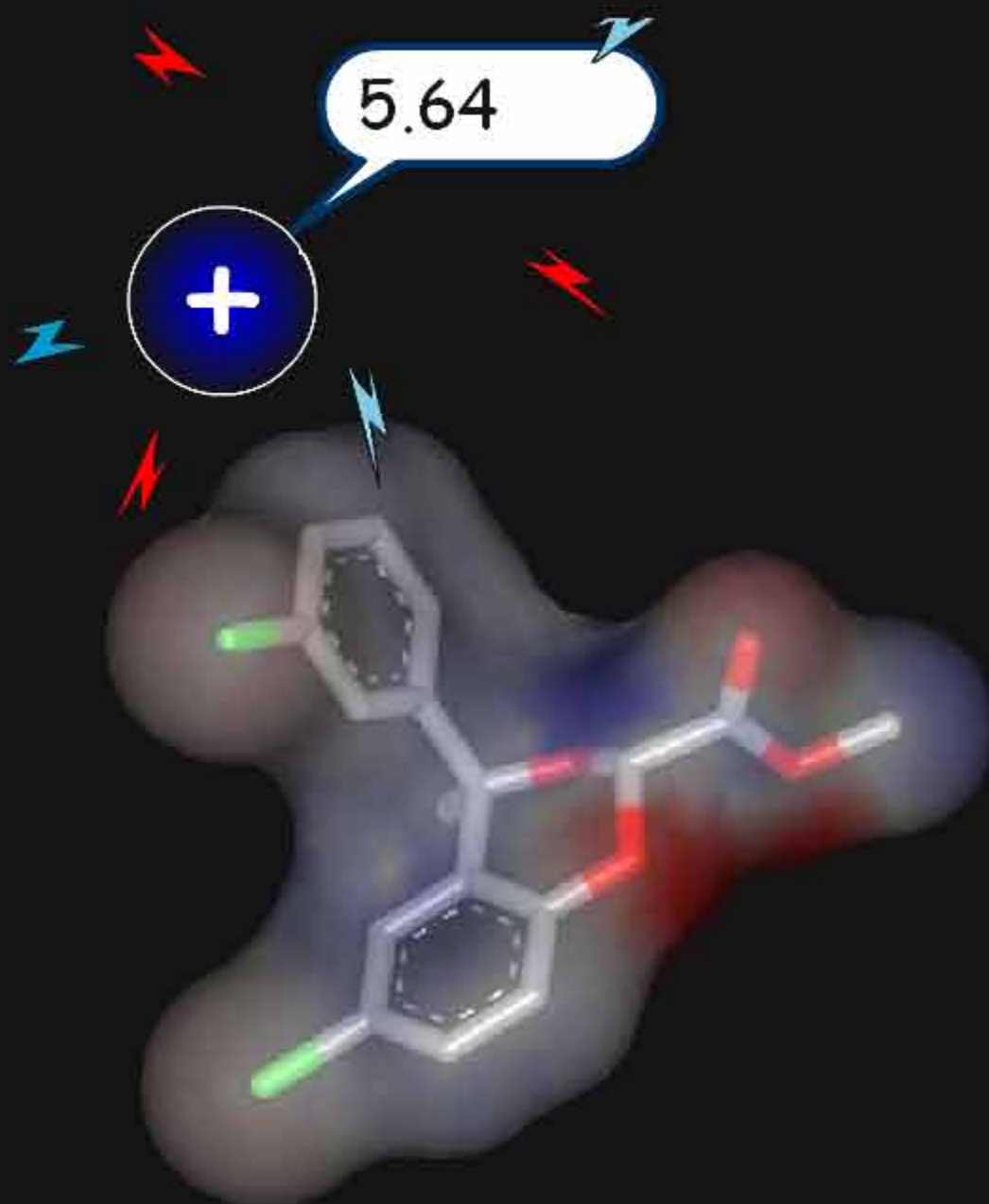
F2.2.4 Probing Steric Field with Single Atom Probe

The value of the steric field is calculated at different points in the space around a given molecule. The probe atom normally used to measure the steric field is a carbon sp^3 .



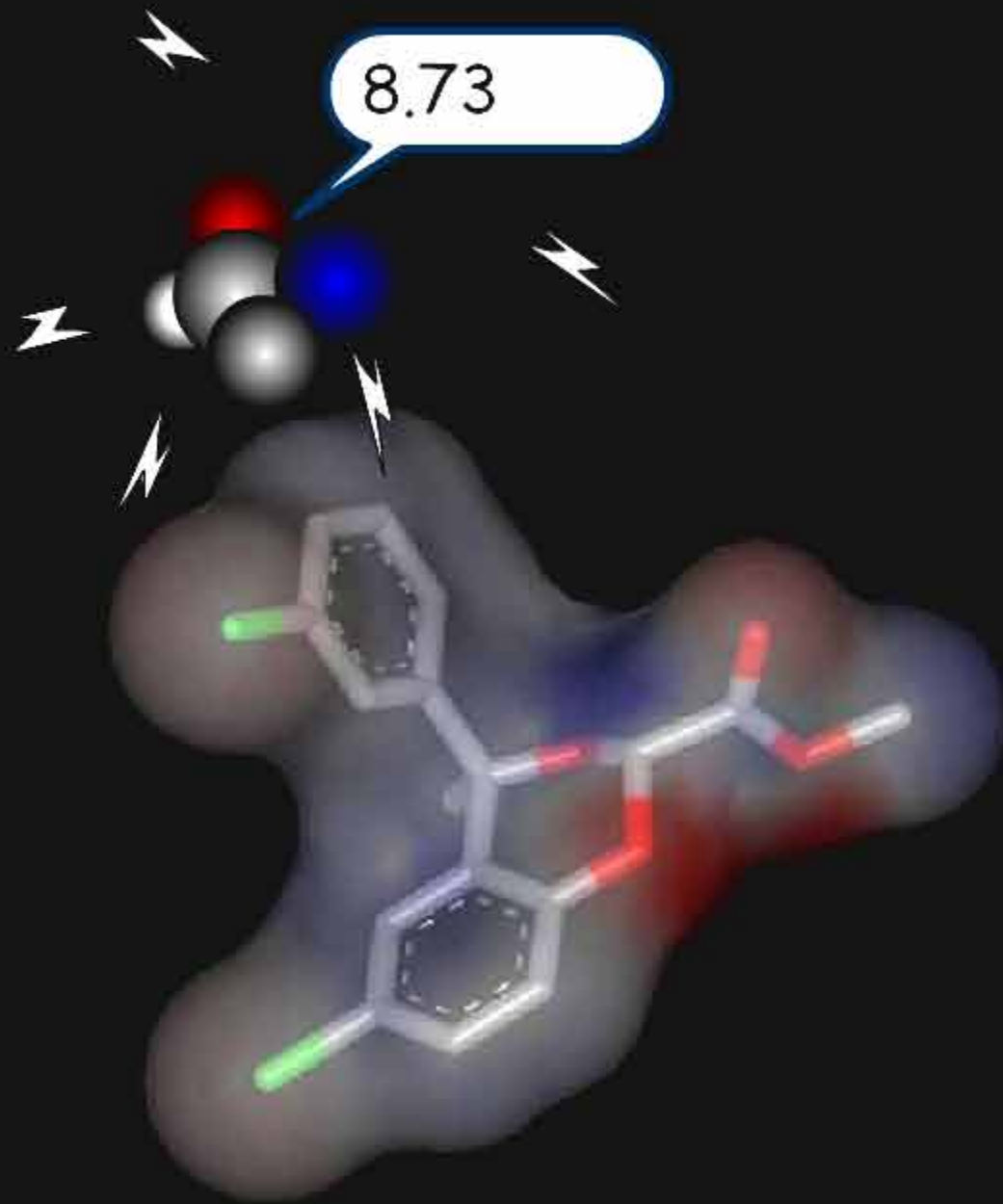
F2.2.5 Probing Electrostatic Field with Single Atom Probe

The electrostatic field is measured at different points in the space around the molecule. The probe atom normally used is a carbon sp^3 with a charge of +1.



F2.2.6 Multi-Atom Probes

The probe concept has been extended to a whole range of molecular probes such as CH_3 , NH_2 , CONH_2 , H_2O , NH_3^+ , COO^- etc...

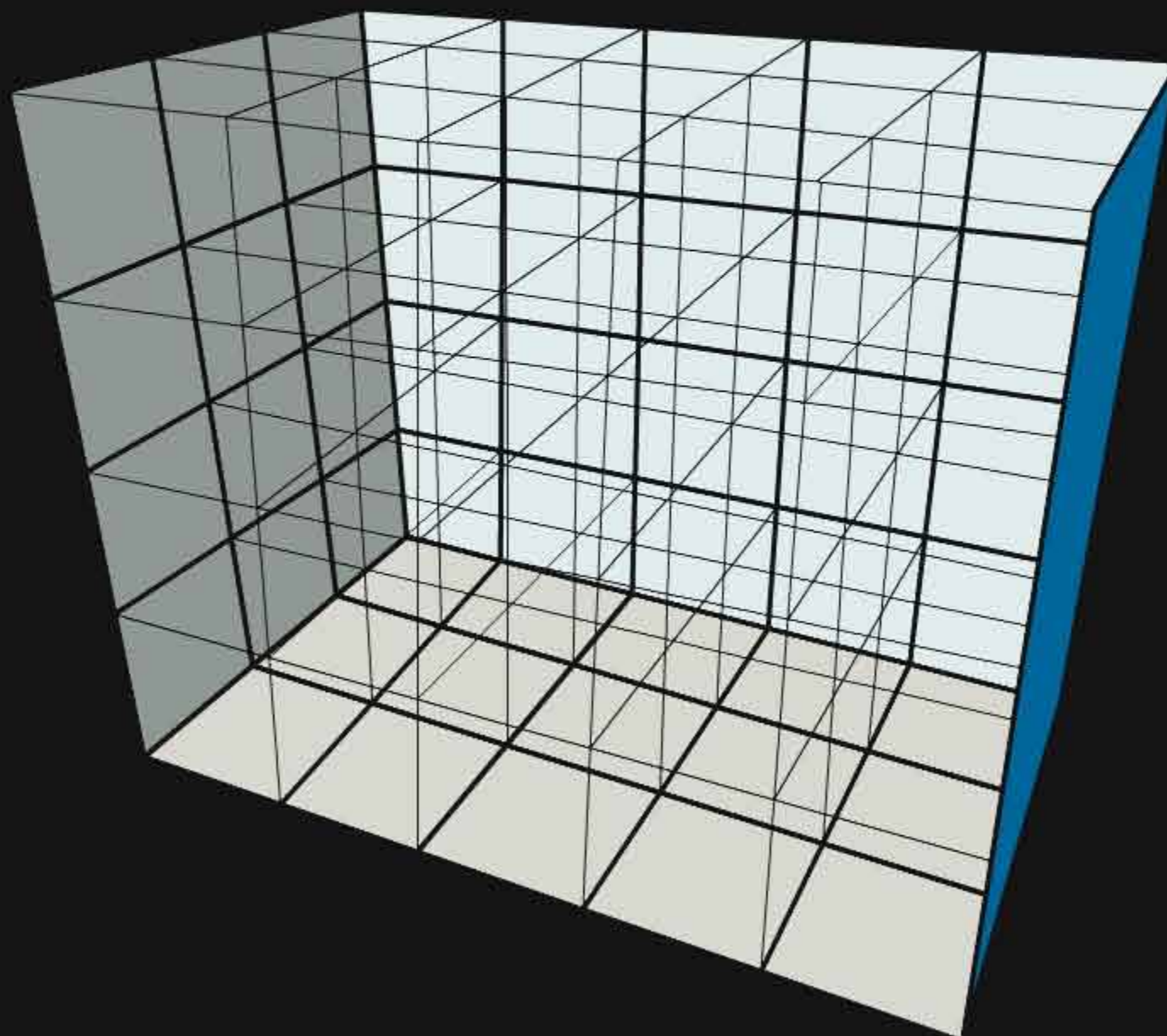


F2.2.7 3D Lattice and Grid Points to Capture the MIFs

To simplify calculations of the fields created around a molecule, the method consists of superimposing a 3D lattice defining grid points regularly distributed in space, and calculating the interaction energy between the molecule and the probe at each grid point, using a potential energy function. The lattice makes it possible to sample the space with a finite number of points with MIFs that can be calculated in a reasonable amount of time.

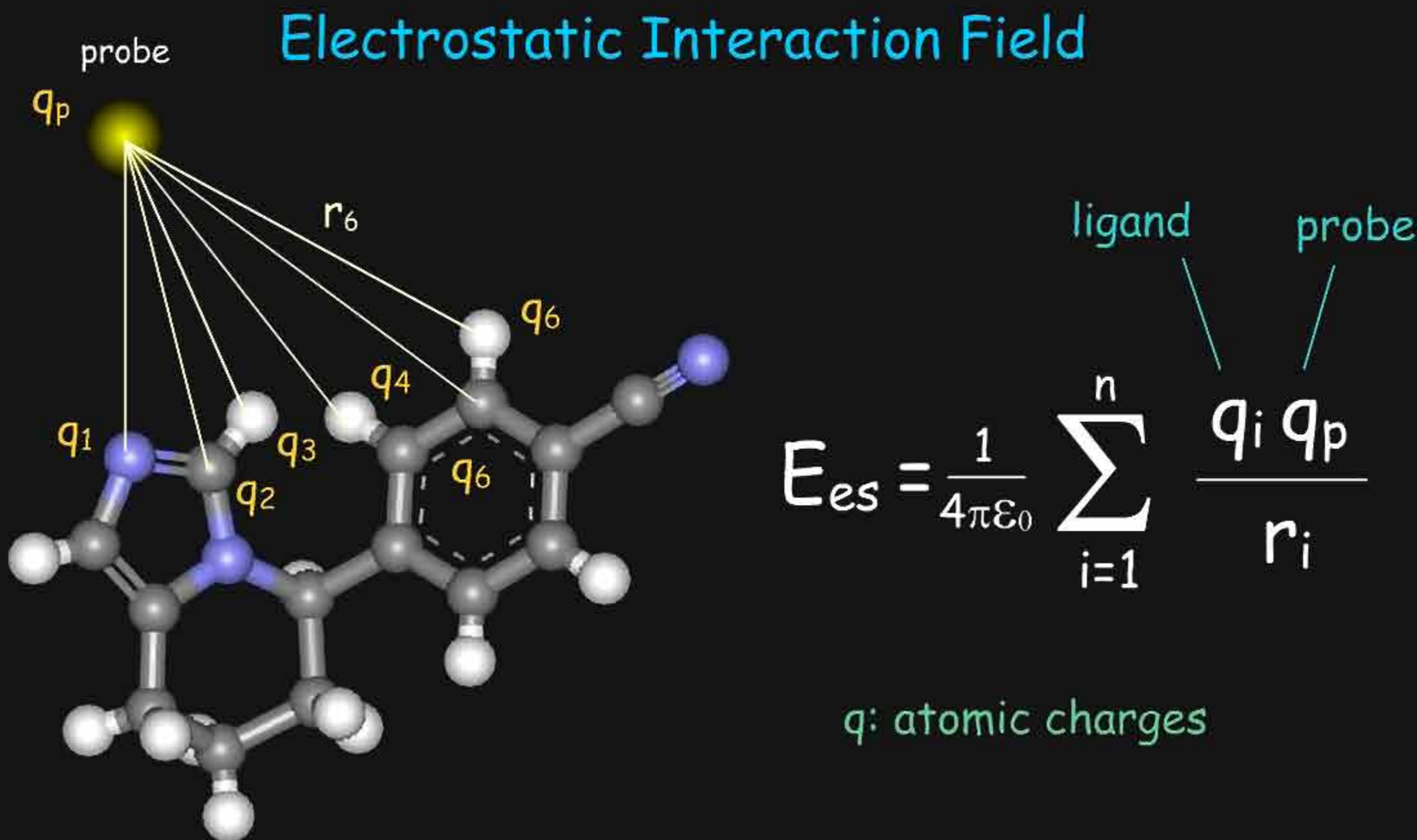
The Grid

The Probe



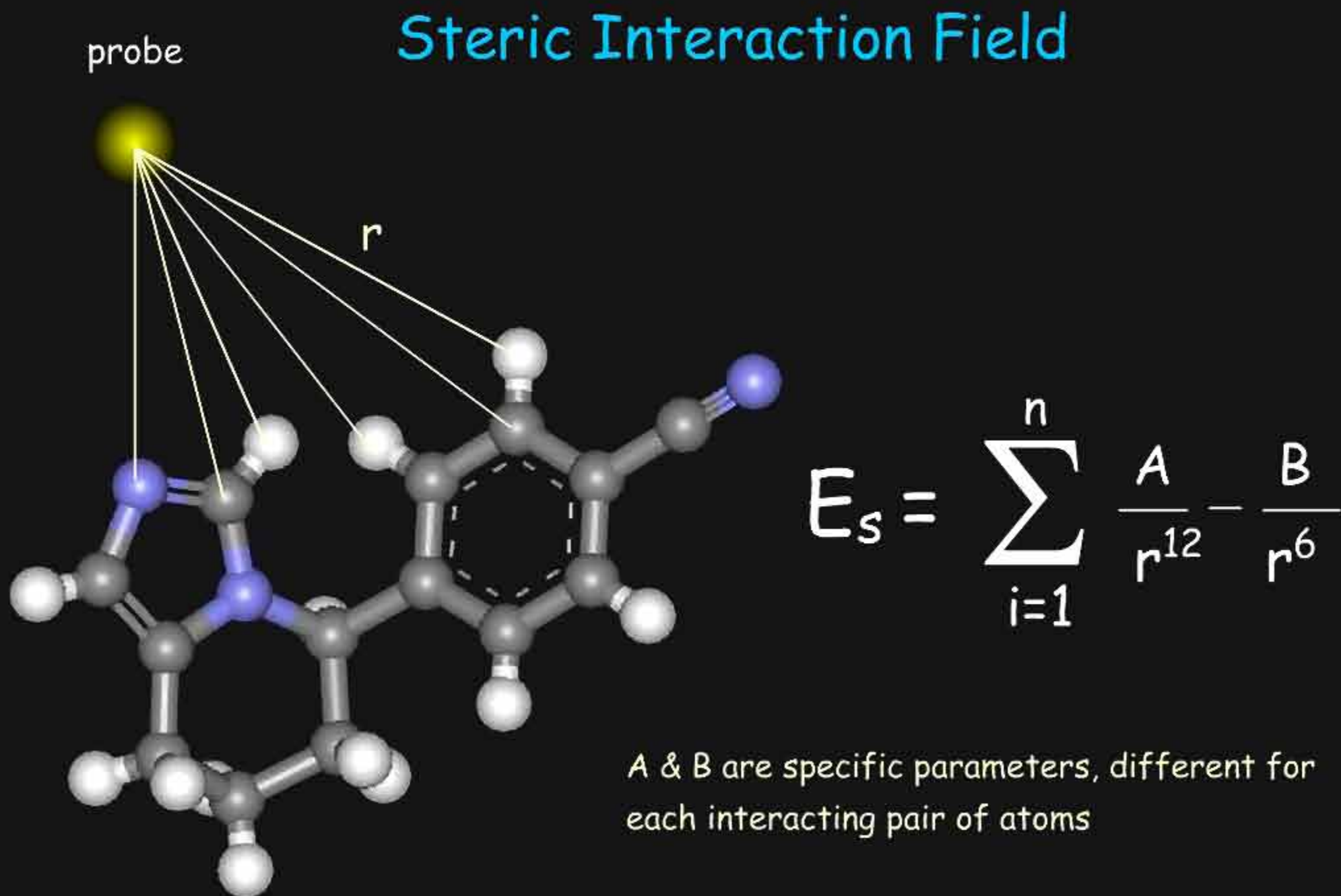
F2.2.8 Calculating the Electrostatic Field

The electrostatic field is obtained by calculating the electrostatic interaction energy between the molecule and the probe at each grid point using Coulomb's law.



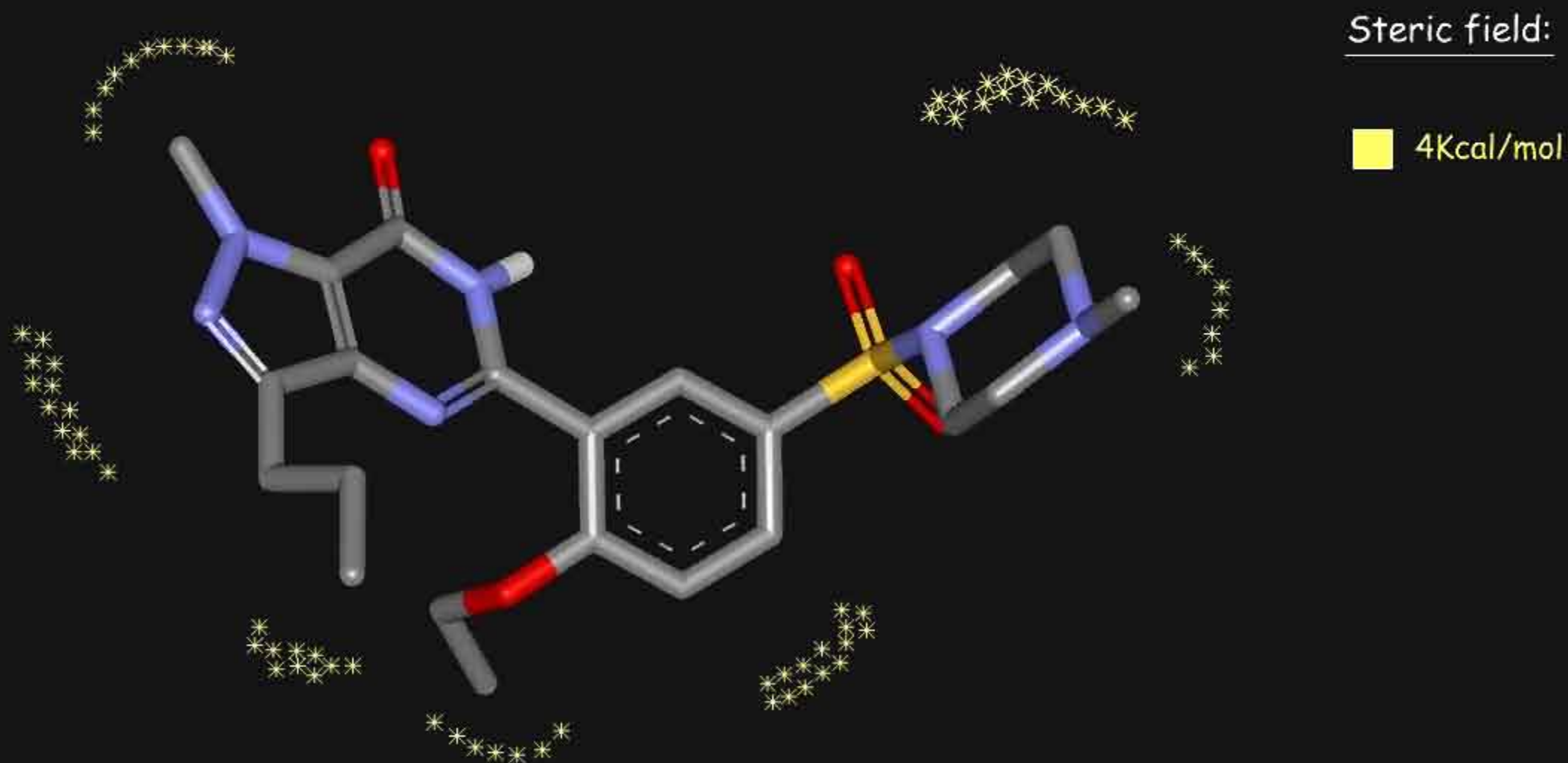
F2.2.9 Calculating the Steric Field

The steric field is obtained by calculating the van der Waals interaction energy between the molecule and the probe at each grid point using for example a 6-12 Lennard-Jones potential.



F2.2.10 Visualization of MIFs with Iso-Potential Surfaces

Molecular interaction fields can be visualized by drawing iso-value surfaces around the molecule. An iso-value surface is a 3-dimensional surface connecting all points of the same value. In the course of a study many such surfaces are analyzed with the aim of deriving useful knowledge in structural terms.



F2.2.11 Other Molecular Interaction Fields

Besides the steric and electrostatic energies, other fields can be generated, depending on the probe and the potential energy function used to calculate the interaction. Additional fields include: interaction energies with functional groups, molecular lipophilicity field, hydrogen bond donor and hydrogen bond acceptor fields.

- Steric field
- Electrostatic field
- Molecular lipophilicity
- Hydrophobic field
- H-bond donor field
- H-bond acceptor field



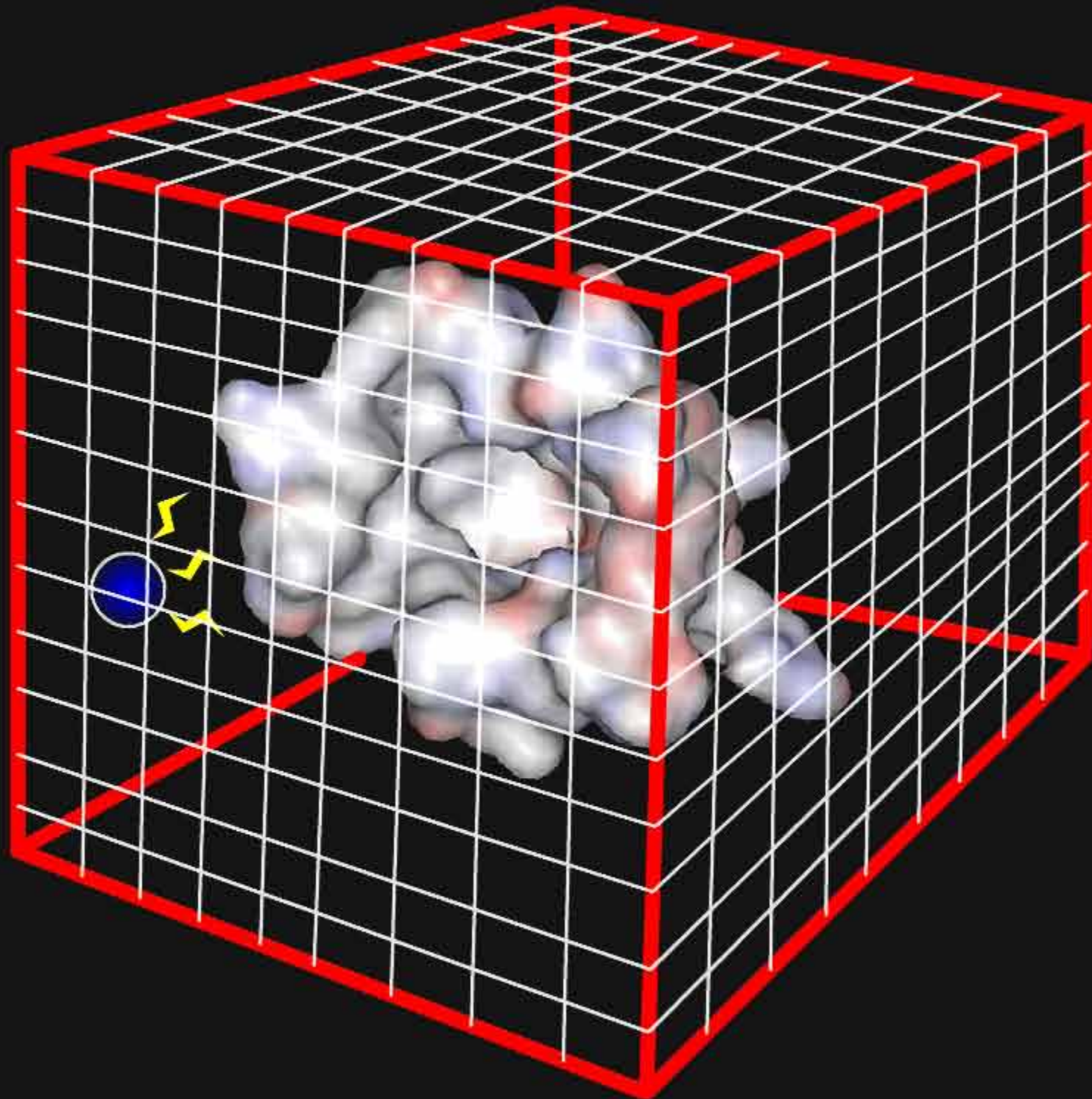
The topic The GRID Approach contains the following 14 pages:

- The GRID Approach
- GRID: a Structure-Based Approach
- Probing the Nature of the Active Site
- The GRID Probes
- Integration of GRID with Other Programs
- Typical Use of GRID
- Outline of a GRID Calculation
 - 3D Coordinates of the Protein
 - Binding Site to be Explored
 - Selection of Probes
 - Run of GRID
 - Output of GRID
- Total Number of Calculations
- ...

For the entire list, see the navigation panel.

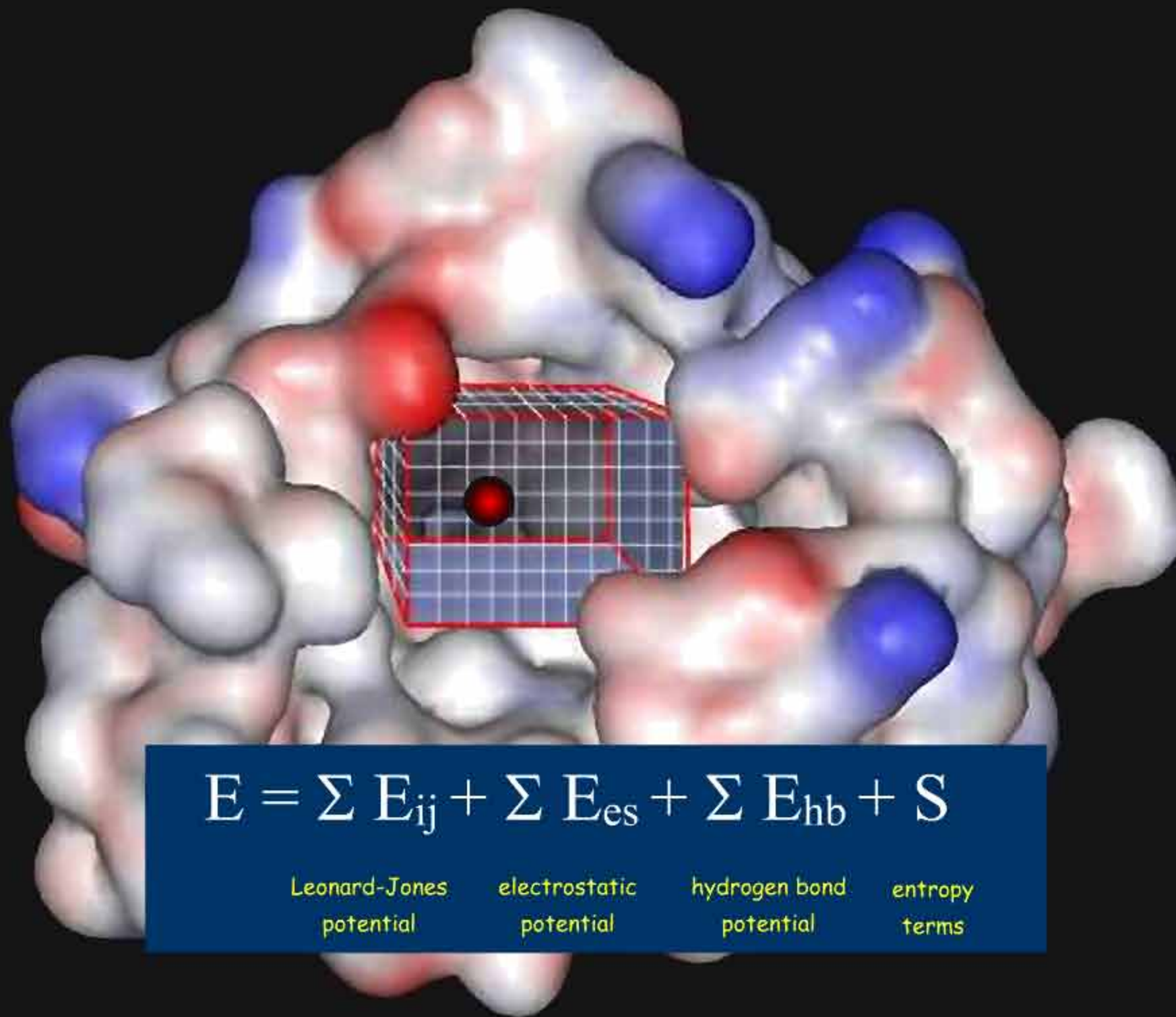
F2.3.1 The GRID Approach

The first program based on the calculations of MIFs was GRID, introduced by Peter Goodford in 1985 as a structure-based method to analyze the active sites of proteins.



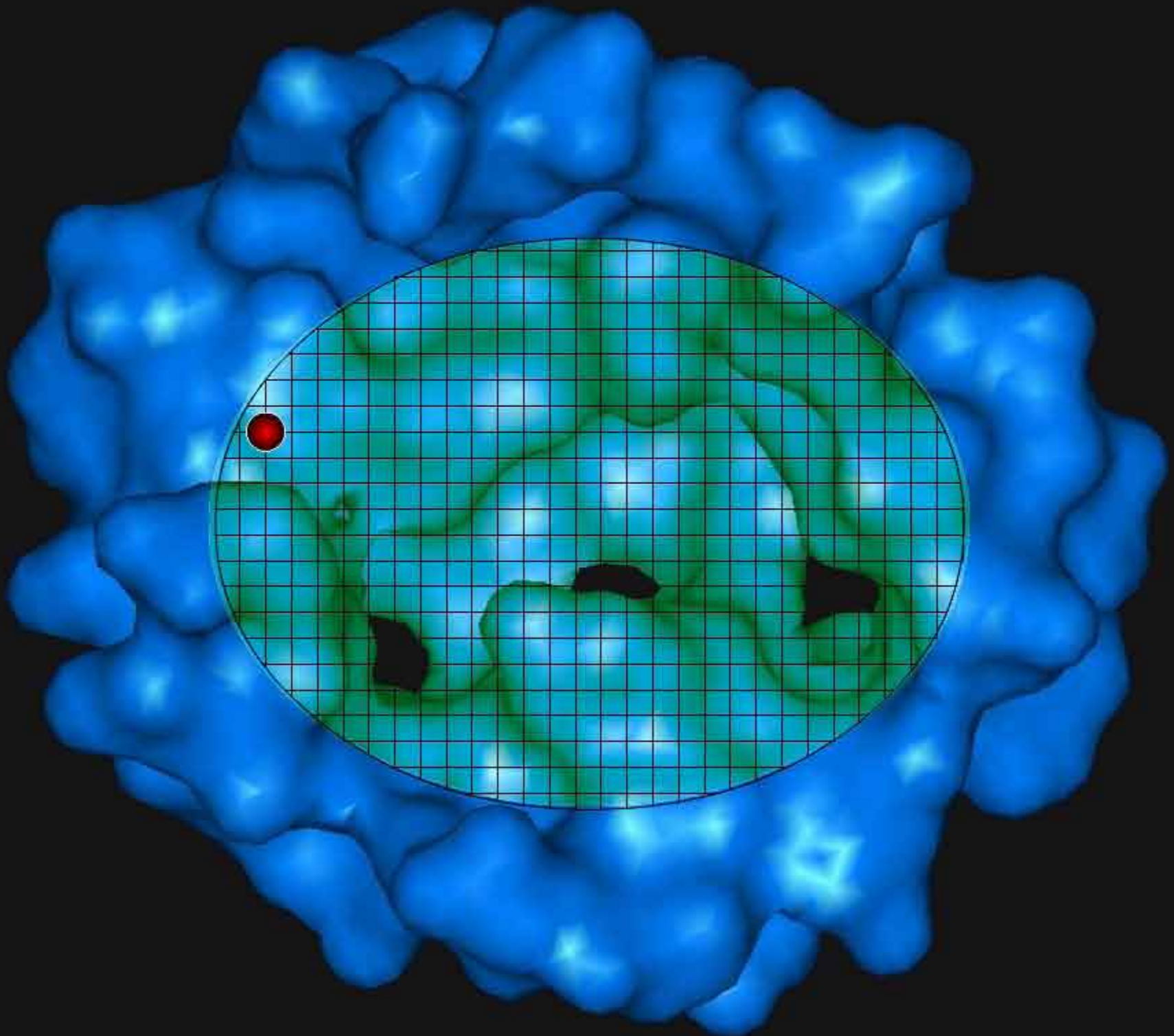
F2.3.2 GRID: a Structure-Based Approach

The active site is systematically explored by calculating the interaction energy between the protein and a chemical probe, at each grid point. A typical GRID interaction energy function between the protein and a probe is shown below.



F2.3.3 Probing the Nature of the Active Site

Underlying the GRID approach is the idea that MIFs of different probes contain relevant information on the nature of the active site of a protein and the forces involved upon binding of a ligand.



F2.3.4 The GRID Probes

Binding sites can be explored by using realistically shaped and charged probes. The GRID program contains several dozen probes such as a single atom, water, the methyl group, amine nitrogen, carbonyl oxygen, carboxylate and hydroxyl etc... More elaborate probes include metal cations (Na^+ , K^+ , Ca^{++} , Fe^{++} , Fe^{+++} , Zn^{++} , Mg^{++}), aliphatic or aromatic (cis or trans) amides, aliphatic or aromatic cationic amidines, meta-diamino-benzene probes etc...

N3+	sp3 Amine NH3 cation
N2:	sp3 NH2 with lone pair
N2	Neutral flat NH2 eg amide
N1:	sp3 NH with lone pair
N1	Neutral flat NH eg amide
N1#	sp NH with one hydrogen
N:-	sp2 N with lone pair

F2.3.5 Integration of GRID with Other Programs

Initially the GRID program generated hundred of pages for each probe, with many tables listing the numerical values of interaction energies at each grid point, and the GRID MIFs proved to be of good quality. The advent of novel numerical statistical methods and progress in computer graphics have enabled the GRID output to be better analyzed.

probe 15			
X	Y	Z	field
2	1	0	-2.3
0	1	2	-12.8
3	1	0	6.1
4	1	2	23.9
3	2	1	-1.2
3	2	0	-7.7



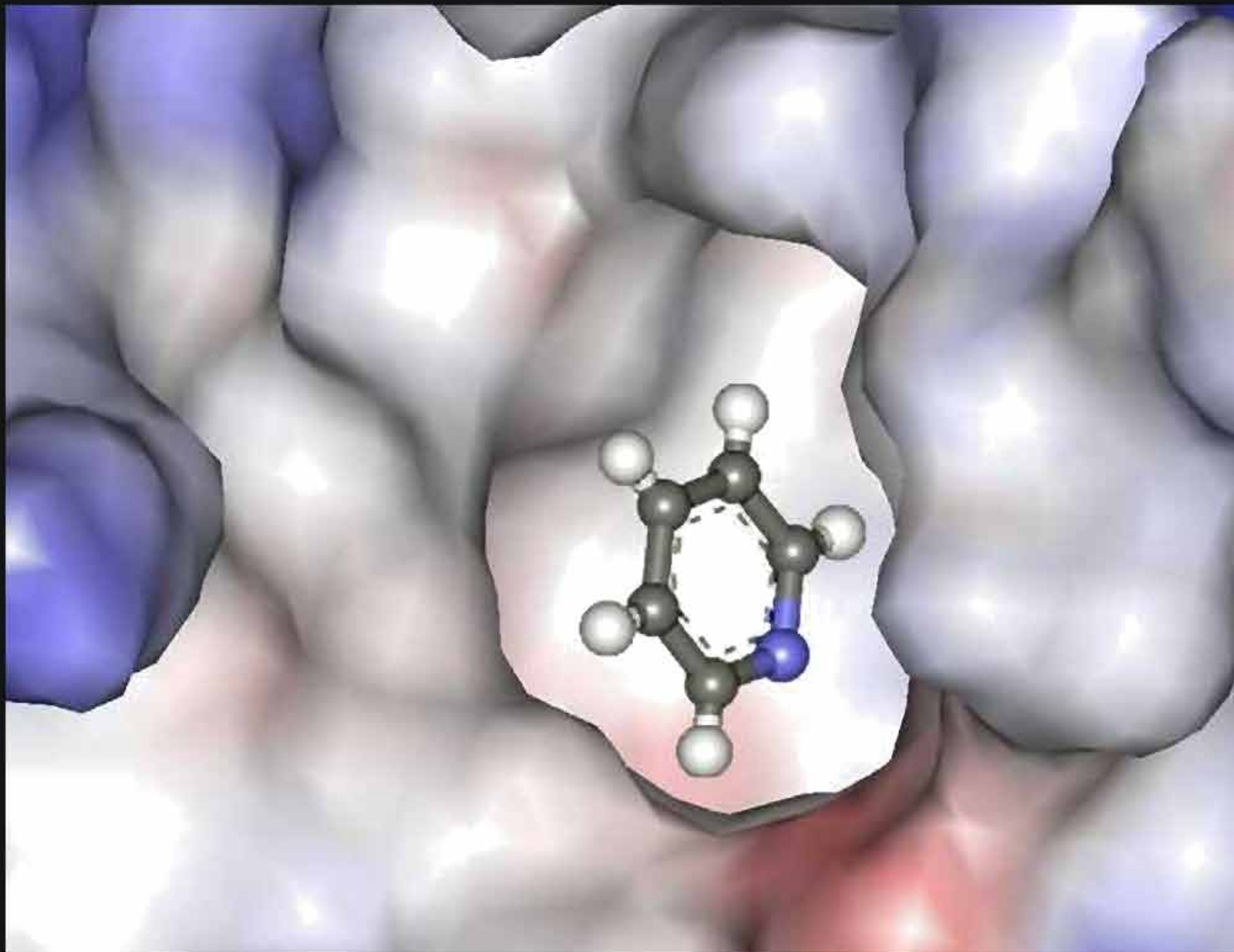
numerical analyses
statistical methods



useful results

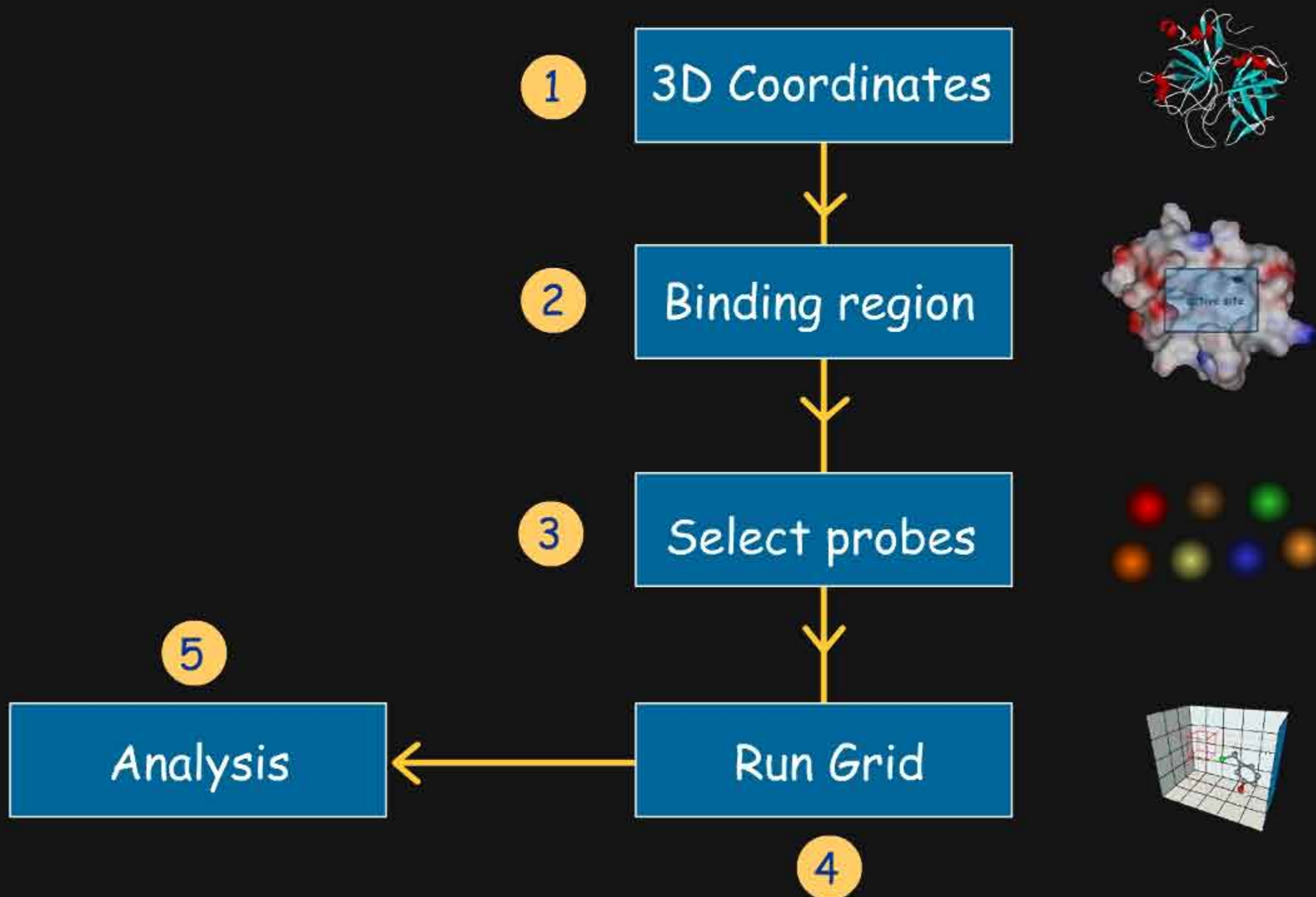
F2.3.6 Typical Use of GRID

GRID predicts favorable interaction positions ("hot spots") with the probes. When fragments are used as probes, the calculations reveal regions in the binding site where this fragment is likely to bind (see pyridine fragment illustrated in the view). This information can be exploited for the de novo design of new molecules.



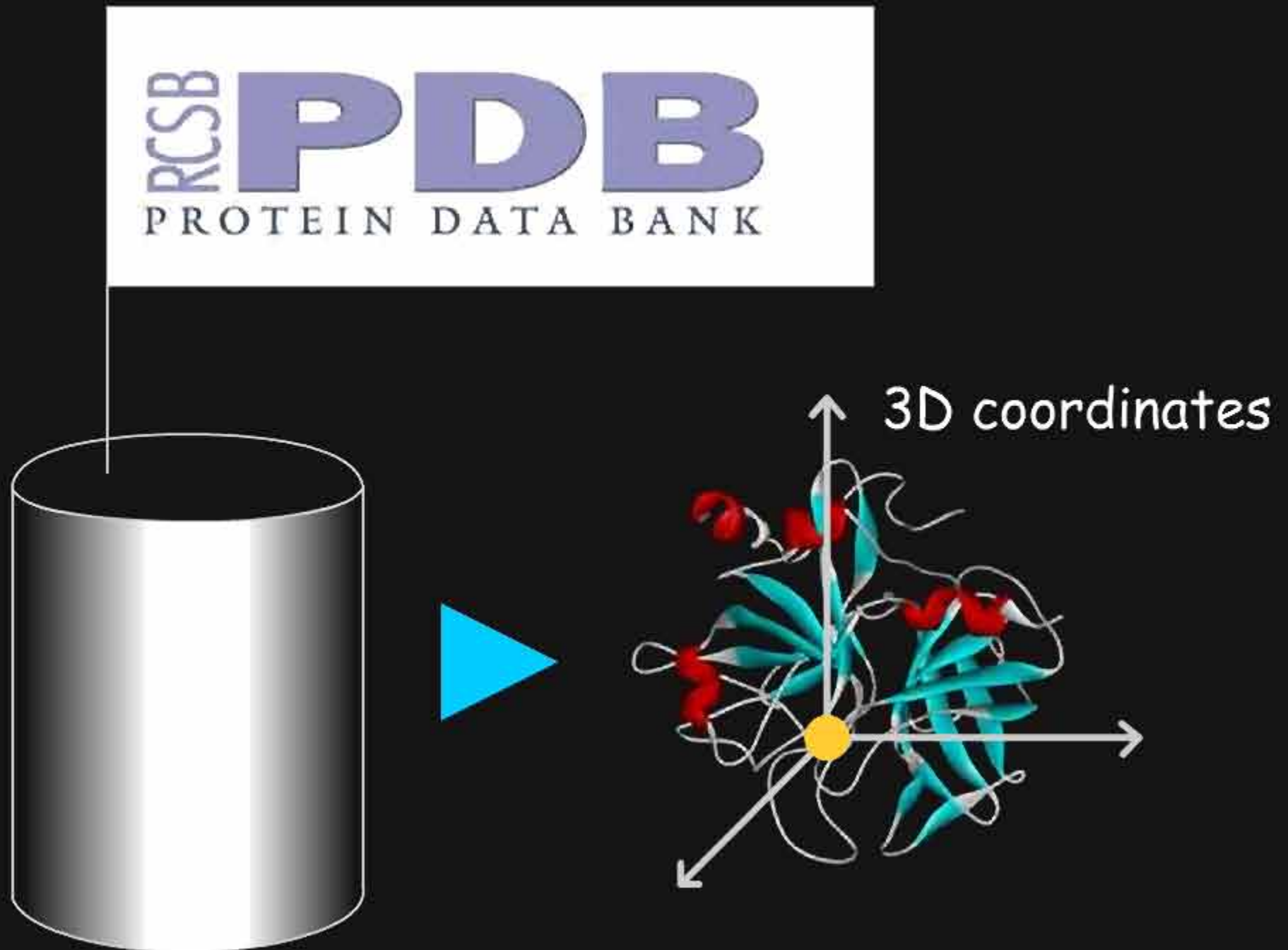
F2.3.7 Outline of a GRID Calculation

The deployment of GRID requires: (1) the 3D coordinates of the atoms of the protein; (2) the binding site to be explored; (3) a selection of several probes; (4) run of GRID leading to the prediction of favorable positions for each probe; (5) analysis of the results with computer graphics.



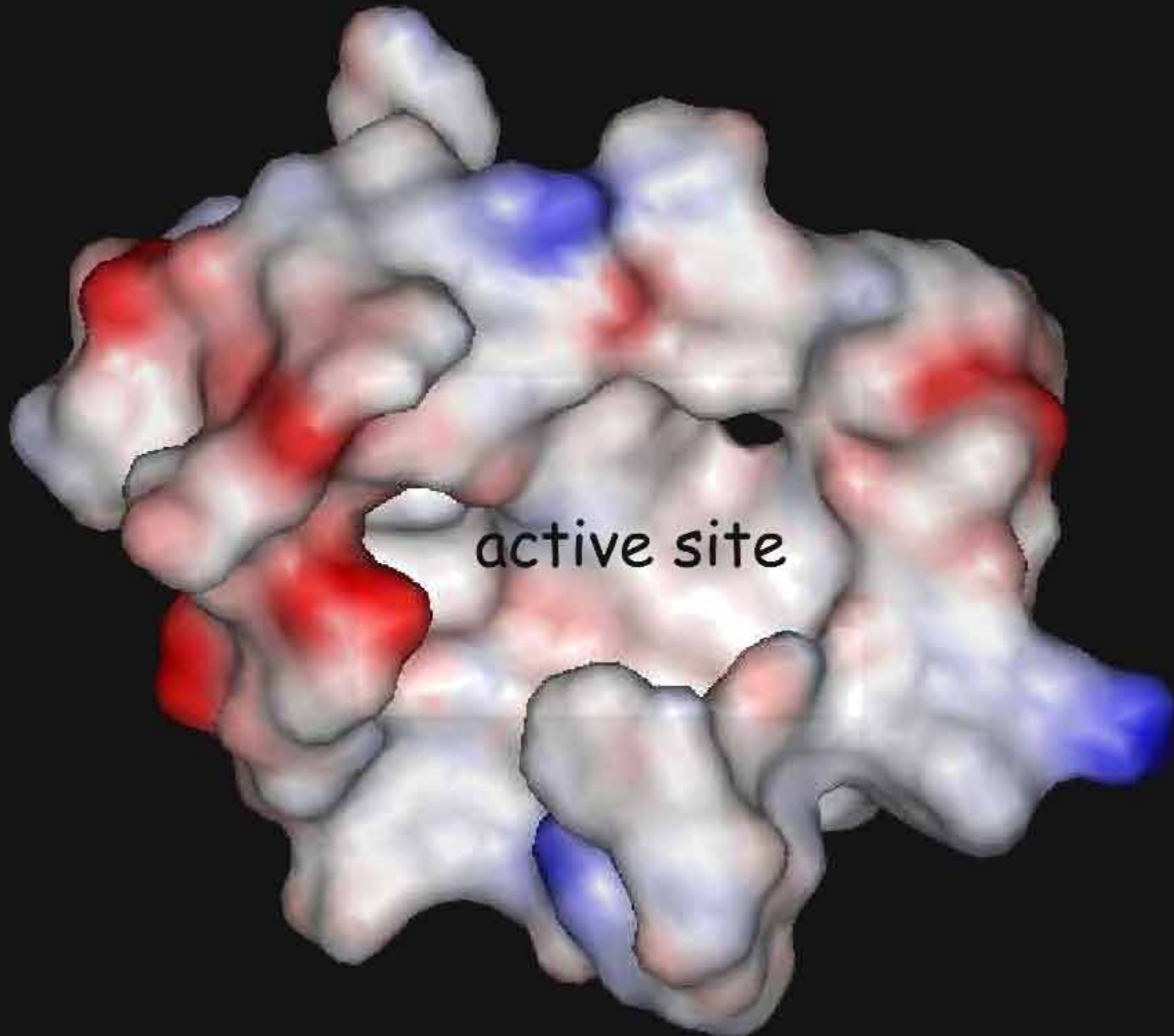
F2.3.8 3D Coordinates of the Protein

The coordinates of the atoms of a protein are obtained from X-ray crystallography or NMR studies (the protein databank, PDB, is the most comprehensive source of high-quality 3D structures). For new proteins of known sequences, 3D models can be derived by homology modeling.



F2.3.9 Binding Site to be Explored

The space surrounding the surface of a protein is huge and it is not necessary to explore the entire space. Normally the exact location of the binding site is known thus the calculation of the interaction fields can be limited to a box containing the active site to be explored.



F2.3.10 Selection of Probes

Several dozen probes are available in GRID. The selection of the probes depends on the nature of the groups that are present in the binding site of the protein. One can address particular types of favorable energy interactions such as hydrophobic, aromatic, polar, salt-bridge interactions etc...

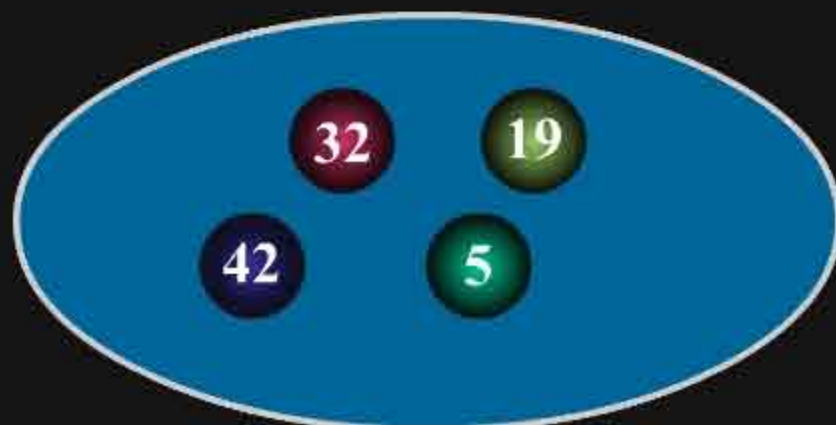
hydrophobic probes



polar probes



aromatic probes

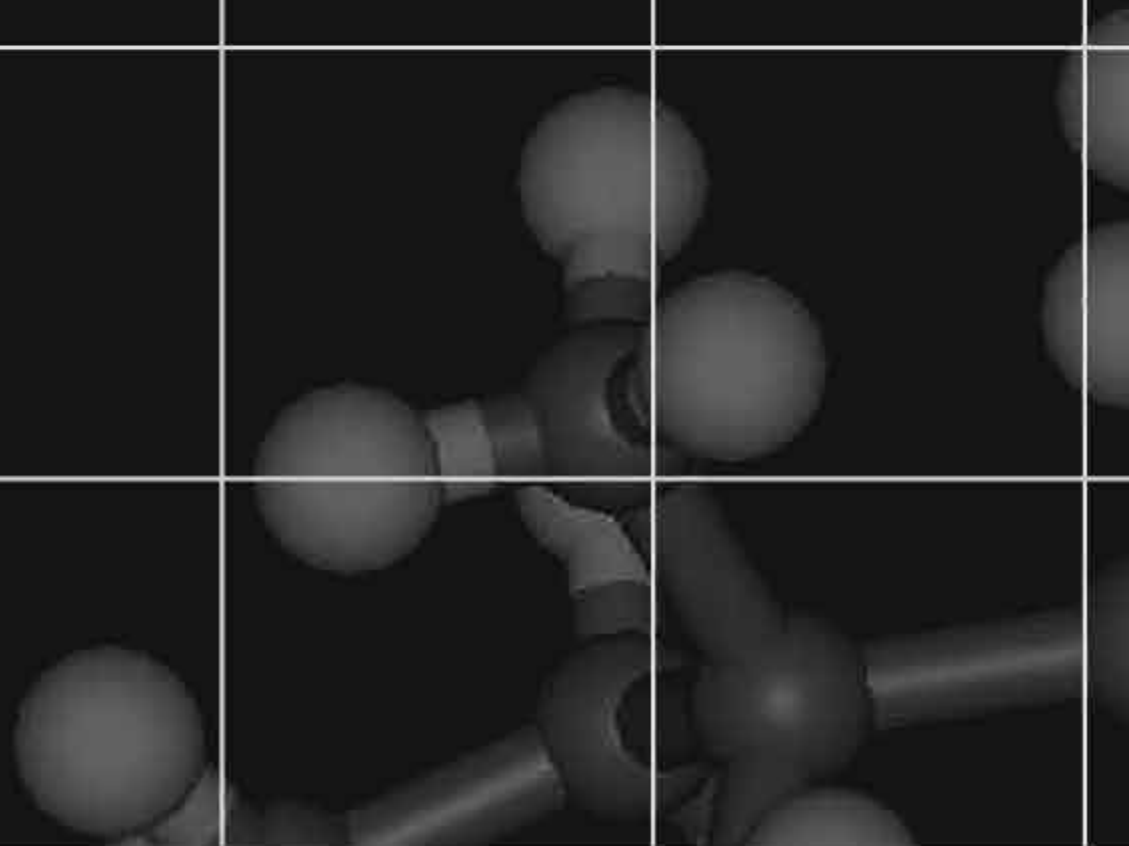
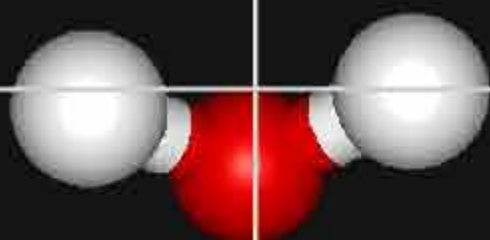


salt-bridge interaction probes



F2.3.11 Run of GRID

For each probe and at each grid point of the lattice the molecular interaction fields are calculated. For multi-atom probes the probe is allowed to rotate at each grid point in order to find the lowest-energy orientation (e.g. essential for hydrogen bonding interactions), an optimization that is very demanding in terms of computing time.



F2.3.12 Output of GRID

The GRID output consists of two types of files (line printer and binary) that need to be used in conjunction with other software. The binary files can serve as input for programs such as CoMFA, GOLPE, SIMCA for the statistical analysis of the GRID maps. Originally, when GRID was introduced visualization tools were not rich enough: advanced visualization programs only started to appear a few years later, enabling the effective visual interpretation of GRID results.

1



<u>probe 1</u>			
X	Y	Z	field
2	1	0	-2.3
0	1	2	-12.8
3	1	0	6.1
4	1	2	23.9
3	2	1	-1.2
3	2	0	-7.7

9



<u>probe 9</u>			
X	Y	Z	field
2	1	0	-1.3
0	1	2	-1.8
3	1	0	-6.1
4	1	2	3.6
3	2	1	-4.0
3	2	0	-4.2

13



<u>probe 13</u>			
X	Y	Z	field
2	1	0	-2.2
0	1	2	-18.8
3	1	0	7.1
4	1	2	3.0
3	2	1	-4.2
3	2	0	-6.5

7



<u>probe 7</u>			
X	Y	Z	field
2	1	0	-4.2
0	1	2	-13.8
3	1	0	4.1
4	1	2	3.9
3	2	1	-0.2
3	2	0	-2.1

F2.3.13 Total Number of Calculations

The total number of GRID calculations is equal to the product of the number of compounds with the number of grid points and the number of probes. This must also be multiplied by the number of rotations, if several orientations at each grid point have been taken into consideration for multi-atom probes.

$$N_{\text{calculations}} = N_{\text{compounds}} \times N_{\text{grid points}} \times N_{\text{probes}} \quad (\times N_{\text{rotation}})$$

Example:

$$N_{\text{compounds}} = 35 \text{ molecules}$$

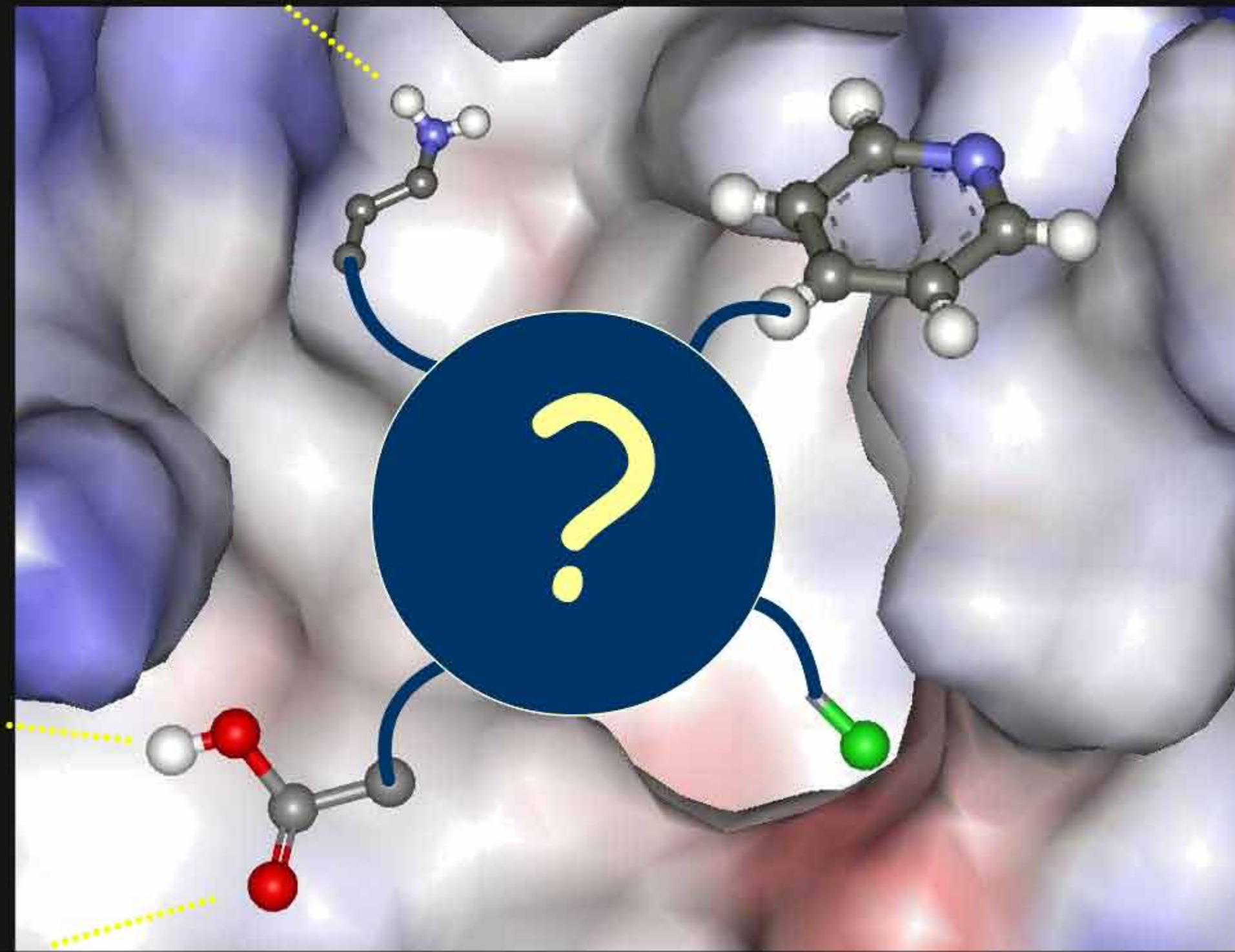
$$N_{\text{probes}} = 8 \text{ probes}$$

$$N_{\text{grid points}} = 21 \times 15 \times 18 \text{ grid points}$$

$$N_{\text{calculations}} = 35 \times 21 \times 15 \times 18 \times 8 = 1,587,600$$

F2.3.14 De Novo Design of New Scaffolds

GRID predicts favorable interaction positions with the probes and reveals where in the binding site a fragment of a given type will prefer to bind. Connecting a maximum of fragments in the correct orientation into a synthetically accessible molecule is not simple. Although GRID provides useful visual clues for creative structure design (de novo design), more advanced computerized approaches enable the systematic exploration of possible solutions (see design methods in chapters D2 and E2).





The topic CoMFA: First 3D-QSAR Method contains the following 41 pages:

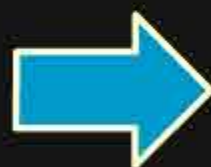
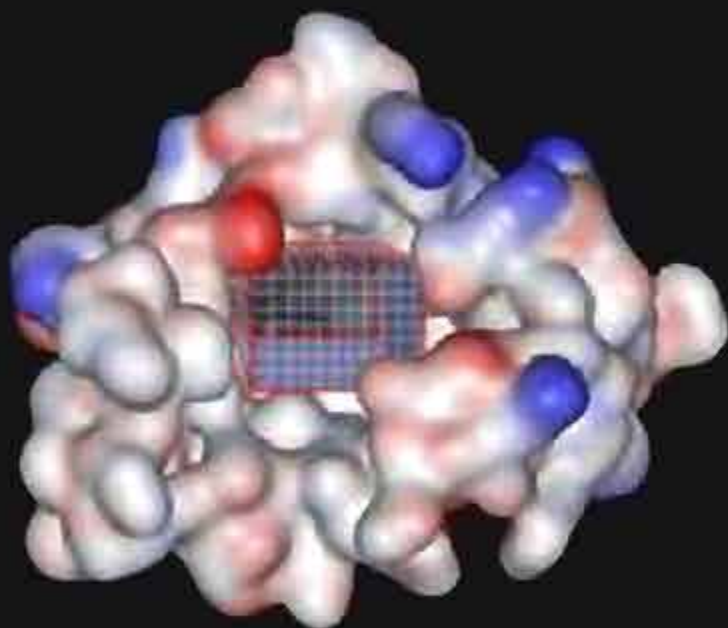
- From GRID to 3D-QSAR
- Comparative Molecular Field Analyses (CoMFA)
- Development of a Correlation Function
- Rapid Outline of a CoMFA Calculation
 - Reference Compounds and Initial Assumptions
 - Superimpose the Structures
 - Calculate the MIF at Grid Each Points
 - Derive a Correlation Function
- Molecular Alignment Issues
 - Template or Atom Alignments
 - Pharmacophore Alignments
 - Shape Alignments
 - Field Fitting
- ...

For the entire list, see the navigation panel.

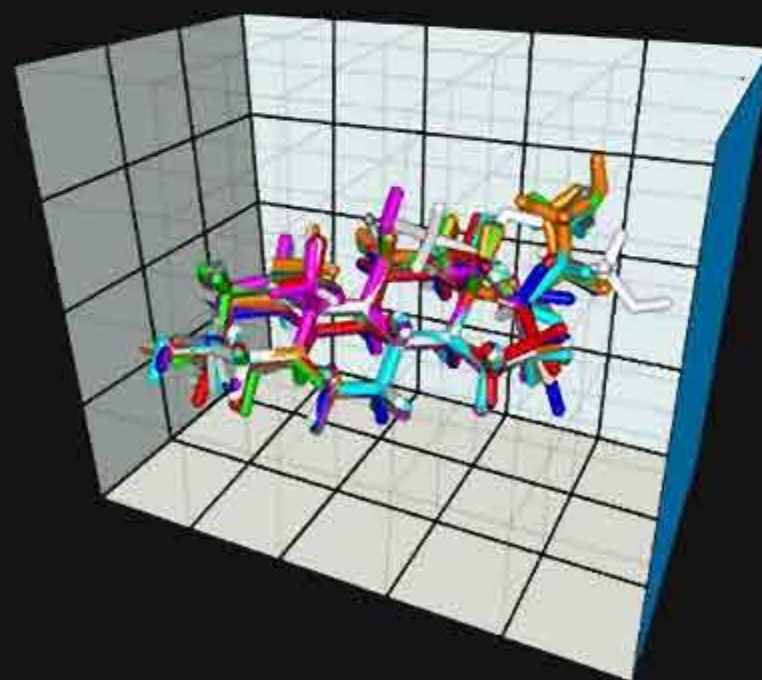
F2.4.1 From GRID to 3D-QSAR

GRID is a MIF-based method developed for the analysis of macromolecule active sites, to reveal "hot spots" in their binding regions. It paved the way for the ideas that led to the development of the 3D-QSAR approaches, which retained the MIF concept and fully exploited it for the study of small molecule ligands, in projects where the 3D structure of the target protein is not known.

GRID Method

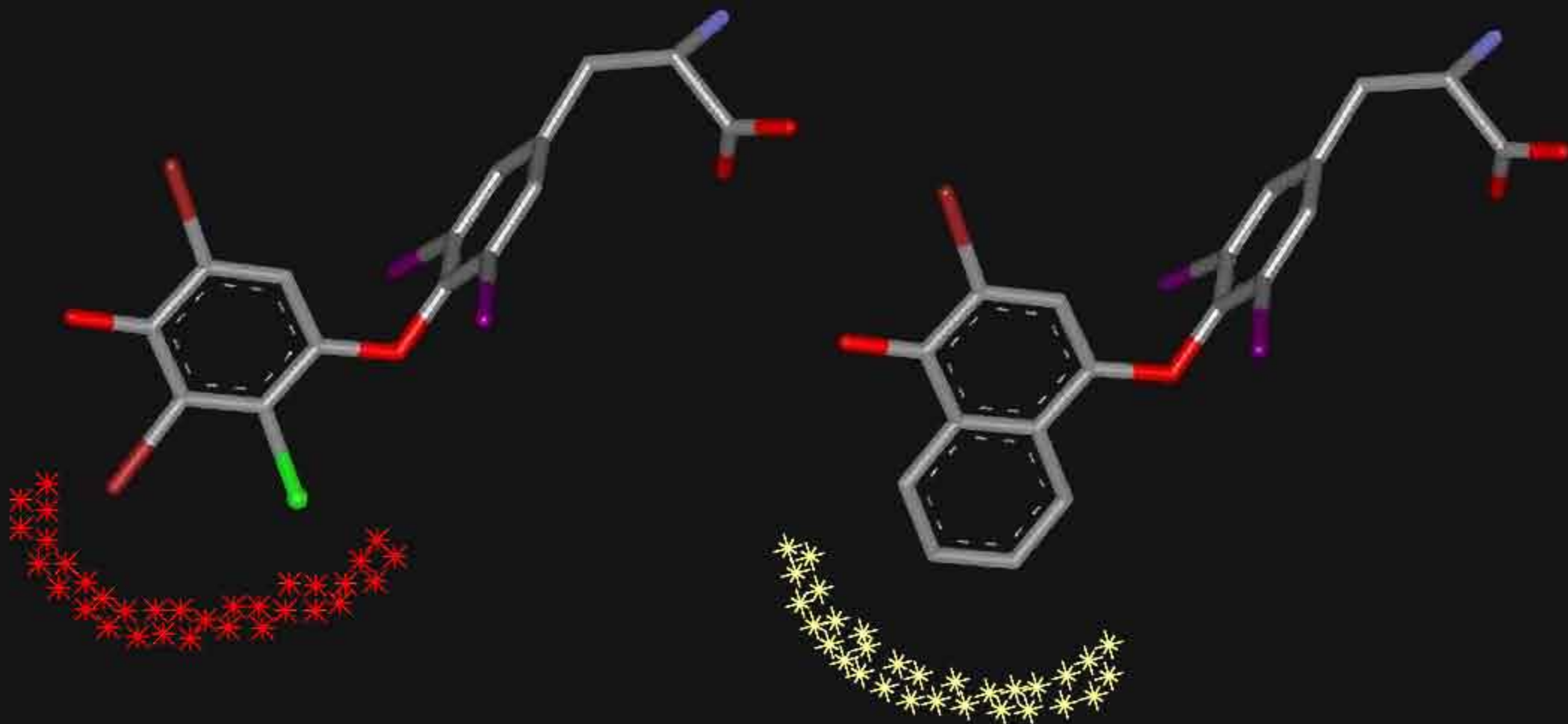


3D-QSAR Method



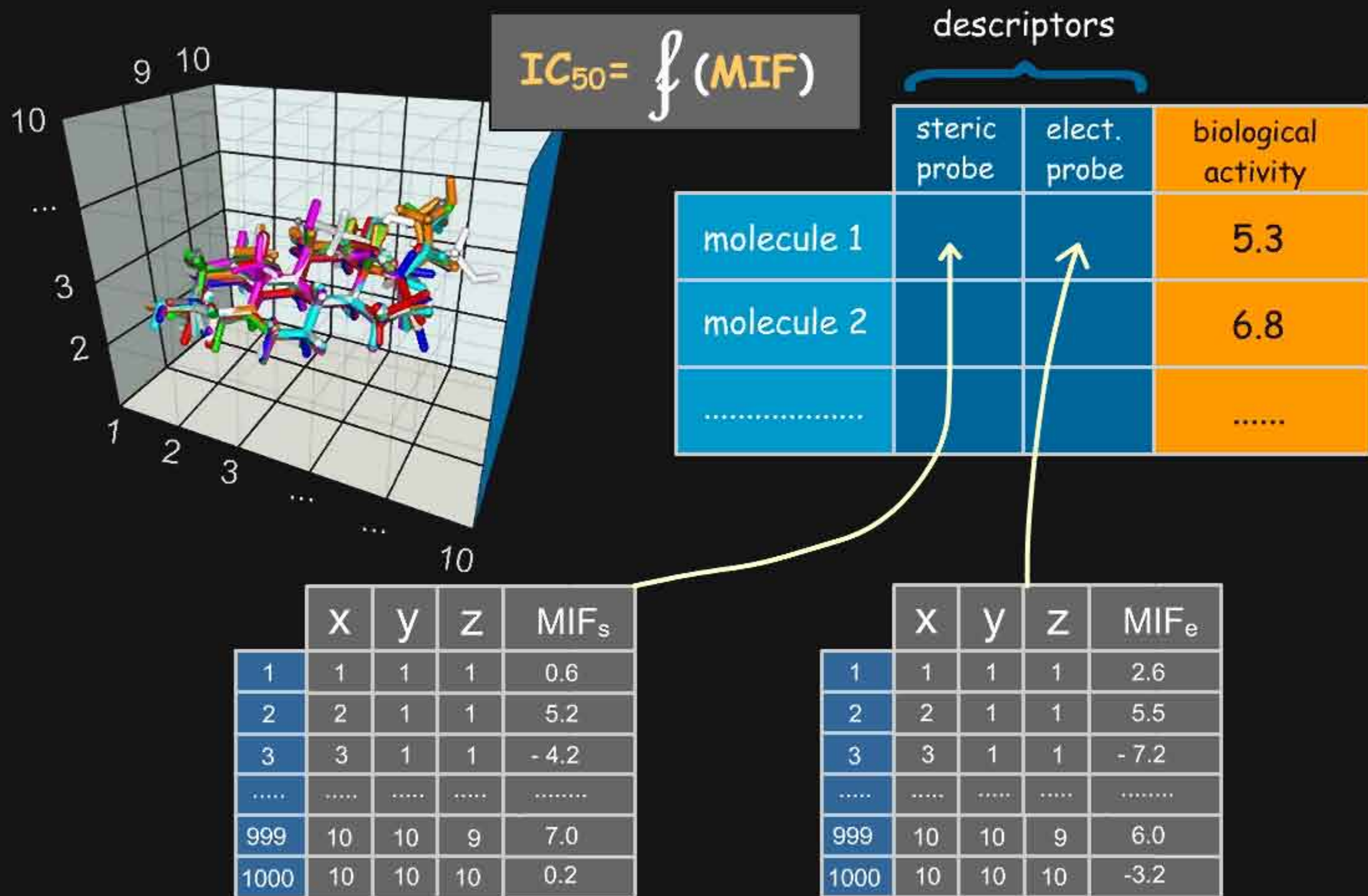
F2.4.2 Comparative Molecular Field Analyses (CoMFA)

CoMFA is a method introduced by Cramer in 1988, for the COmparison of Molecular Field Analyses of different molecules. The method is based on the assumption that the 3D distribution of the interaction fields of a compound contains relevant information for understanding its biological activities. Comparison of the fields is expected to reveal important features concerning the activities of the molecules and can facilitate their optimization.



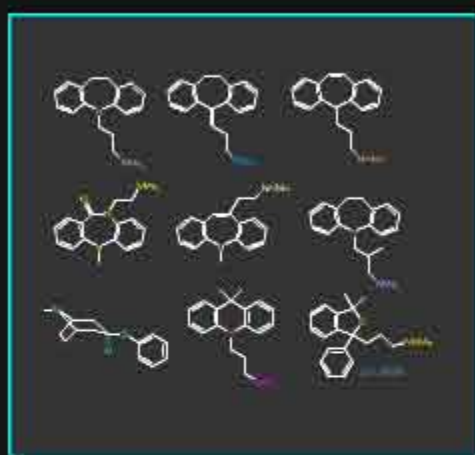
F2.4.3 Development of a Correlation Function

Beyond the visual analysis of the MIFs of the active and inactive molecules, CoMFA aims at formulating a linear equation, correlating the biological activities with the values of the fields in each point. Here, each value calculated for a field on a given point (x_i, y_i, z_i) is a descriptor.

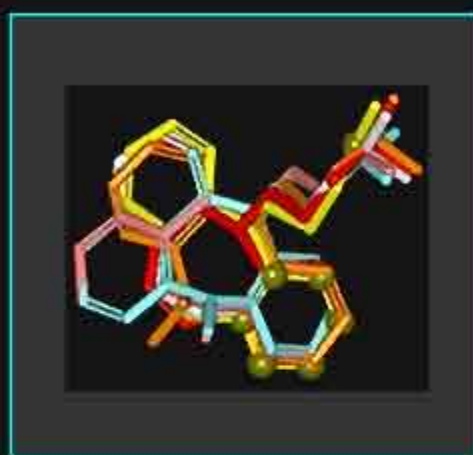


F2.4.4 Rapid Outline of a CoMFA Calculation

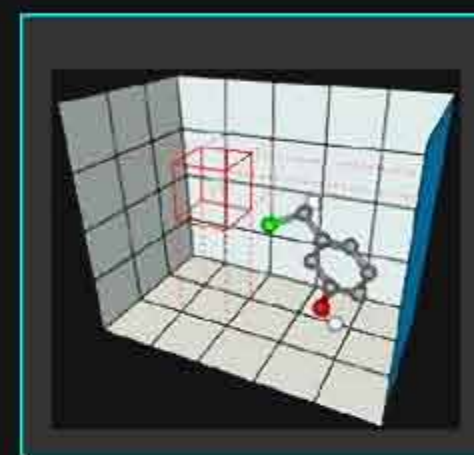
The development of CoMFA requires: (1) a set of related analogs; (2) defining a rule for superimposing them; (3) constructing a lattice of grid points and computing for each molecule the interaction with the probe at each point; (4) deriving a correlation function; (5) assessing the predictability of the model and (6) exploiting the results.



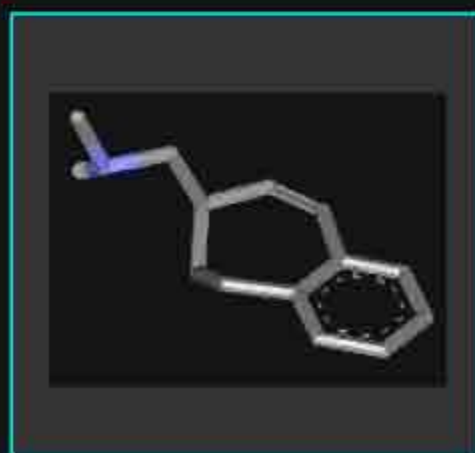
Reference compounds



Superimpose molecules



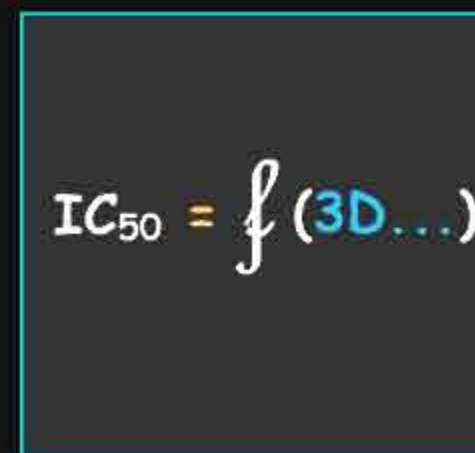
Construct lattice
Calculate interactions fields



Exploit results



Quality of model



3D-QSAR
Correlation function

F2.4.5 Reference Compounds and Initial Assumptions

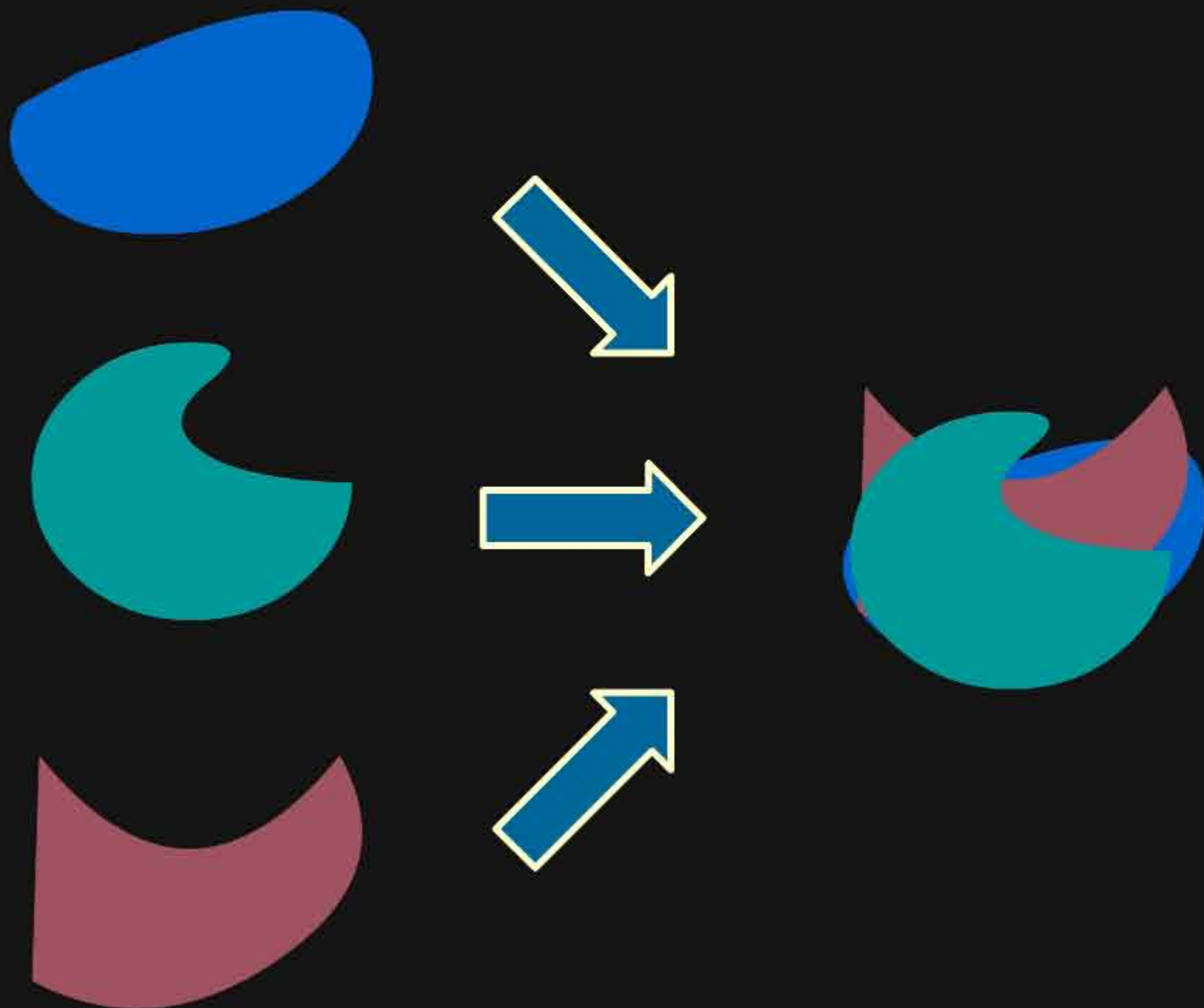
A CoMFA study starts by selecting of a set of active/inactive molecules with their associated biological properties. Implicitly it is assumed that they share the same mechanism of action and that they are active for the same reason. All molecules are assumed to bind in the same way; it is also assumed that the biological action is enthalpically driven and that the entropic terms and desolvation energies are similar for all the compounds.

CoMFA Assumptions

- Same mechanism of action
- Active for the same reason
- Bind in the same way
- Biological process enthalpically driven
- Entropic terms similar for all compounds
- Desolvation energies similar for all compounds

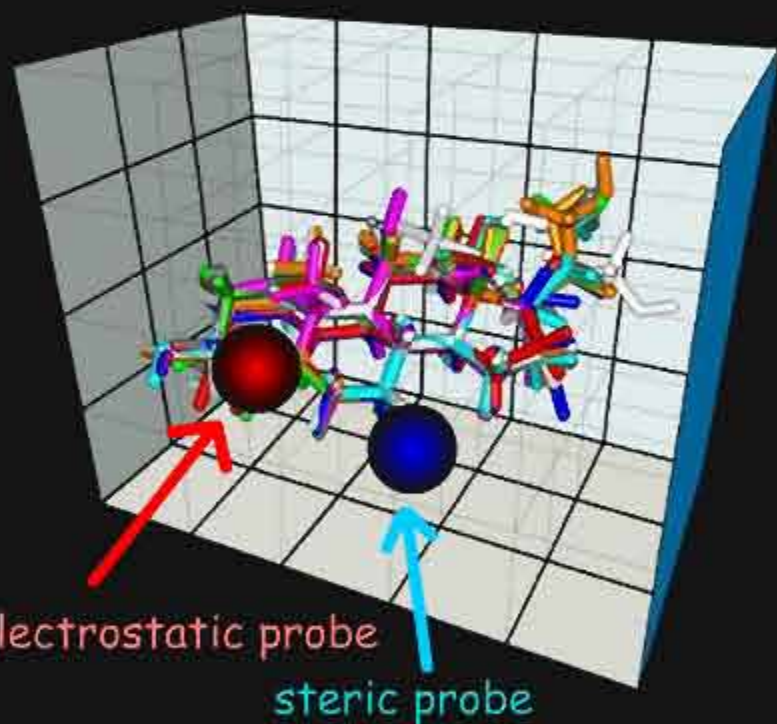
F2.4.6 Superimpose the Structures

All molecules in the reference data set should be aligned to one another prior to the MIF calculation. Most typically used methods are based on the superimposition of their common chemical scaffolds or their common pharmacophores. Molecular alignment methods will be presented in more detail in the course of this chapter.



F2.4.7 Calculate the MIF at Grid Each Points

First, a common lattice is constructed for the molecules superimposed. Then, for each separate molecule the molecular interaction fields are calculated for each probe and at each grid point. In CoMFA, only two probes are used: one for measuring the steric field and one for measuring the electrostatic field.



Molecule 1

steric probe				
X	electrostatic probe			
	X	Y	Z	field
2	2	1	0	-2.3
0	0	1	2	-12.8
3	3	1	0	6.1
0	4	1	2	23.9
4	3	2	1	-1.2
	3	2	0	-7.7

Molecule 2

steric probe				
X	electrostatic probe			
	X	Y	Z	field
2	2	1	0	-4.2
0	0	1	2	-13.8
3	3	1	0	4.1
0	4	1	2	3.9
4	3	2	1	-0.2
	3	2	0	-2.1

F2.4.8 Derive a Correlation Function

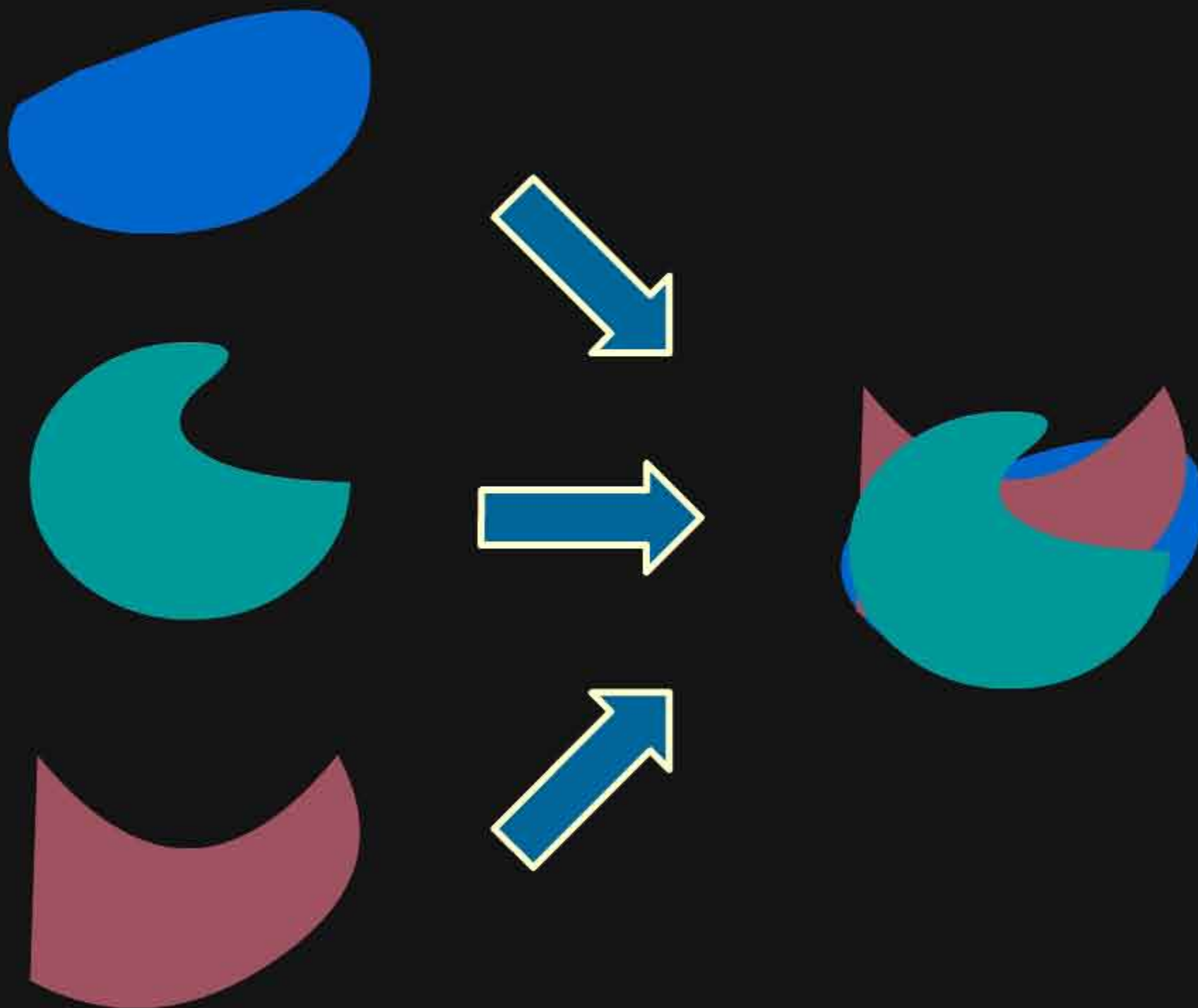
The numerical values generated by the calculations are then processed by sophisticated mathematical statistical methods, with the aim of revealing a linear relationship between the field parameters and the biological activities. Normally the partial least-squares method (PLS) proves to be the method of choice in 3D-QSAR (statistical methods normally used in QSAR cannot handle the huge amount of data generated by CoMFA calculations).



$$\text{Biological effect} = f(3\text{D Molecular Fields})$$

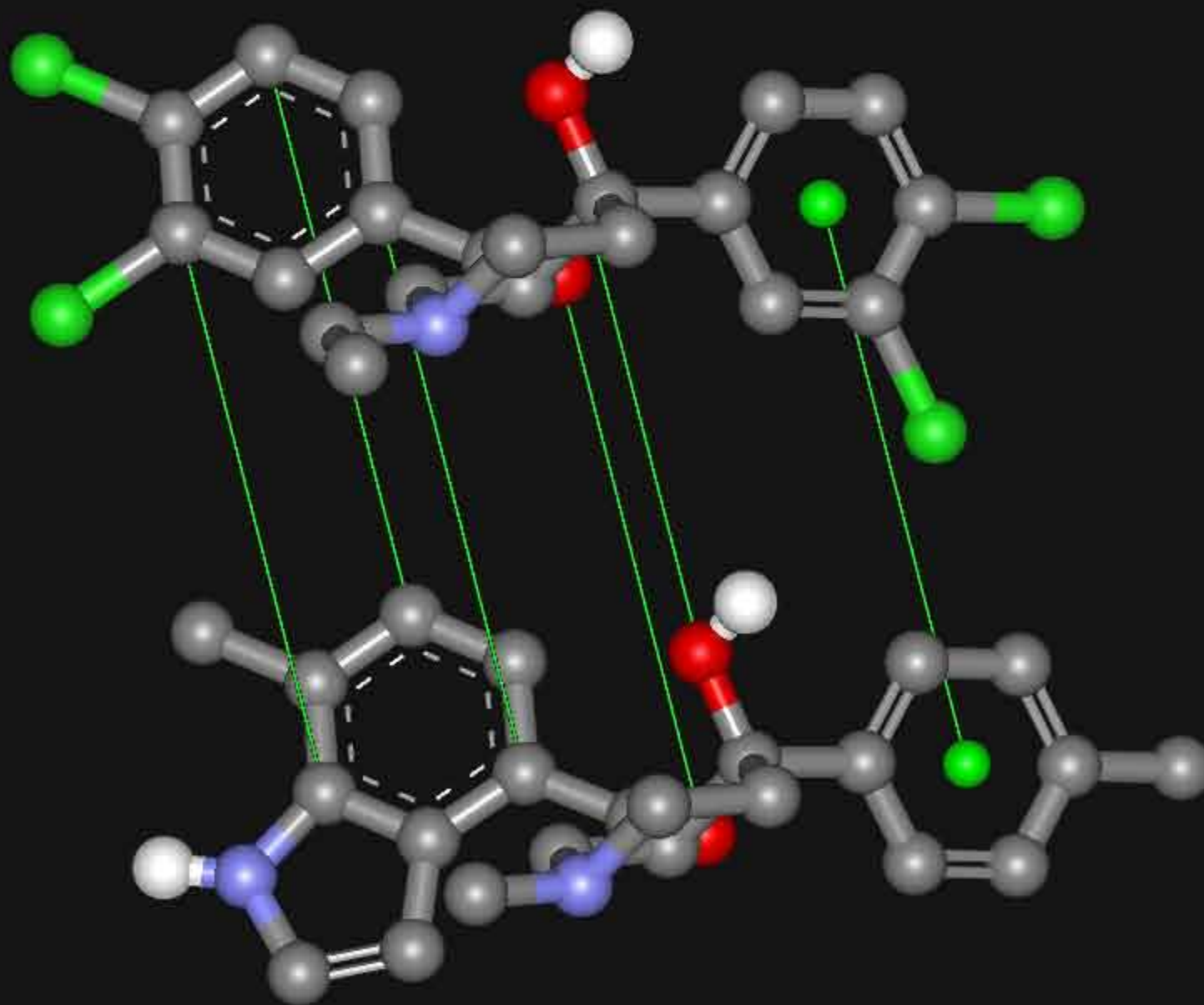
F2.4.9 Molecular Alignment Issues

Unfortunately CoMFA models are highly dependent on the way the molecules are aligned. The assumptions used in deriving alignments are therefore a difficult problem. Even for a series of related analogs, their exact orientation in the active site might be different (depending on the particular forces exerted by the protein on each ligand). The most common alignment methods are discussed in the following pages.



F2.4.10 Template or Atom Alignments

A simple method consists of superimposing a set of atoms common to all the compounds. A molecule is chosen as a reference fixed template and the other molecules are moved to their new positions, corresponding to the minimum of the sum of the squared distances between a chosen set of atom pairs.

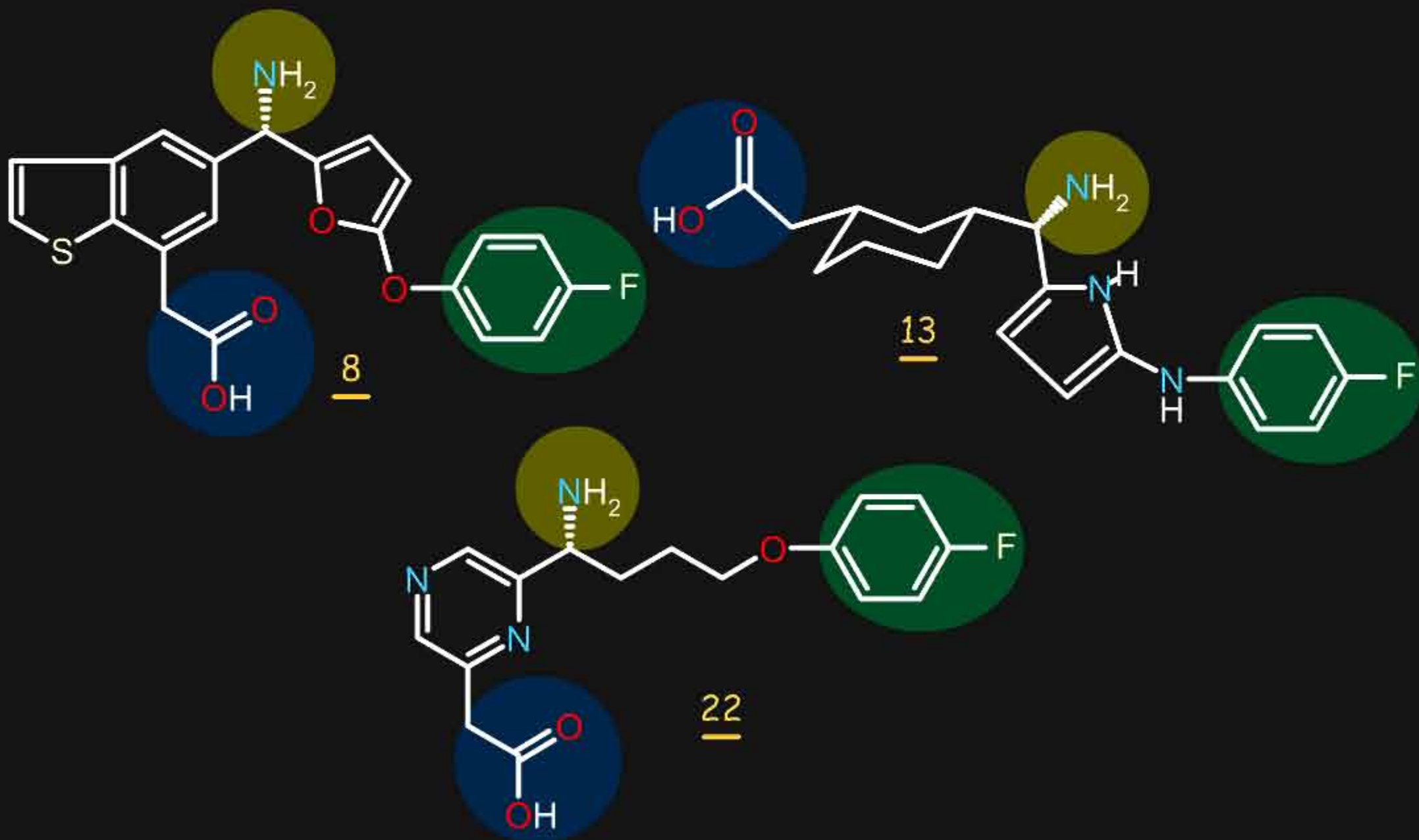


F2.4.11 Pharmacophore Alignments

When there is no common scaffold shared by all the molecules it is possible to select pairs of atoms based on a common pharmacophore (or on chemical similarity assumptions). The molecules below illustrate a situation where the alignment of the molecules is not straightforward however this can be done by superimposing their common pharmacophore consisting of a fluoro-phenyl, an amino group and a carboxyl moiety.

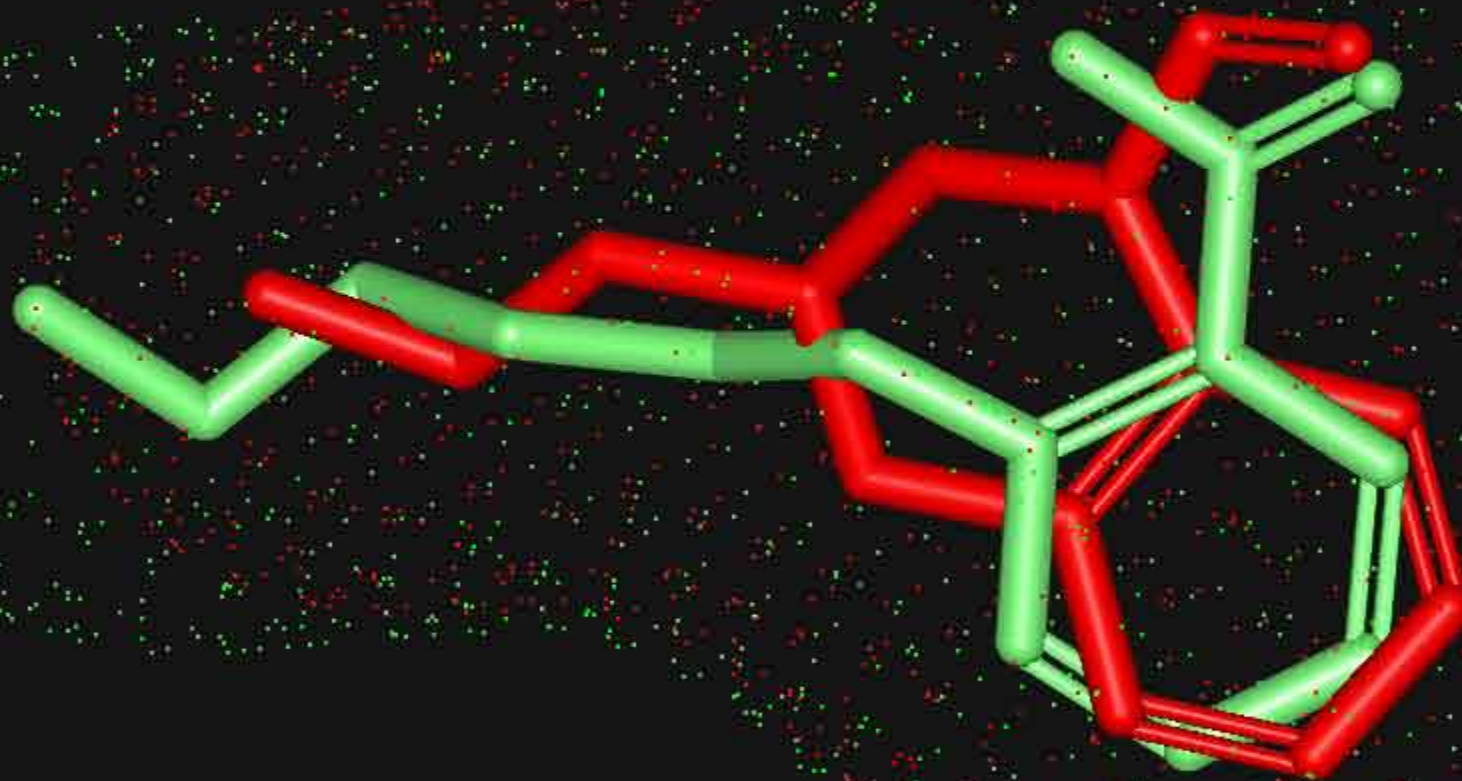
2D

Superimposition



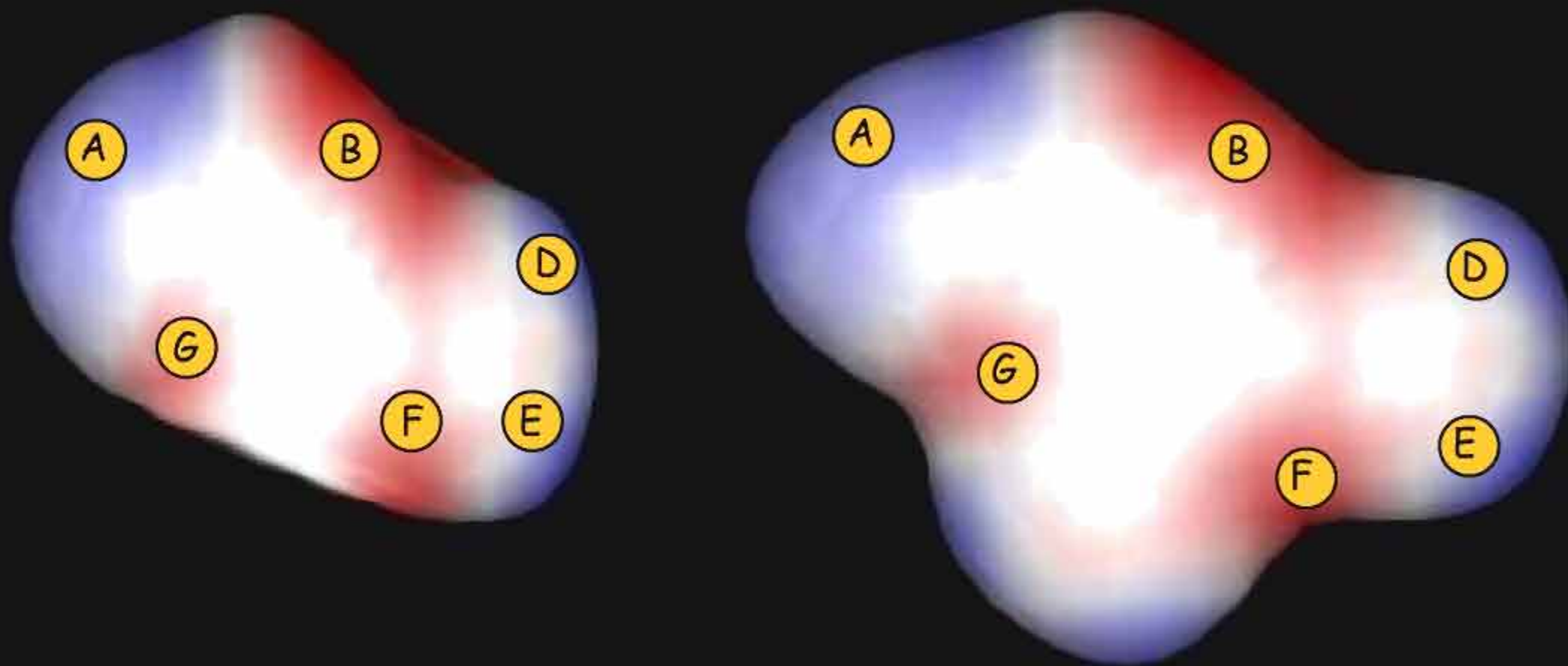
F2.4.12 Shape Alignments

In the absence of obvious rules, molecular modeling always enables the superimposition of molecules, based on shape alignments.



F2.4.13 Field Fitting

After the steric and electrostatic fields around the molecules are calculated, it is possible to align the molecules by maximizing the overlap of the two fields. This method does not require the definition of any matching atoms.

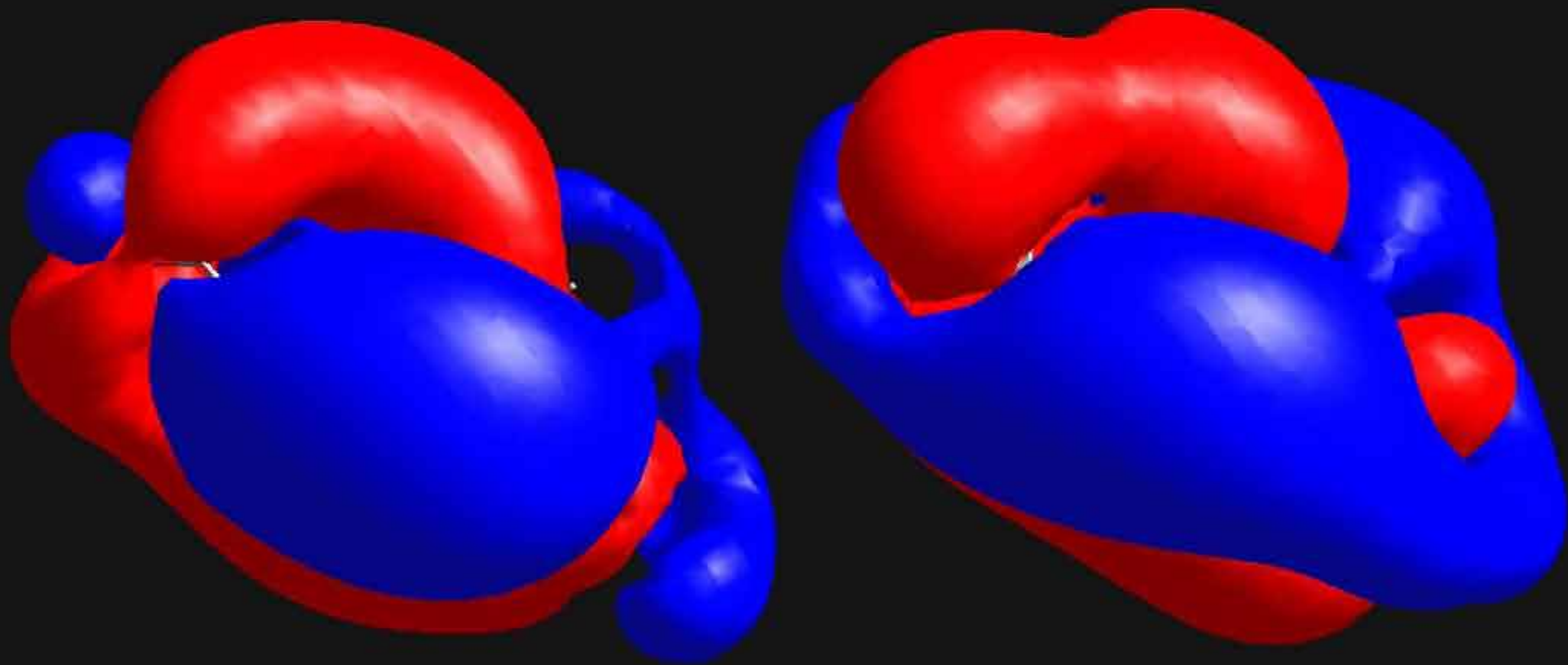


F2.4.14 Electrostatic Field Alignment

This example from the EON program illustrates two chemically unrelated molecules (from MDDR) that could be superimposed by aligning the electrostatic fields created by the molecules in their vicinity (blue lobes are positive and red lobes are negative).

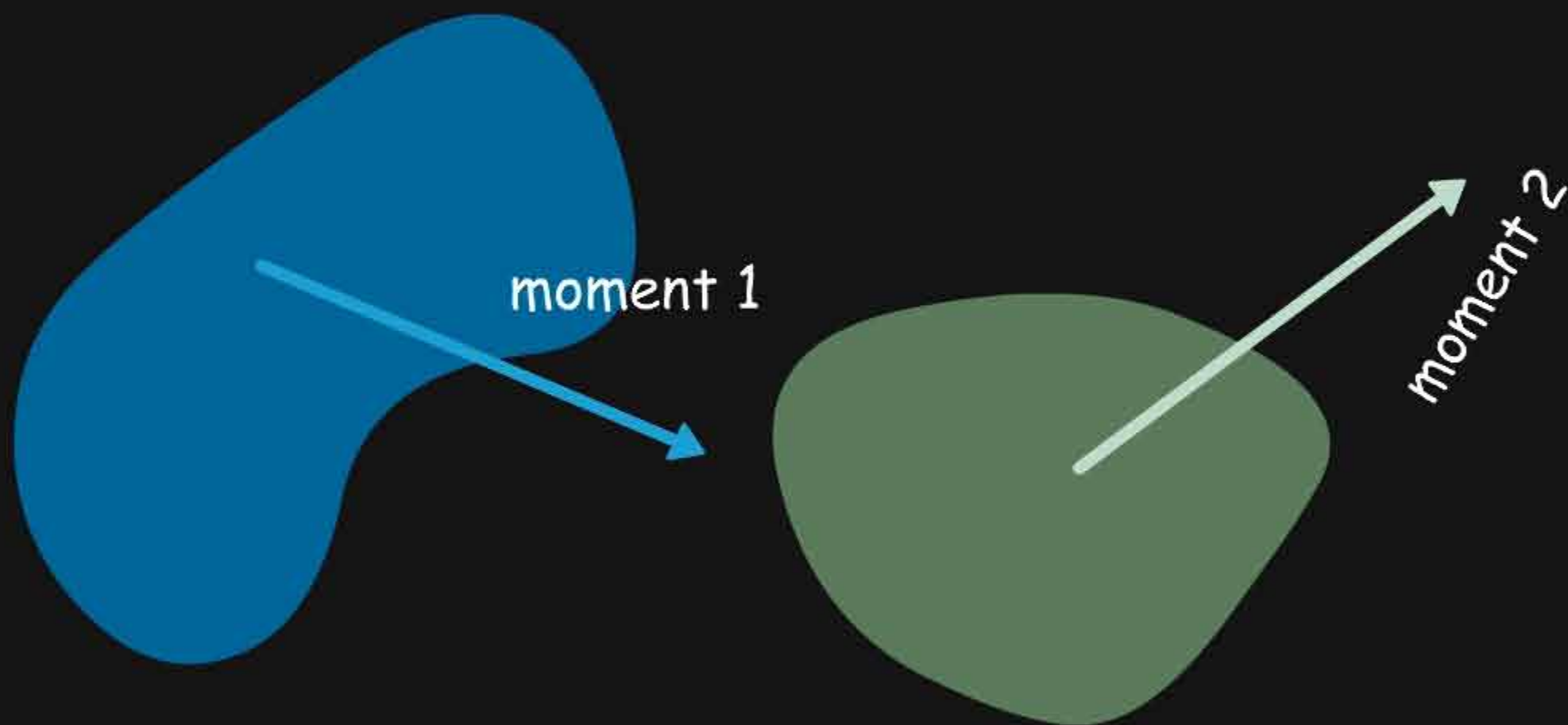
Field Alignment

2D



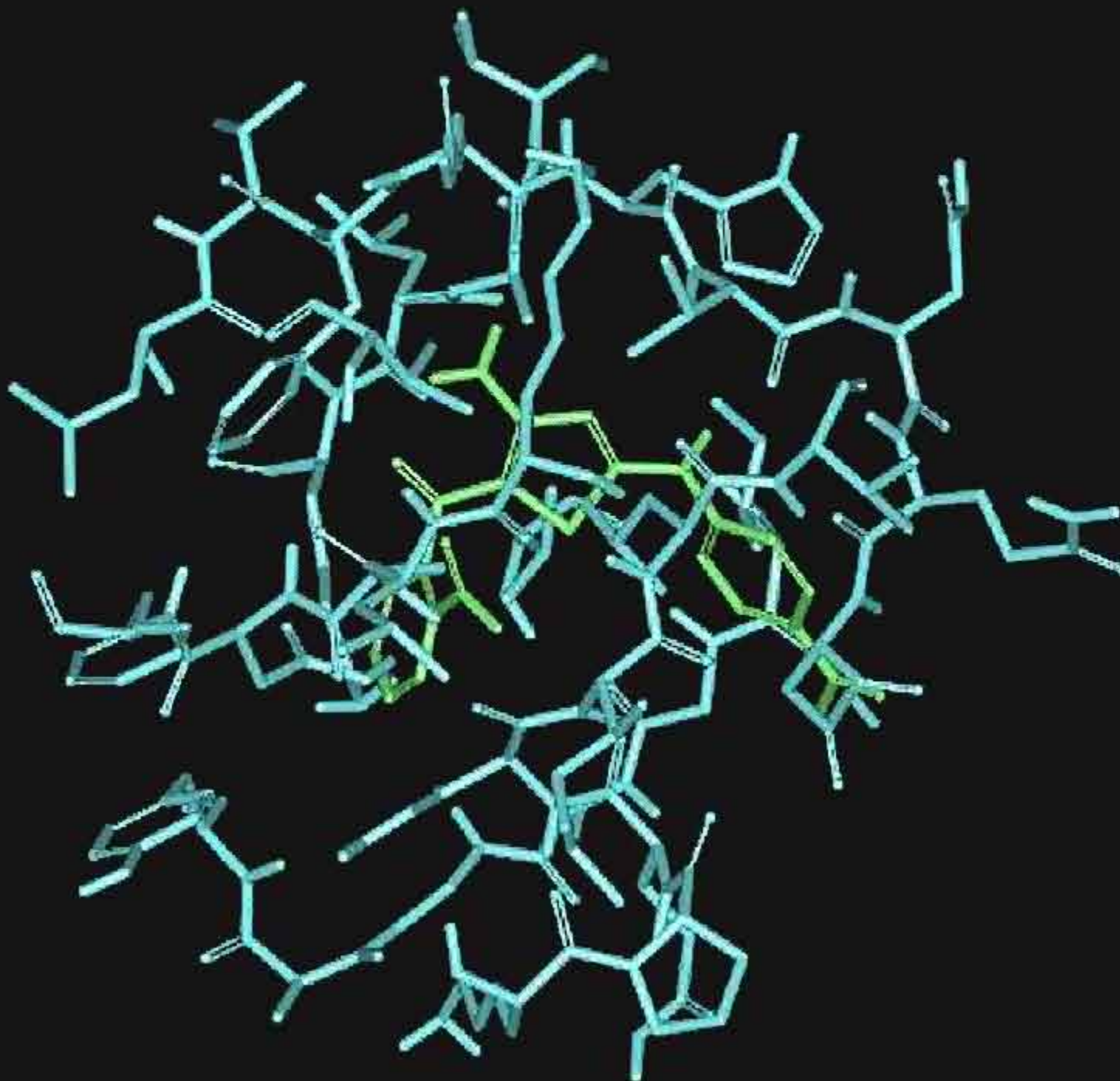
F2.4.15 Moment Alignments

This is a less accurate alignment method; it consists of aligning the molecules based on molecular moments such as the molecular dipole, the principal moments of inertia or a field similarity moment.



F2.4.16 Receptor Based Alignments

The exact orientation of the molecules (from X-ray data or docking calculations with an homology model of the protein) would be ideal to have. Indeed, the position of each molecule in the active site depends on so many interactions that it would be excellent to conduct 3D-QSAR analyses where the orientation of the molecules is close to reality. This method is only possible when the 3D structure of the receptor is known; however this is not the case in many CoMFA projects.

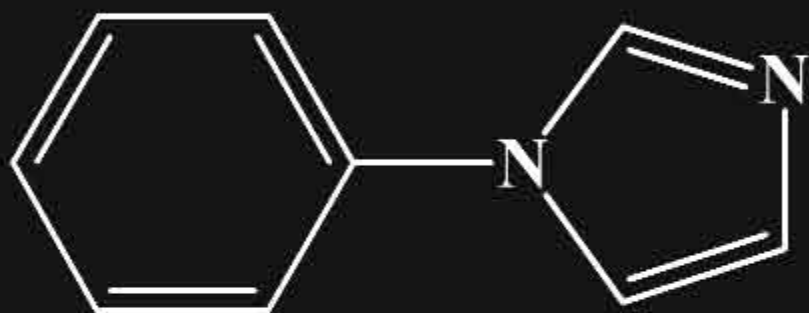


F2.4.17 Alignment from X-ray Data

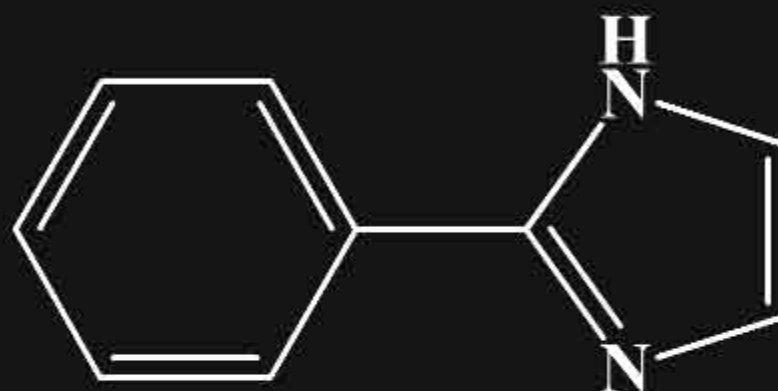
The following analogs are inhibitors of CP450-cam, and it is reasonable to assume that they bind to a sub-pocket having a shape corresponding to their common similar volumes. However X-ray studies reveal a somewhat different alignment of the three molecules (see button "X-rays").

● 2D

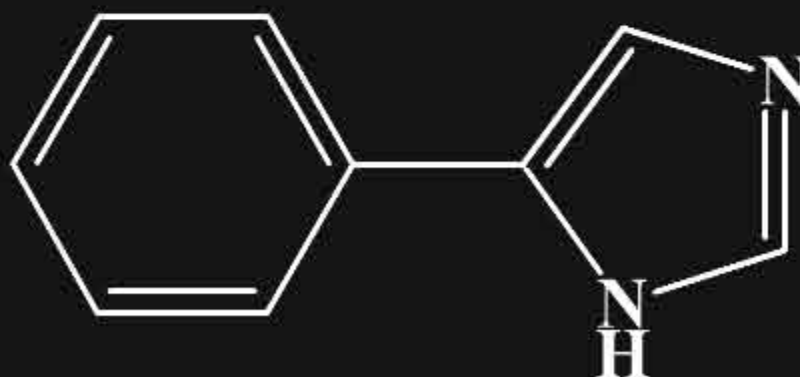
● X-rays



Phenyl-1 imidazole



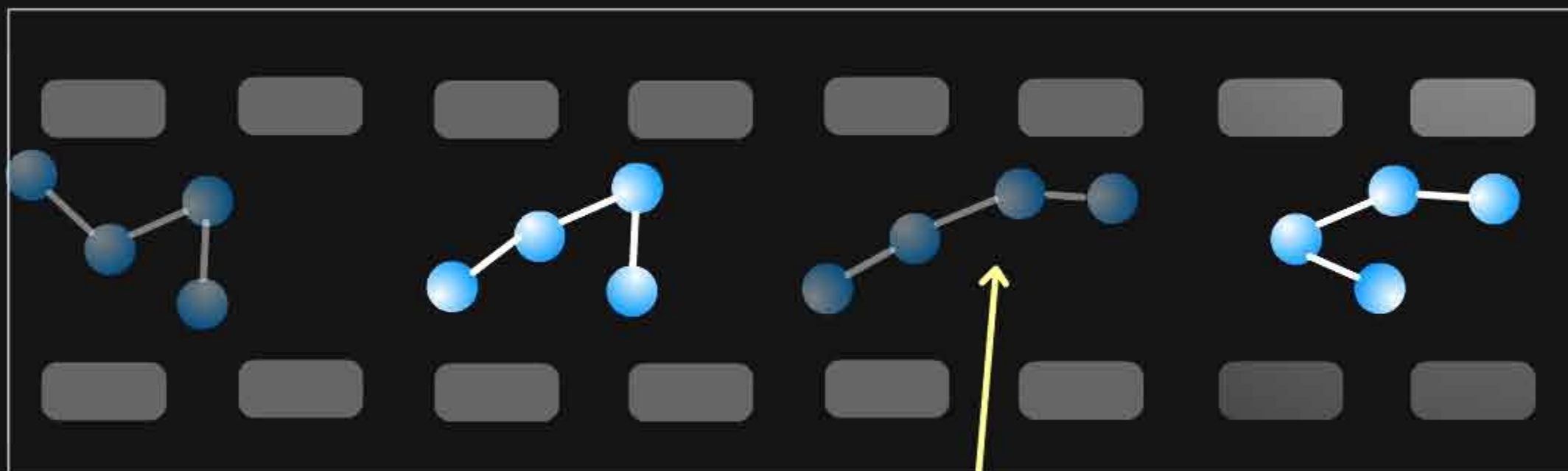
Phenyl-2 imidazole



Phenyl-4 imidazole

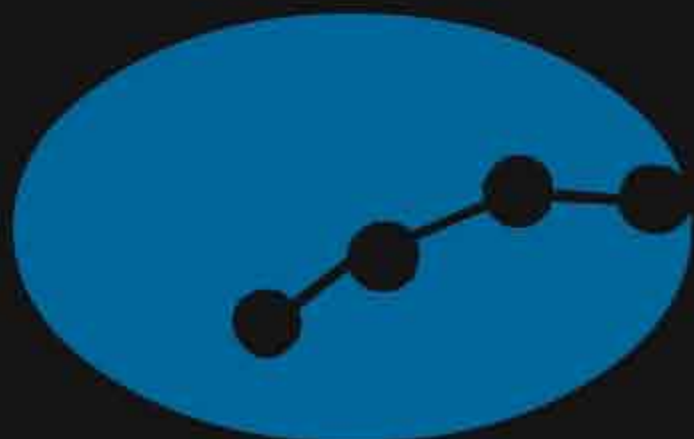
F2.4.18 The Bioactive Conformation Issue

Bear in mind that whatever method is used for superimposing the molecules, they must be aligned in their bioactive conformations. Preliminary analyses are needed to establish these, and in some cases it may be necessary to consider several hypotheses. CoMFA studies are always safer with a set of homolog compounds. The historical CoMFA paper by Cramer was on steroid analogs, a case where the conformational "risk" was minimal.



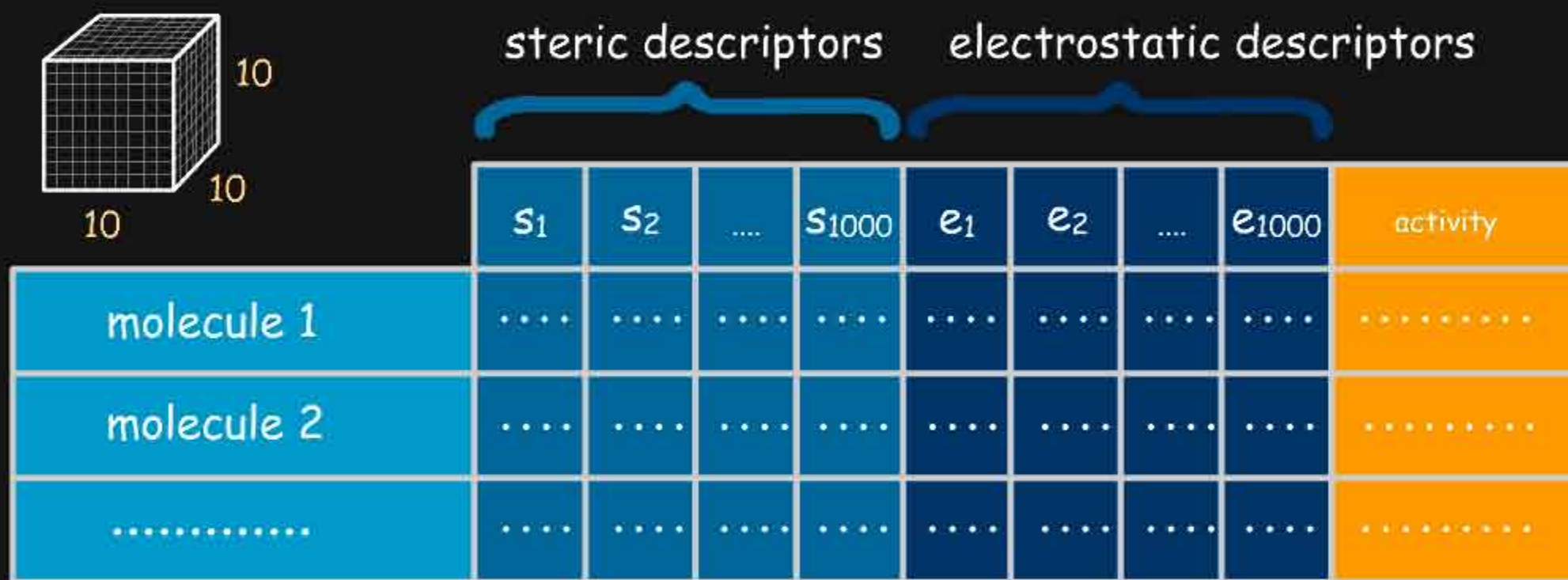
receptor

bioactive conformation



F2.4.19 Deriving the 3D-QSAR Correlation Function

The aim of the 3D-QSAR is to derive a linear function, which predicts the biological activities of the molecules in terms of the individual values calculated for the fields (s_i and e_i are the steric and electrostatic descriptors, respectively).



$$\text{biological activity} = a_1 s_1 + a_2 s_2 + a_3 s_3 + a_4 s_4 + \dots + a_{1000} s_{1000} + b_1 e_1 + b_2 e_2 + b_3 e_3 + b_4 e_4 + \dots + b_{1000} e_{1000} + k$$

F2.4.21 PLS: the Partial Least-Squares Method

The statistical method known as Partial Least Squares (PLS) has proven to be suited for handling this complex multivariate problem by eliminating the correlation between the descriptors, reducing their number and enabling the generation of a linear relationship between the field parameters and the biological activities.

many correlated descriptors

few uncorrelated latent variables

	s_1	s_{1000}	e_1	e_{1000}	activity
mol 1
mol 2
....

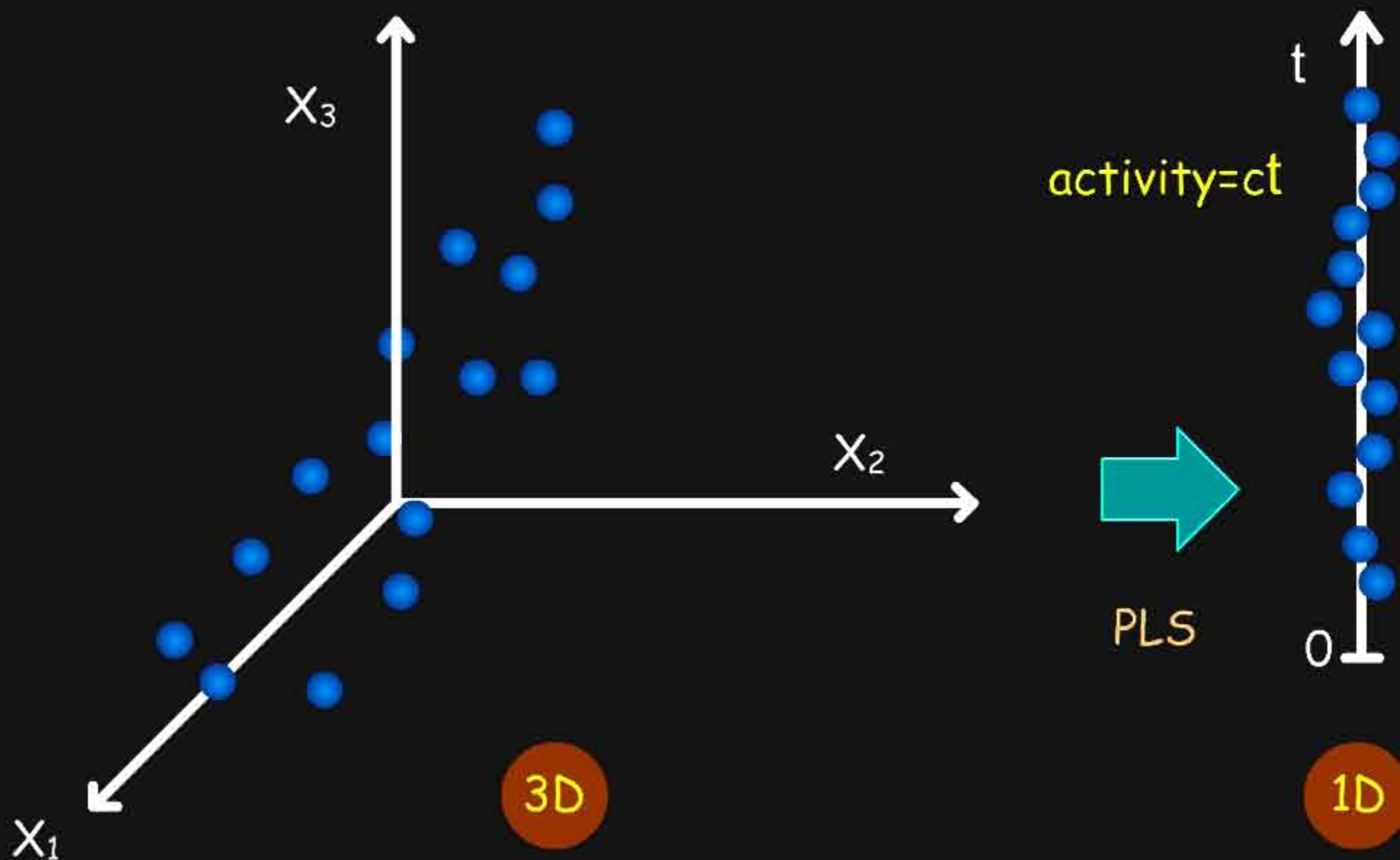
PLS



	t_1	t_2	t_{40}	activity
mol 1
mol 2
....

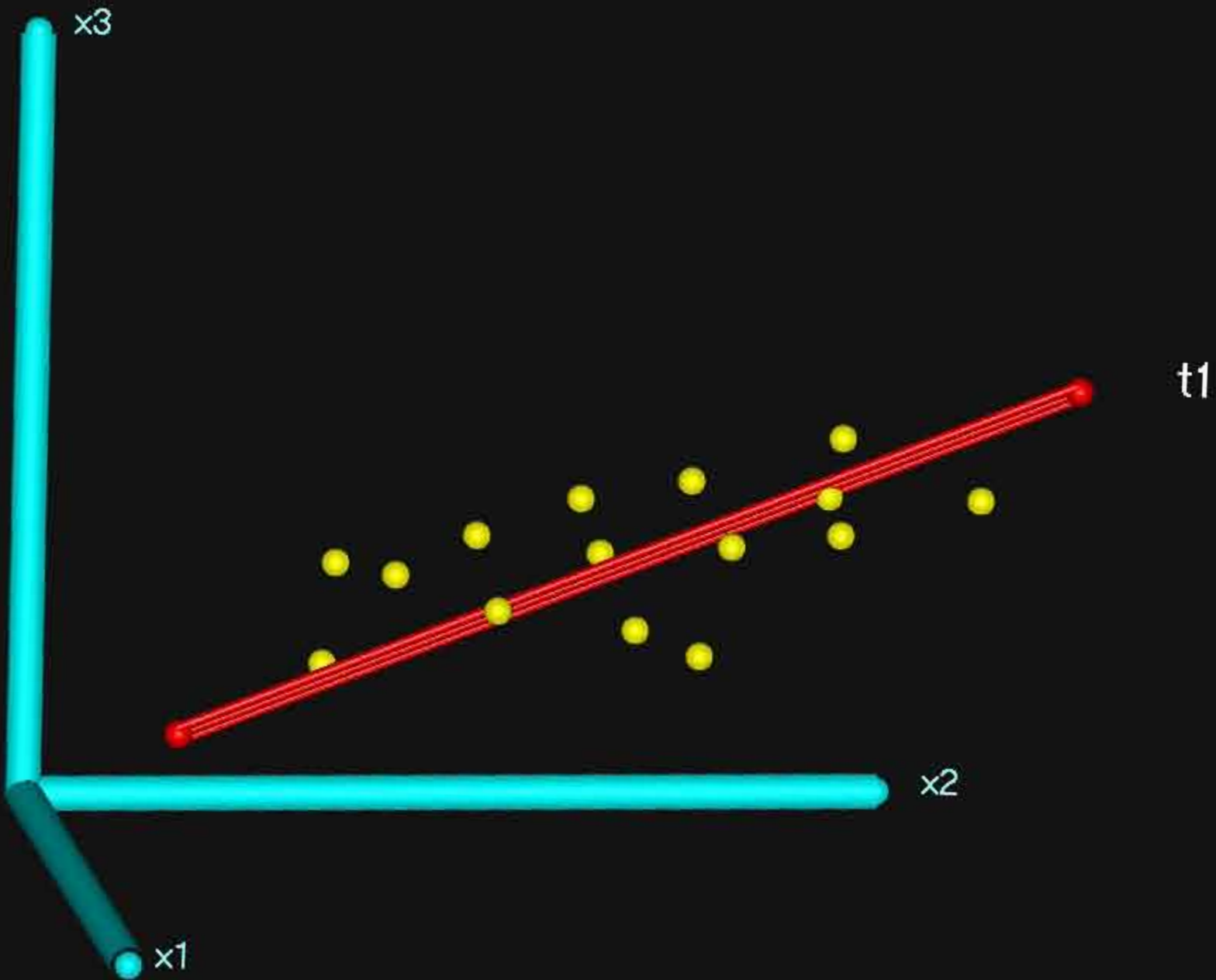
F2.4.22 Geometrical Interpretation of PLS

Consider a case where we want to predict the activity for n molecules using values of 3 descriptors (X_1, X_2, X_3). PLS aims at reducing the initial referential to a space of reduced dimension and generating a reduced set of variables t (the t 's being a linear combination of the original X 's). In the example below, the 3D space has been reduced to 1D where the biological activities are predicted in a simple linear fashion, $\text{activity} = ct$.



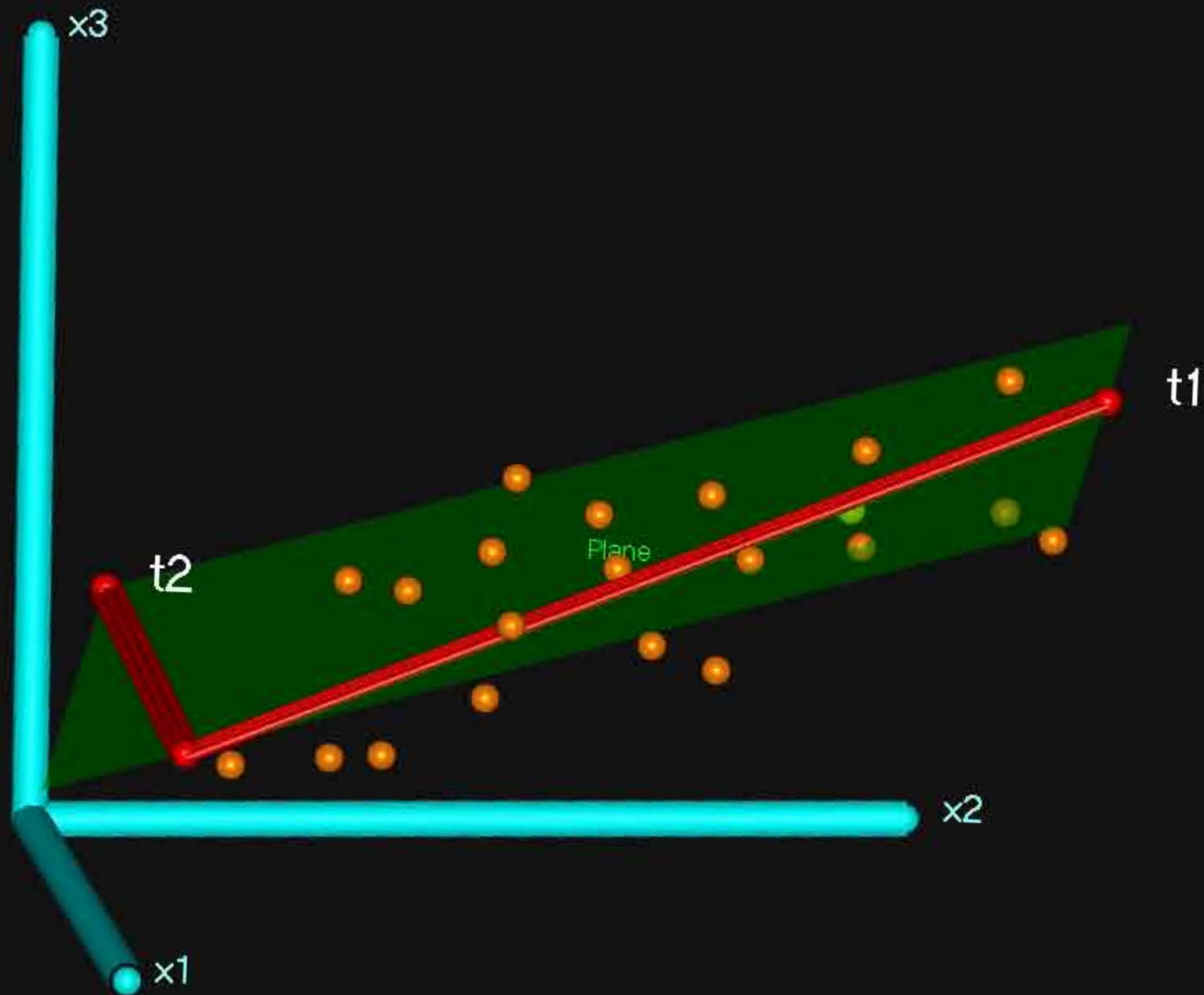
F2.4.23 The First PLS Component

The first PLS component (t_1 axis) is a line in the initial X -space which satisfactorily approximates the points in terms of least squares and at the same time provides a good correlation in the t -space.



F2.4.24 The Second PLS Component

The second PLS component (t_2 axis) is a line orthogonal to the first PLS component, which correctly approximates the points in the (t_1, t_2) plane, and at the same time provides an improved correlation in the t -space. Subsequent PLS components are derived in a similar manner.



F2.4.25 3D-QSAR Equation in the PLS Space

Thanks to the reduction of the number of terms by PLS calculations, an equation in the form of a linear relationship between structure and activity is obtained. Note that the new variables " t_i " do not have a structural meaning.

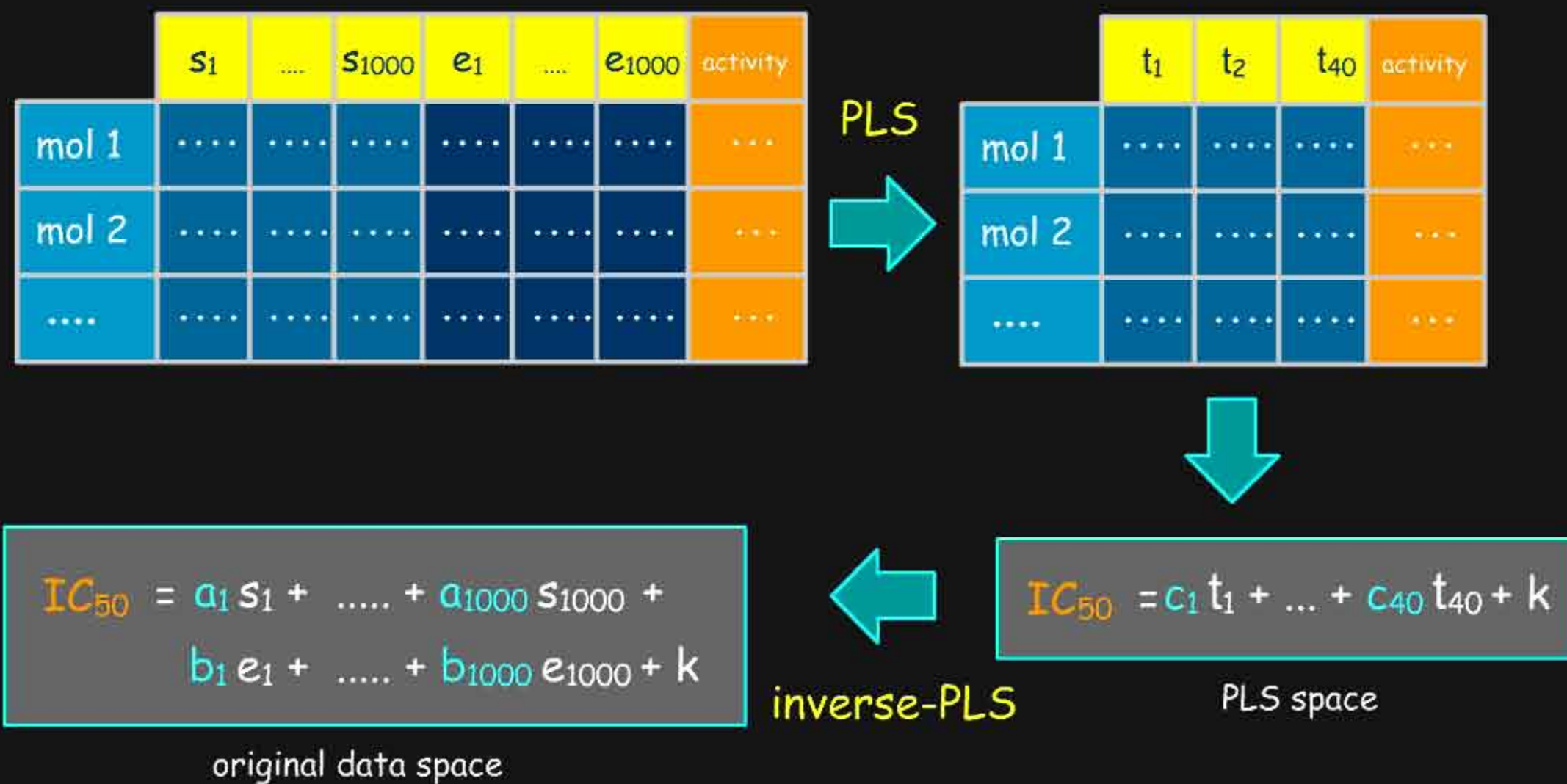
	t_1	$t_2 \dots$	t_n	activity
mol 1
mol 2
.....



$$\text{biological activity} = c_1 t_1 + c_2 t_2 + c_3 t_3 + \dots + c_n t_n + k$$

F2.4.26 Back to Space of Original Descriptors

The 3D-QSAR PLS equation is a good mathematical model, with however a poor structural content. To circumvent this drawback, this equation is projected back into the space of the original descriptors, to yield an equivalent CoMFA equation that can be better exploited in structural terms.



F2.4.27 The 3D-QSAR Equation in the Original Data Space

The CoMFA equation consists of a linear combination of S_i (the steric field) and e_i (the electrostatic field) calculated at each point of the lattice. This equation can be used to highlight regions in space where steric or electrostatic interactions are critical for the activity, and this will be presented a few pages later.

$$\text{Activity} = c + \sum_{i=1}^N a_i S_i + \sum_{i=1}^N b_i e_i$$

S_i = steric field at the i^{th} grid point


e_i = electrostatic field at the i^{th} grid point

a_i & b_i = regression weights

F2.4.28 Many Terms in the 3D-QSAR Equation

Unlike classical QSAR equations that can be clearly written, 3D-QSAR equations never appear explicitly. This equation contains thousands of terms, each of which is associated to a particular (x_i, y_i, z_i) , which is impossible to represent in a linear expression. The CoMFA equation can be fully assessed and exploited however its linear form remains in the computer. The number of terms collected in the data space may be 25,000-55,000, depending on the number of molecules, the lattice dimension and spacing. PLS analyses can reduce them by a factor in the range of 4 to 40.

$34 \times 41 \times 18 = 25092$ grid points


$$\text{IC}_{50} = \left. \begin{array}{l} 3.2s_1 - 0.2s_2 + \dots - 4.1s_{25092} \\ 6.2e_1 - 4.0e_2 + \dots + 0.7e_{25092} + 4.6 \end{array} \right\} \begin{array}{l} 50185 \\ \text{terms} \end{array}$$

↓ PLS

$$\text{IC}_{50} = \left. 0.1t_1 - 4.2t_2 + \dots + 2.7t_{576} + 0.3 \right\} \begin{array}{l} 577 \\ \text{terms} \end{array}$$

F2.4.29 Measuring the Quality of the Relationship

The indexes used for assessing the quality of a 3D-QSAR model are the same as those already presented for the classical QSAR regression model: r^2 (squared correlation coefficient); TSS (total sum of squares), ESS (explained sum of squares), RSS (residual sum of squares), SDEP (standard deviation in error prediction), F_{value} (F-test of statistical significance), q^2 (cross-validated correlation coefficient) provide measures of the degree of correlation between the activity values calculated by the model and those measured experimentally. Usually different models are envisaged and these numbers serve to assess their respective quality and predictability.

$$IC_{50} = 0.1t_1 - 4.2t_2 + \dots + 2.7t_{576} + 0.3$$

correlation coefficient for
assessing the quality of the model

F-value for assessing
the statistical significance

$n = 31$; $r^2 = 0.92$; SDEP = 0.155; $F_{\text{value}} = 66.4$; $Q^2 = 0.80$

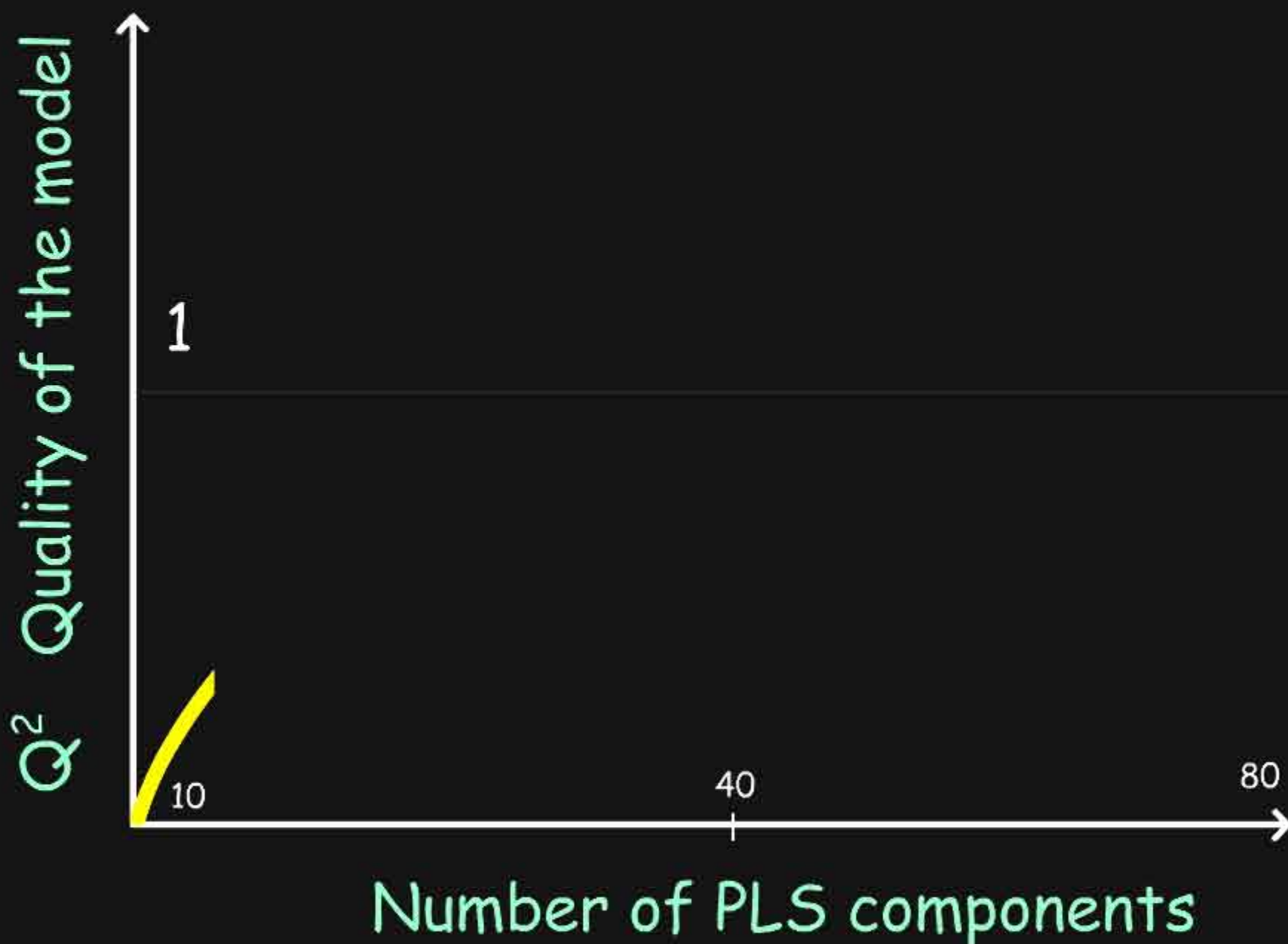
number of molecules

Total sum of squares
Standard deviation
in error prediction

regression coefficient for
measuring the predictability

F2.4.30 Total Number of PLS Components

The method known as 'cross validation' is used to fix the total number of PLS components (axis). The calculations are repeated with a randomly chosen set of compounds, and the resulting model is used to predict the missing activity data. Q^2 is the correlation between the measured activities and the predicted ones. New PLS components are added as long as Q^2 continues to increase.



F2.4.31 Two Equivalent 3D-QSAR Equations

We now have two equivalent 3D-QSAR equations: the first one is the PLS equation (in the t-space), and the second is the projection of the first in the space of the original descriptors. The former is useful for numerical calculations, and the later for 3D visualizations to provide insight into the structural content of the resulting 3D-QSAR model.

t-space

Predicting the Activities of New Compounds

$$\text{biological activity} = c_1 t_1 + c_2 t_2 + c_3 t_3 + \dots + c_n t_n + k$$

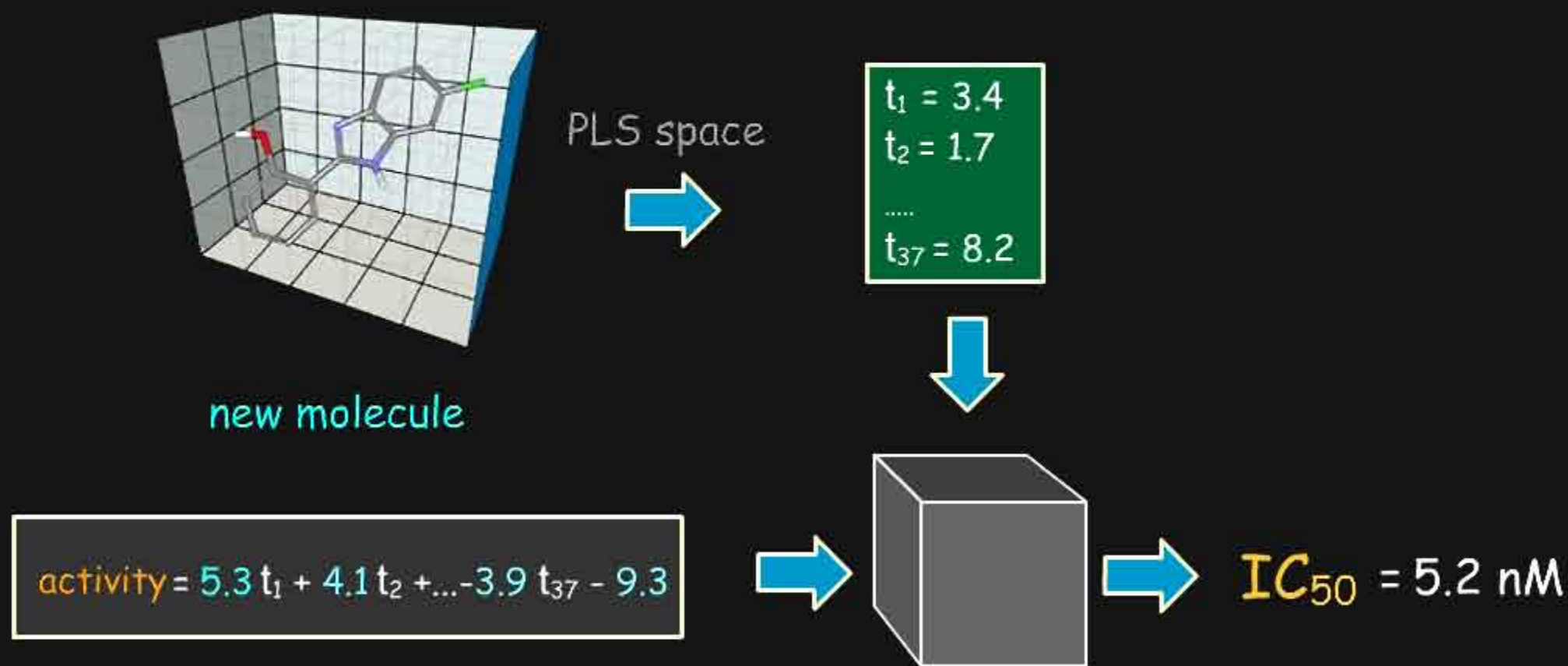
original data space

Visualization and 3D analysis of results

$$\text{biological activity} = a_1 s_1 + a_2 s_2 + a_3 s_3 + a_4 s_4 + \dots + a_{1000} s_{1000} + b_1 e_1 + b_2 e_2 + b_3 e_3 + b_4 e_4 + \dots + b_{1000} e_{1000} + k$$

F2.4.32 Predicting the Activities of New Compounds

Once a model is formed it can be used to predict the biological activity of a molecule which has yet to be synthesized and tested. The same sequence is repeated: alignment, interaction fields and projection in the PLS space to predict its activity. If a designed molecule is structurally new, care must be taken when aligning it with the reference compounds.



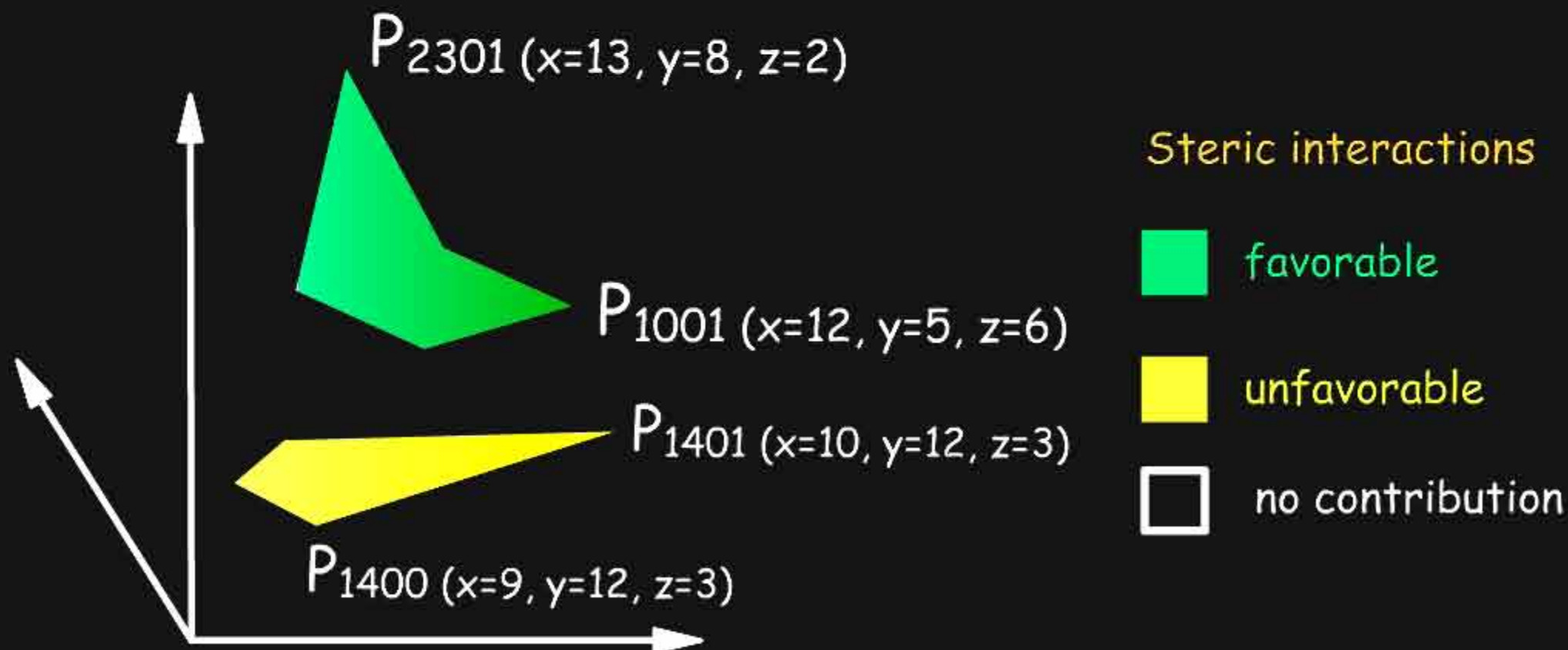
F2.4.33 CoMFA Coefficient Contour Maps

Thanks to the back-projection of the PLS equation into the space of the original data, the resulting equation can be exploited for the construction of CoMFA contour maps. A typical contour map is created by connecting points of the 3D grid with similar favorable or unfavorable coefficient values of a given field, multiplied by the standard deviation. The visualization of the contour map highlights regions where interactions are critical for activity. In the example below, point P2301 contributes sterically to 0.1% of the activity.

● Steric Component

● Electrostatic Component

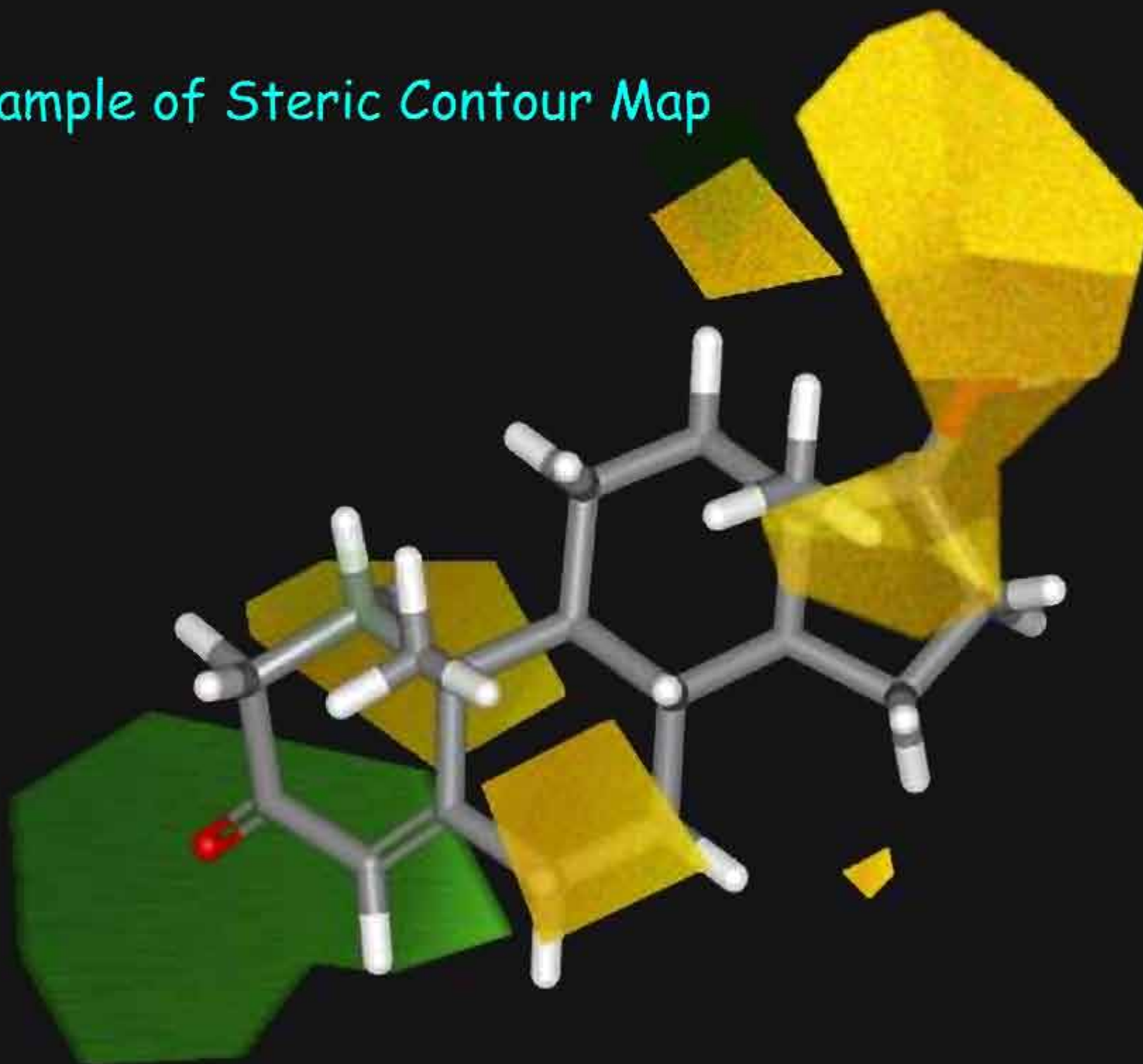
$$\dots 0.2S_{1001} + \dots + \boxed{0.0001S_{1010}} + \dots - 0.04S_{1400} - 0.05S_{1401} - \dots + 0.1S_{2301}$$



F2.4.34 CoMFA Steric Contour Map

In the example below a steric coefficient contour map is visualized, on top of a reference molecule. Green and yellow contours (a color code now widely adopted) indicate regions where bulky groups increase or decrease activity, respectively.

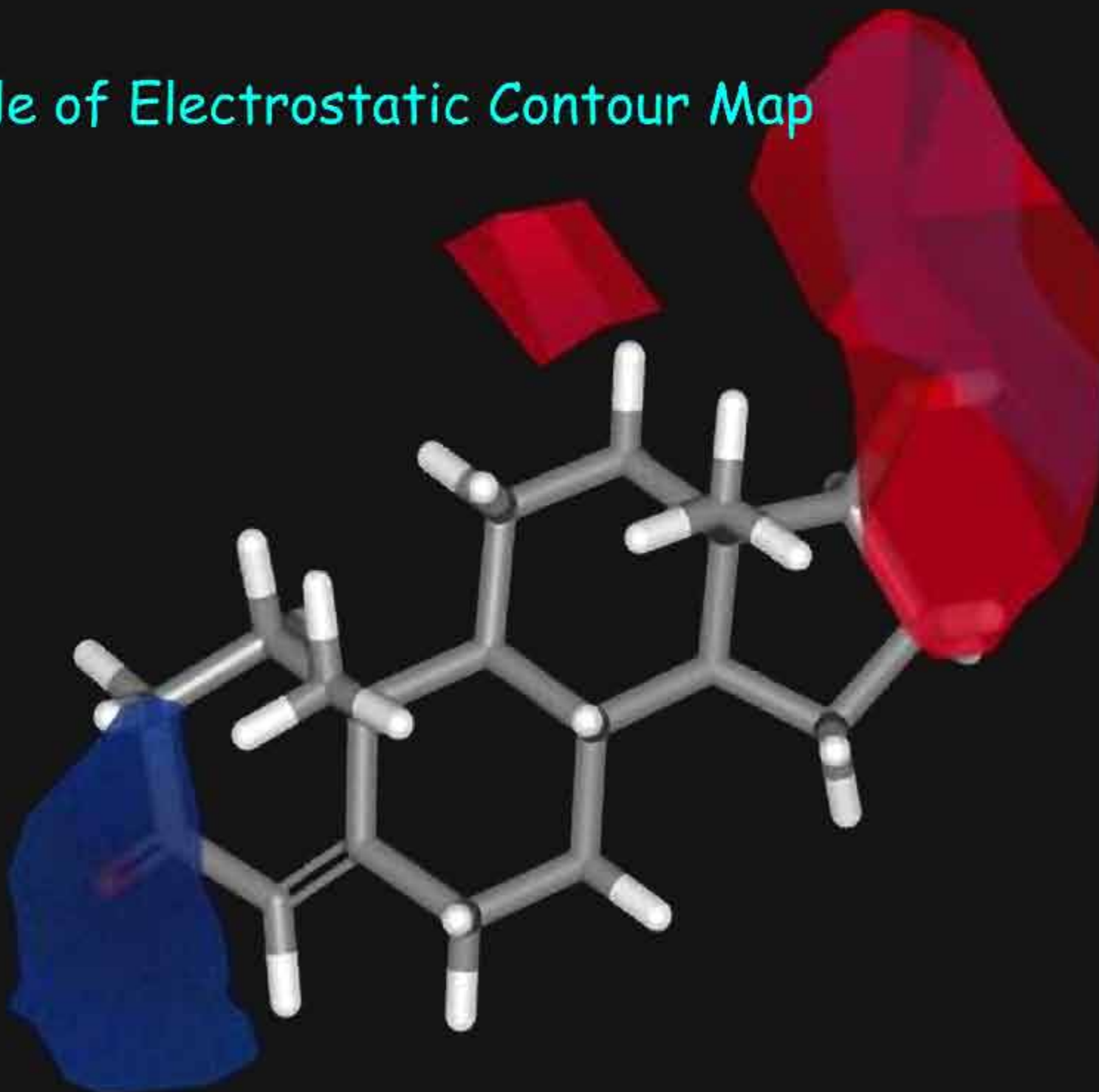
Example of Steric Contour Map



F2.4.35 CoMFA Electrostatic Contour Map

Electrostatic contour maps are constructed in a similar manner, and can be visualized with a now widely adopted color code. Blue contours indicate regions where electropositive groups increase activity, whereas red contours represent regions where electronegative groups increase activity.

Example of Electrostatic Contour Map



F2.4.36 CoMFA Contour Maps vs. MIF Contour Maps

The CoMFA maps help understand the SAR, give an idea of the nature and the regions where interactions with the receptor are critical, and also to exploit them for designing new molecules. These maps have high informational content, contrary to MIF iso-potential maps that only visualize a molecular property, with no relationship whatsoever with the biological effects.

MIF map

Molecular Property



Visual representation of a property for a single molecule

Poor informational content

CoMFA map

3D-QSAR Equation

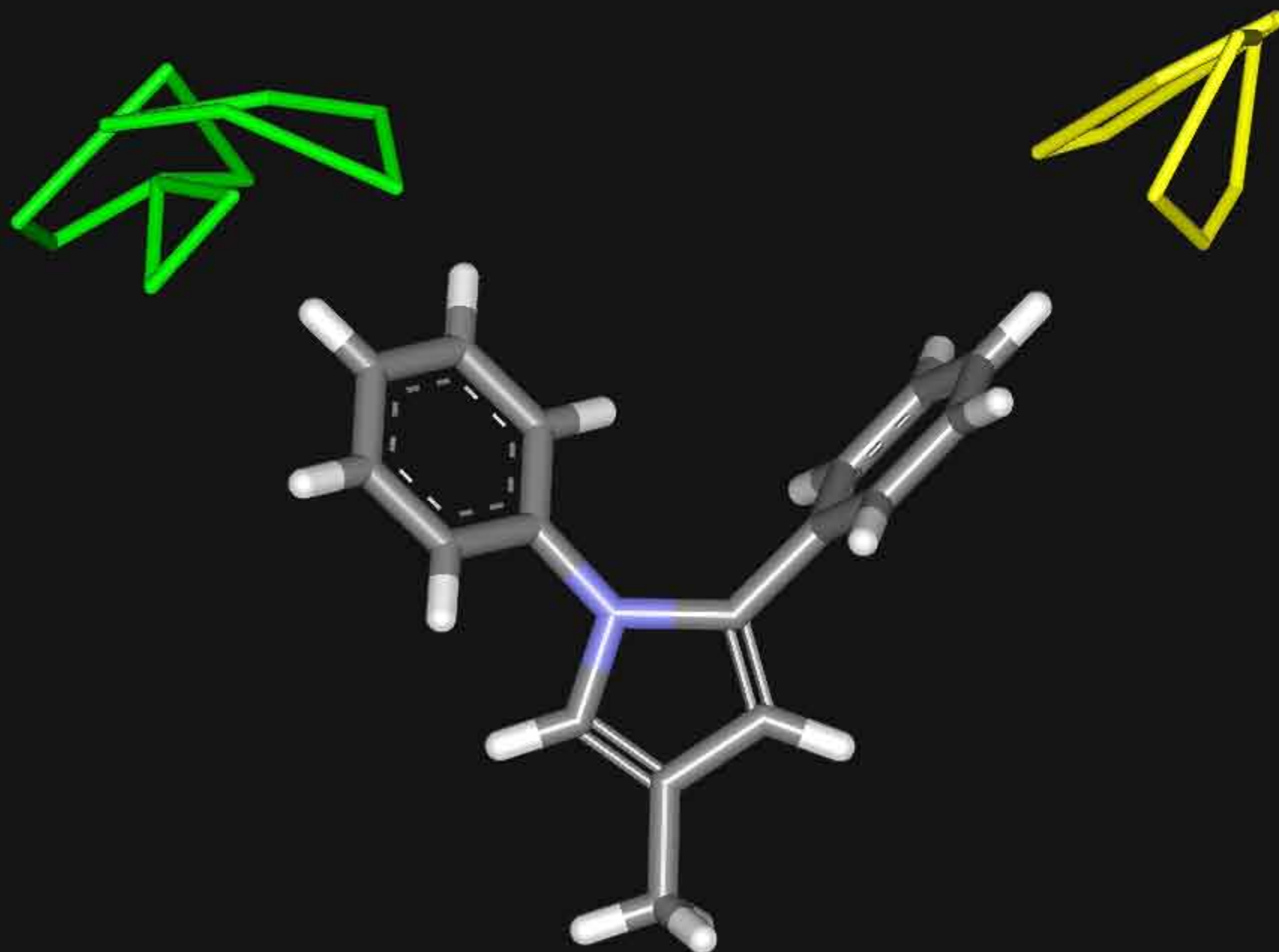


Visual understanding of the structure-activity relationships for a set of molecules

High informational content

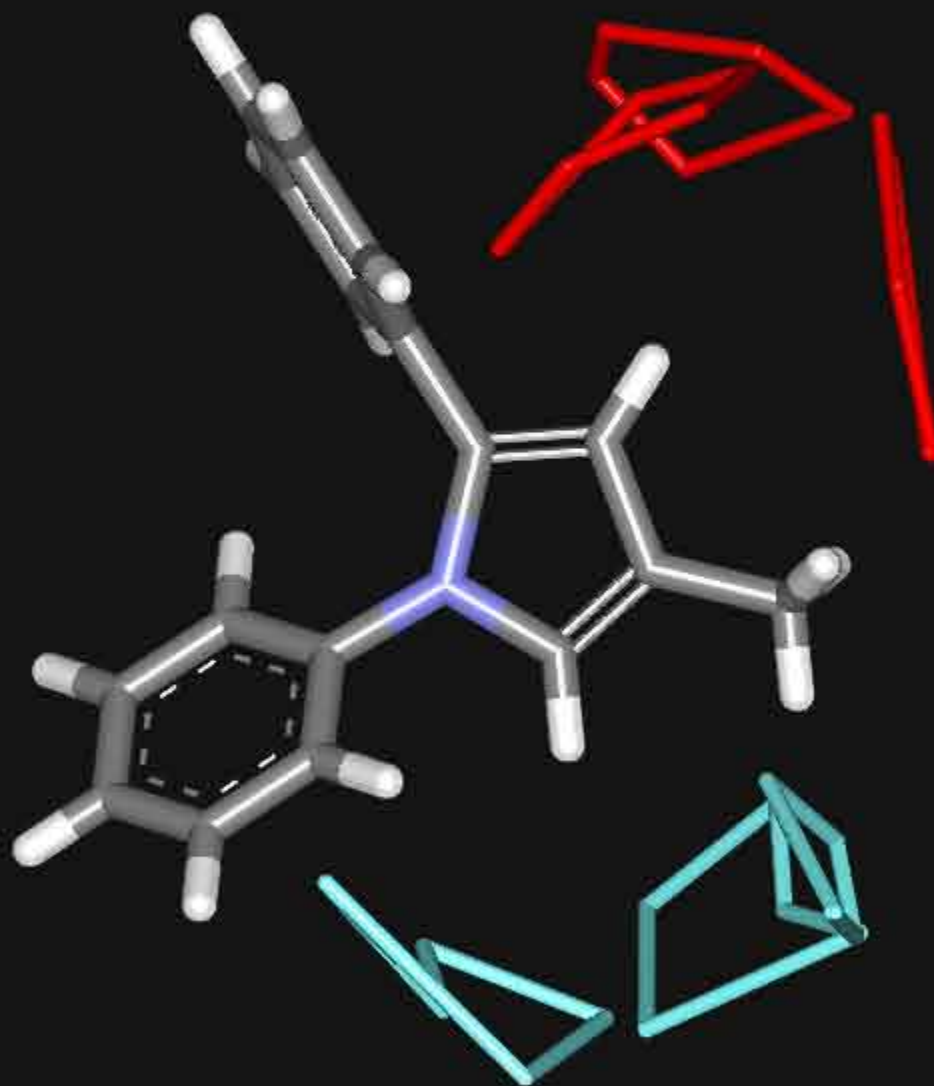
F2.4.37 Analysis of Steric Contour Maps

The visual analysis of a CoMFA steric map is straightforward. One can recognize regions where steric interactions with the receptor are favorable (green regions) or unfavorable (yellow regions). It is easy to imagine what kind of analogs are likely to be more potent or less potent than the reference compound.



F2.4.38 Analysis of Electrostatic Contour Maps

A ComFA electrostatic contour map reveals critical regions: the blue areas correspond to regions where electronegative groups decrease activity and electropositive groups increase activity; by contrast red areas correspond to regions where electronegative groups increase activity and electropositive groups decrease activity. One can imagine what kind of analogs are likely to be more potent or less potent than the reference compound.



F2.4.39 Exploitation of the Steric Contour Map

In the example below, when bulky para substituents are introduced in the phenyl ring pointing towards the green area, biological activities are increased; whereas para substitution of the second phenyl ring is detrimental to these activities.

Steric Contours

Green: steric bulk favorable

Yellow: steric bulk unfavorable

Reference Compound

IC₅₀= 1100 nM

Activities increased

IC₅₀= 46 nM

IC₅₀= 35 nM

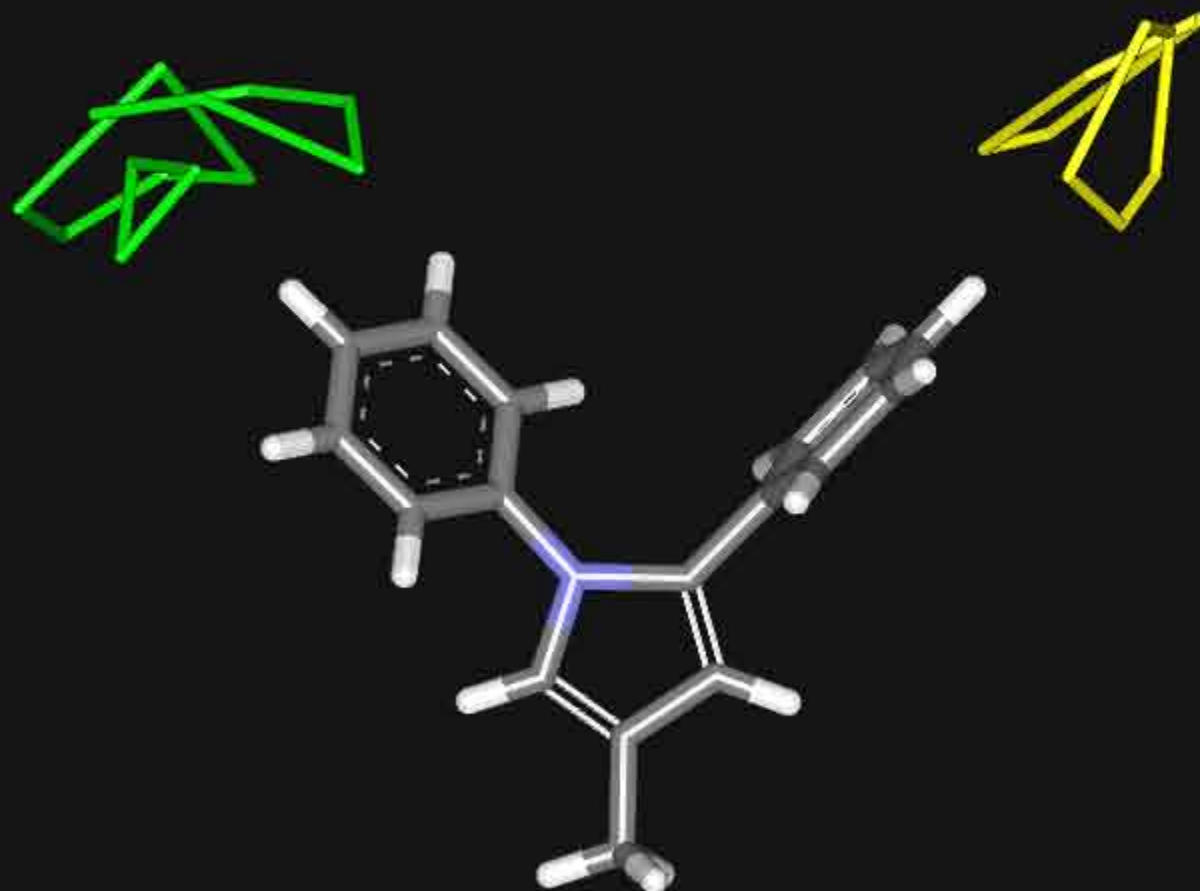
Activities decreased

IC₅₀= 2120 nM

IC₅₀= 3600 nM

Color by Molecule

Stick



F2.4.40 Exploitation of the Electrostatic Contour Map

When electronegative groups (Cl, CF₃) are introduced into the red areas biological activities are increased; and when electronegative groups (Br, CN) are in the blue areas biological activities are decreased.

Electrostatic Contours

- Blue: act.decr. by neg. charge
- Red: increased neg. charge

Reference Compound

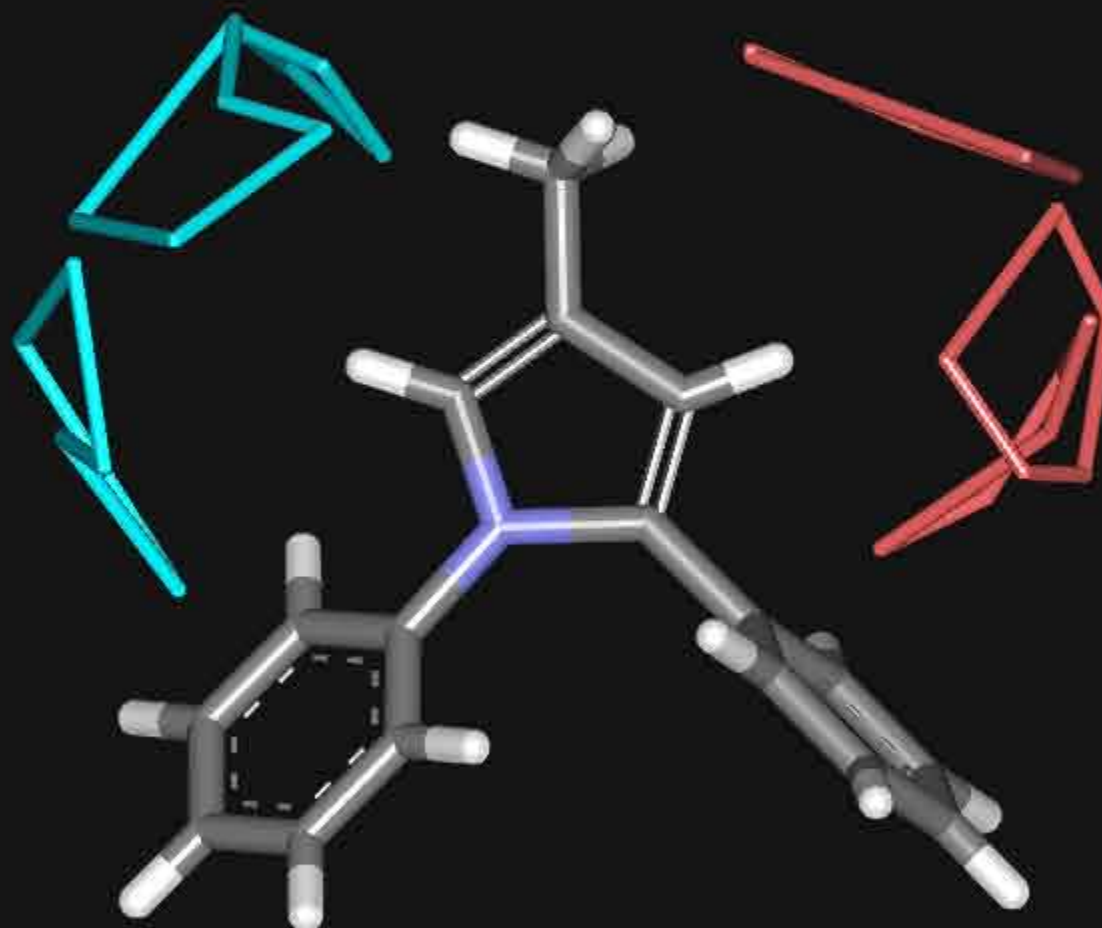
- IC₅₀= 1100 nM

Activities increased

- IC₅₀= 22 nM (Cl)
- IC₅₀= 9 nM (CF₃)

Activities decreased

- IC₅₀= 2300 nM (Br)
- IC₅₀= 5200 nM (CN)

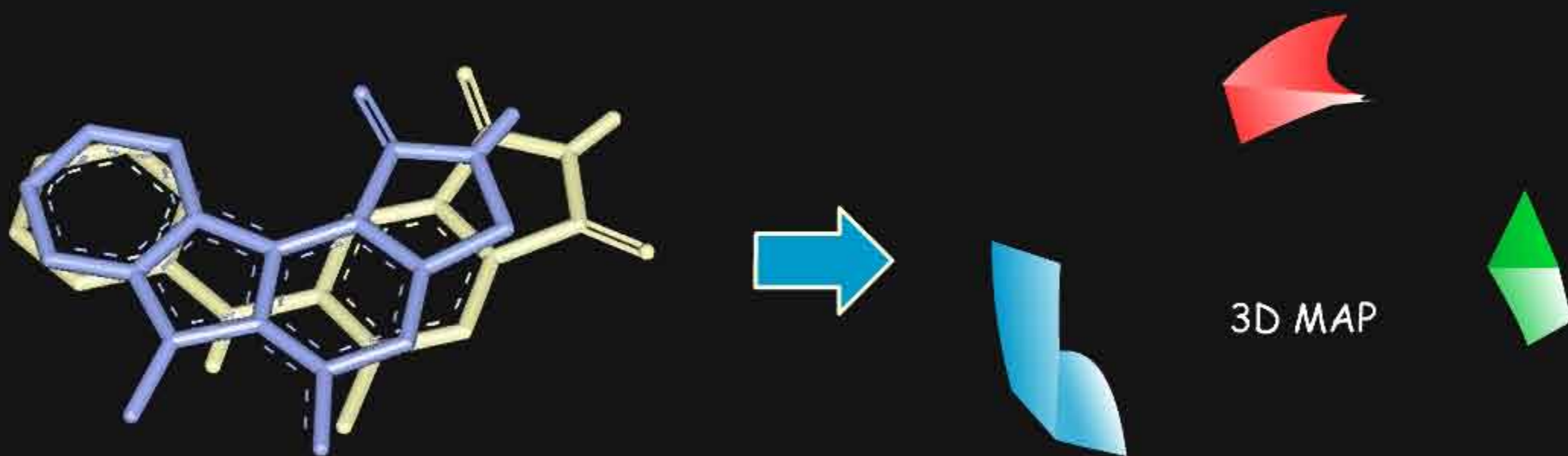


Color by Molecule

Stick

F2.4.41 Stability Problem of CoMFA Models

CoMFA models are sensitive to the parameters used to construct them. For example, very small changes in the relative alignment of the molecules can result in substantially different regression models. Other parameters include the lattice (spacing, orientation and dimensions) or the methods used for calculating the fields. The stability of the models is a difficult issue which is explicitly addressed by the new methods currently in development.





The topic Example of CoMFA Analysis: Steroids contains the following 10 pages:

- The Reference Compounds
- The Biological Data
- Molecular Alignment
- CoMFA Field Calculations
- CoMFA and PLS Results vs. Classical QSAR
- Steric CoMFA Map for Binding to TBG
- Electrostatic CoMFA Map for Binding to TBG
- CBG Affinities of New Steroids
- Predicting the CBG Affinities of New Steroids
- A Benchmark Set for 3D-QSAR

F2.5.1 The Reference Compounds

This example was presented by Cramer in the original paper introducing CoMFA. This historical contribution has been cited repeatedly in the literature and often used to illustrate the methodology. The 3D-QSAR approach was developed to analyze the binding affinities of 21 steroid compounds that interact with the following two receptors: the human corticosteroid-binding globulin (CBG) receptor and the human testosterone-binding globulin (TBG) receptor.

● 1-4

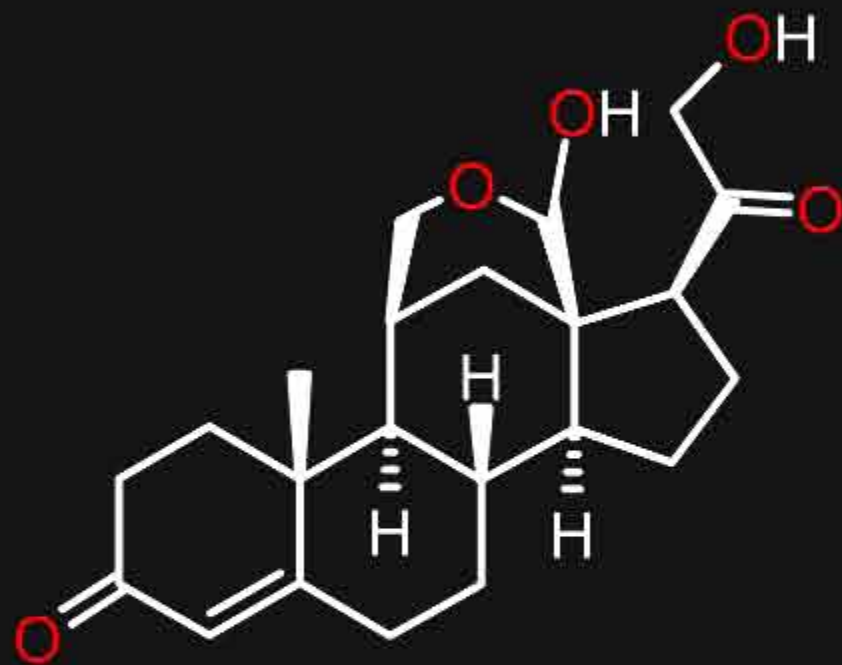
● 5-8

● 9-12

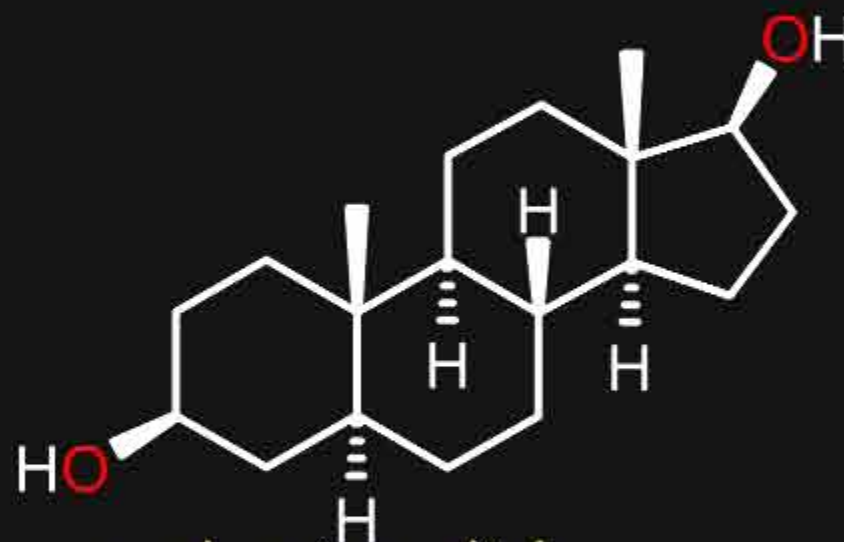
● 13-16

● 17-20

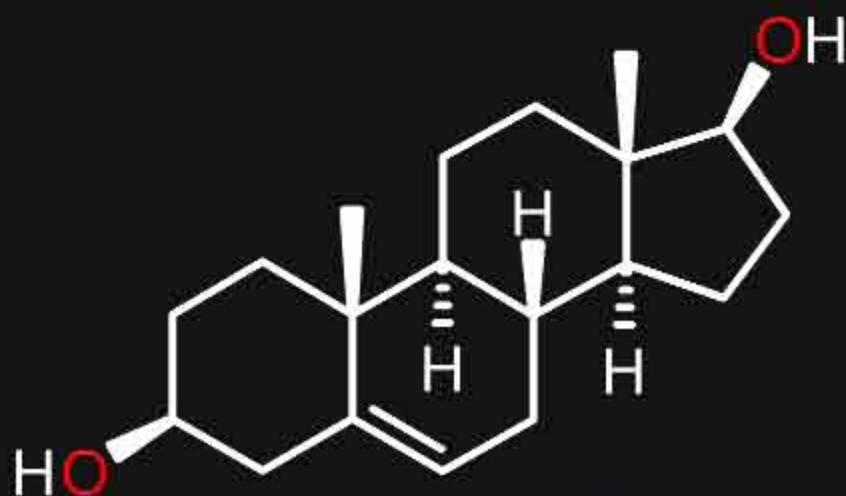
● 21



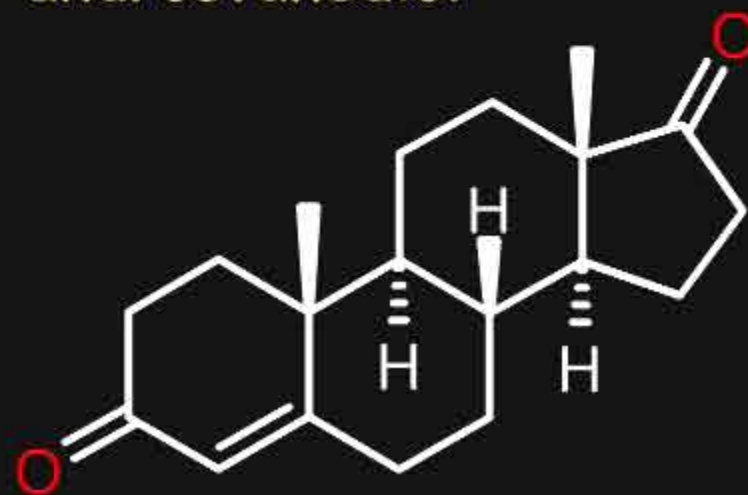
aldosterone



androstanediol



androstenediol



androstenedione

F2.5.2 The Biological Data

The binding affinities (expressed as pK_i values) of the 21 steroids for the two receptors CBG and TBG are shown below.

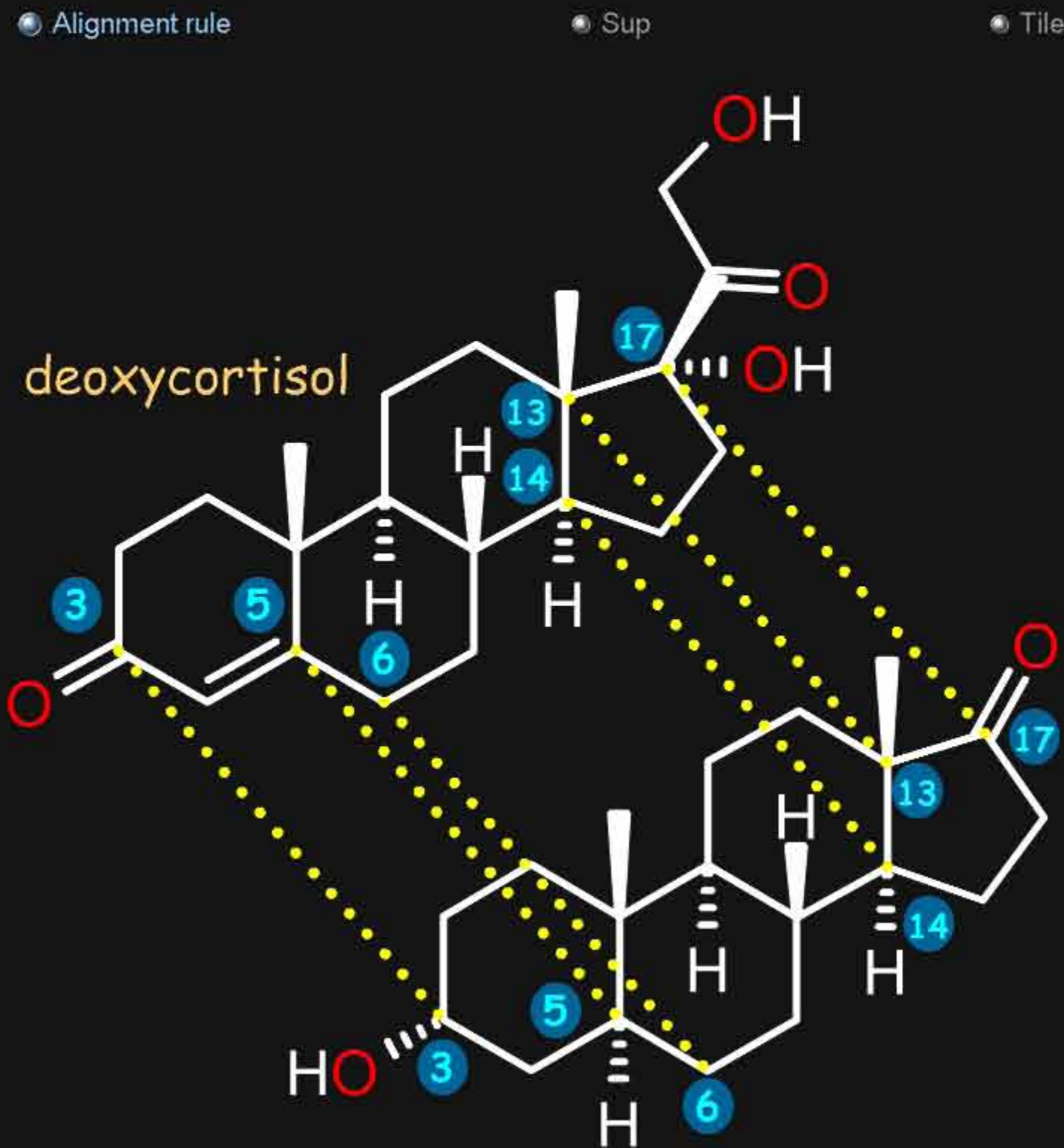
● 1-11

● 12-21

	Molecules	TBG	CBG
1	aldosterone	5.322	6.279
2	androstanediol	9.114	5.000
3	androstenediol	9.176	5.000
4	androstenedione	7.462	5.763
5	androsterone	7.146	5.613
6	corticosterone	6.342	7.881
7	cortisol	6.204	7.881
8	cortisone	6.431	6.892
9	prasterone	7.819	5.000
10	deoxycorticosterone	7.380	7.653
11	deoxycortisol	7.204	7.881

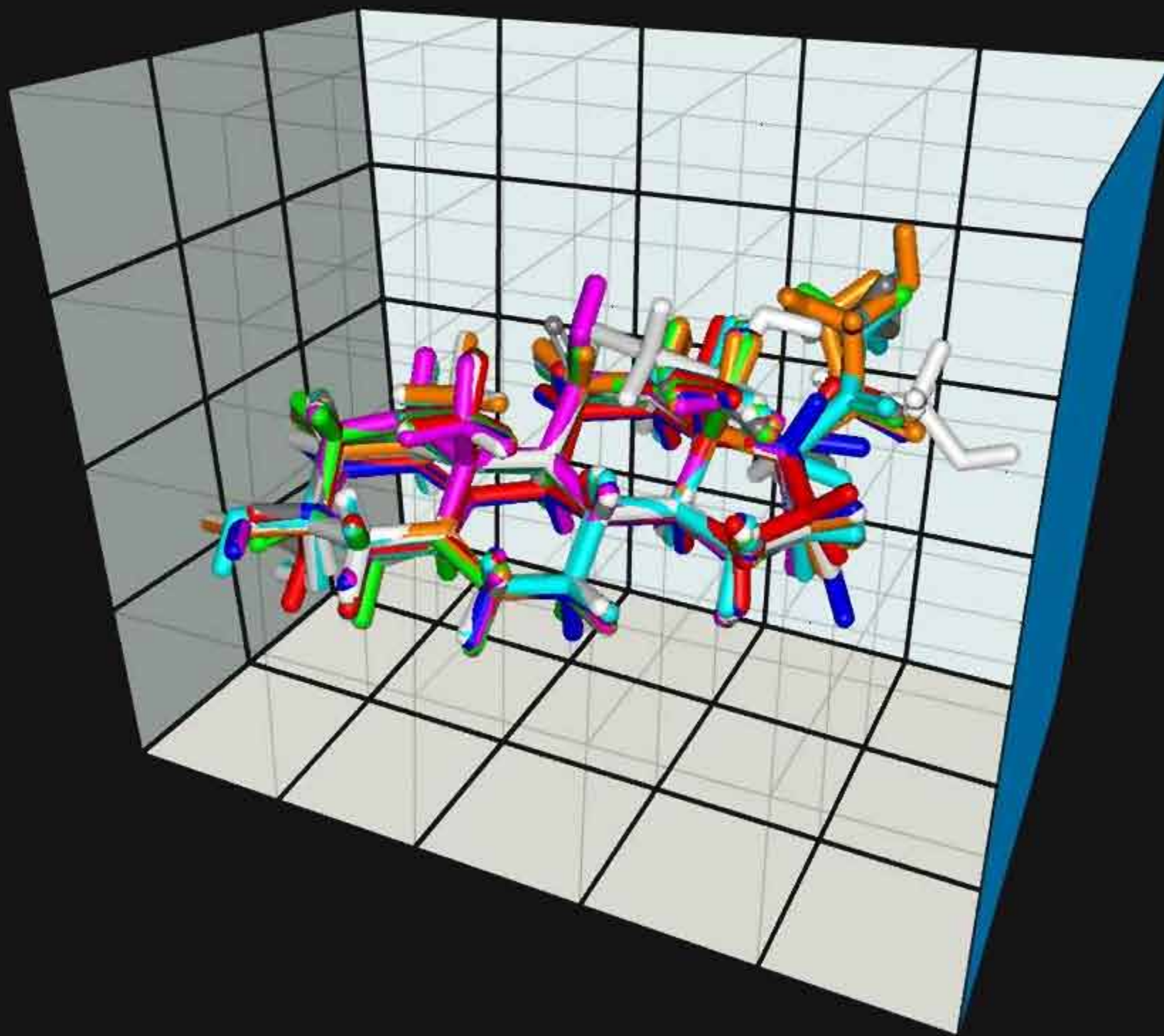
F2.5.3 Molecular Alignment

Due to the relative rigidity of the steroid skeleton, alignment was based on a simple rigid-body least-squares fitting of the 3, 5, 6, 13, 14 and 17 carbon atoms of each steroid to the corresponding atoms of deoxycortisol, a steroid with good binding affinity to both globulins. The 21 superimposed steroids are shown below (button "Sup").



F2.5.4 CoMFA Field Calculations

The set of 21 aligned steroids was placed in a 3D grid large enough to encompass all structures. The distance between neighboring grid points was about 2Å. Steric and electrostatic grids were calculated using the 6-12 Lennard-Jones and Coulomb potentials respectively, by using an sp^3 carbon atom with a +1 charge as a probe.



F2.5.5 CoMFA and PLS Results vs. Classical QSAR

Binding affinities to human CBG and TBG were predicted through 3D-QSAR and PLS. The results obtained with the CoMFA methodology were better than those obtained with classical QSAR. Note that only a few descriptors (MR, LogP, MP...) were used for the QSAR; note also that MLR (multiple linear regression) cannot handle the huge amount of data generated by CoMFA calculations.

		CBG			TBG	
		Method	r^2	Q^2	r^2	Q^2
classical QSAR	Molar refractivity	MLR	0.43	0.31	0.30	0.22
	LogP + Melting Point (CBG)	MLR	0.03	0.18	0.37	0.23
	Molar refractivity + freeWilson indicator variables	MLR	0.34	0.20	0.69	0.42
3D-QSAR	COMFA	PLS	0.90	0.66	0.87	0.56

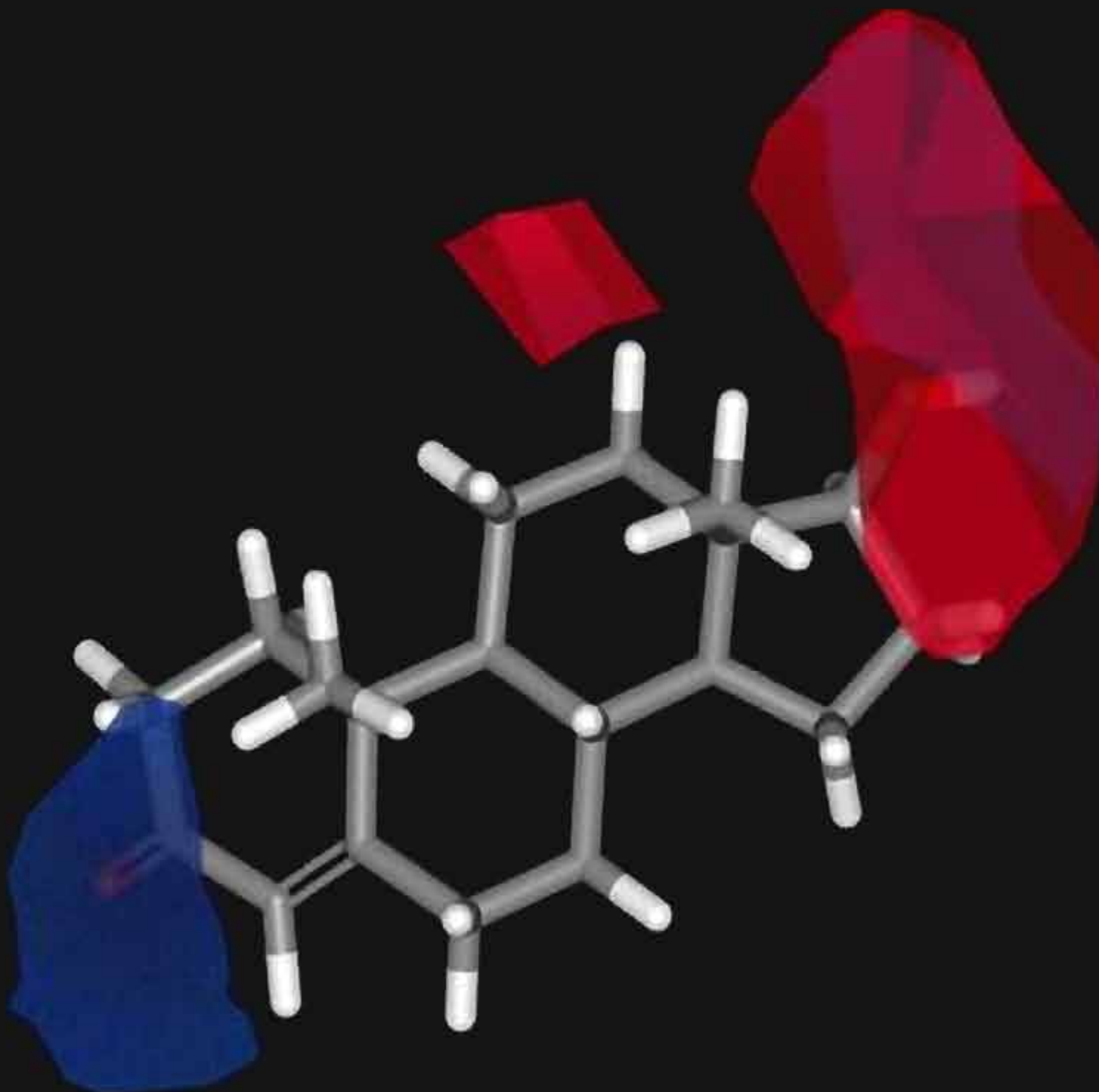
F2.5.6 Steric CoMFA Map for Binding to TBG

As revealed by the CoMFA regression coefficients, the steric 3D contour maps around testosterone for TBG affinities are presented below. Regions where bulky substituents enhance the binding affinity are shown in green, whereas regions where bulky substituents reduce the binding affinity are shown in yellow.



F2.5.7 Electrostatic CoMFA Map for Binding to TBG

The electrostatic 3D contour maps around testosterone for TBG affinities are presented below. Regions where electronegative substituents enhance the binding affinity are shown in red, and regions where electronegative substituents reduce the binding affinity are shown in blue.



F2.5.8 CBG Affinities of New Steroids

After completion of the CoMFA analyses, the binding affinities to human CBG (corticosteroid-binding globulin) of a set of 10 new steroids became available.

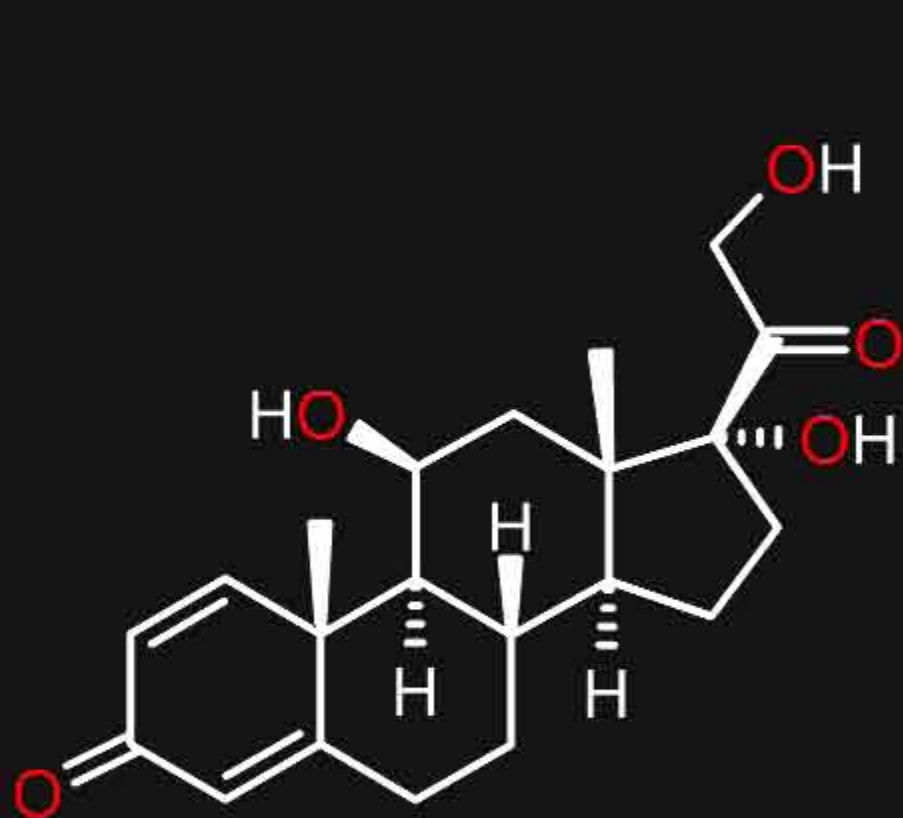
● 1-2

● 3-4

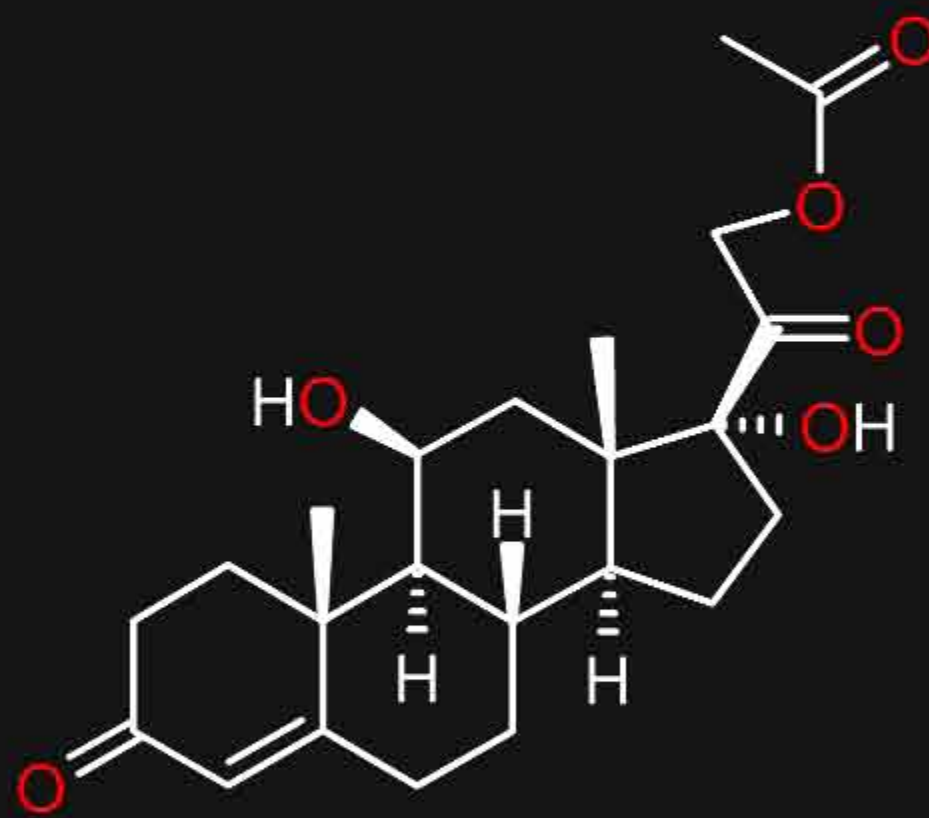
● 5-6

● 7-8

● 9-10



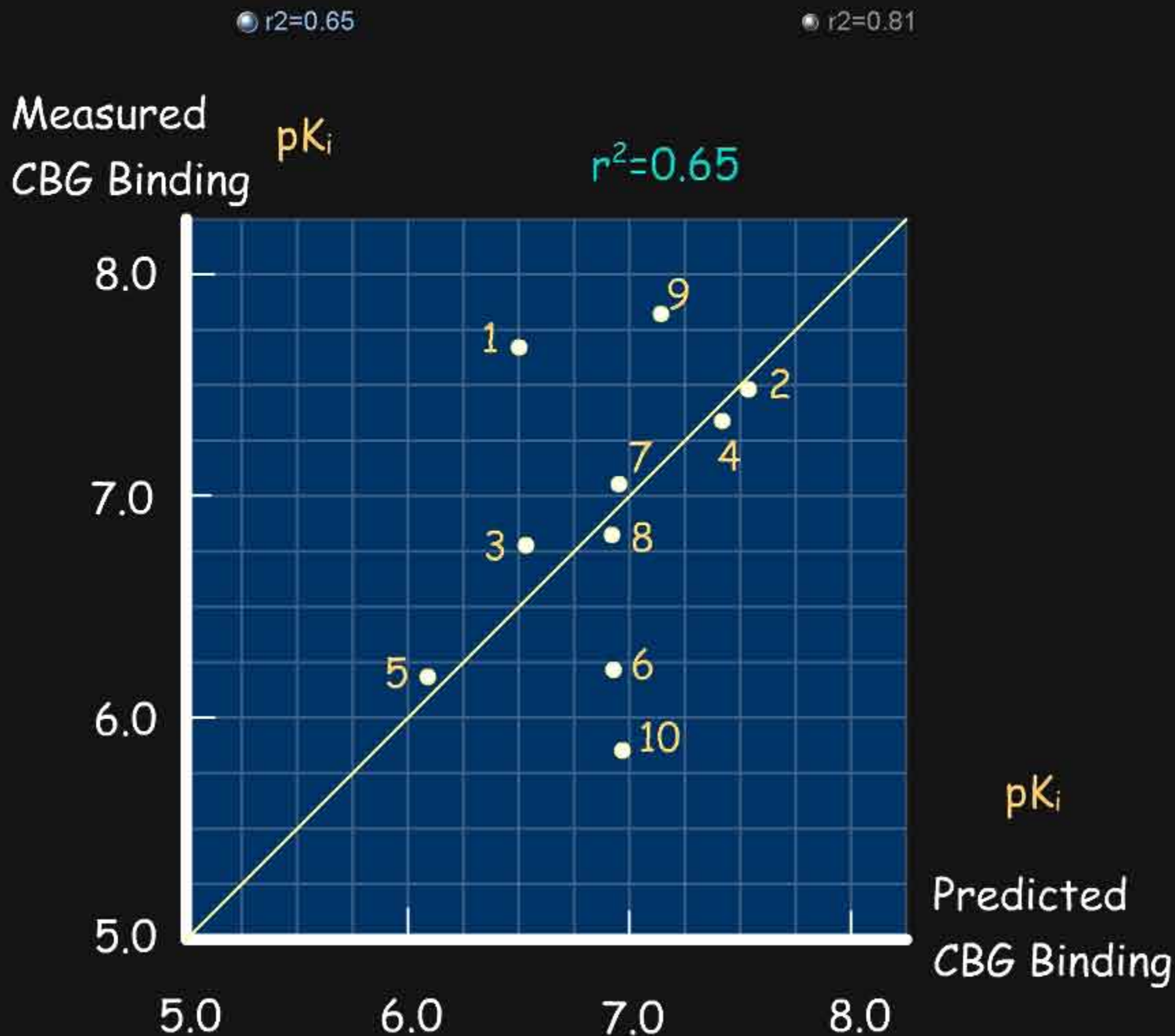
1



2

F2.5.9 Predicting the CBG Affinities of New Steroids

The 3D-QSAR equation derived by CoMFA was used to predict the binding affinities to the human CBG receptor of these 10 new steroids leading to a predictive r^2 value of 0.65; if compounds 1, 9 and 10 are removed from this analysis, r^2 increases to 0.81.

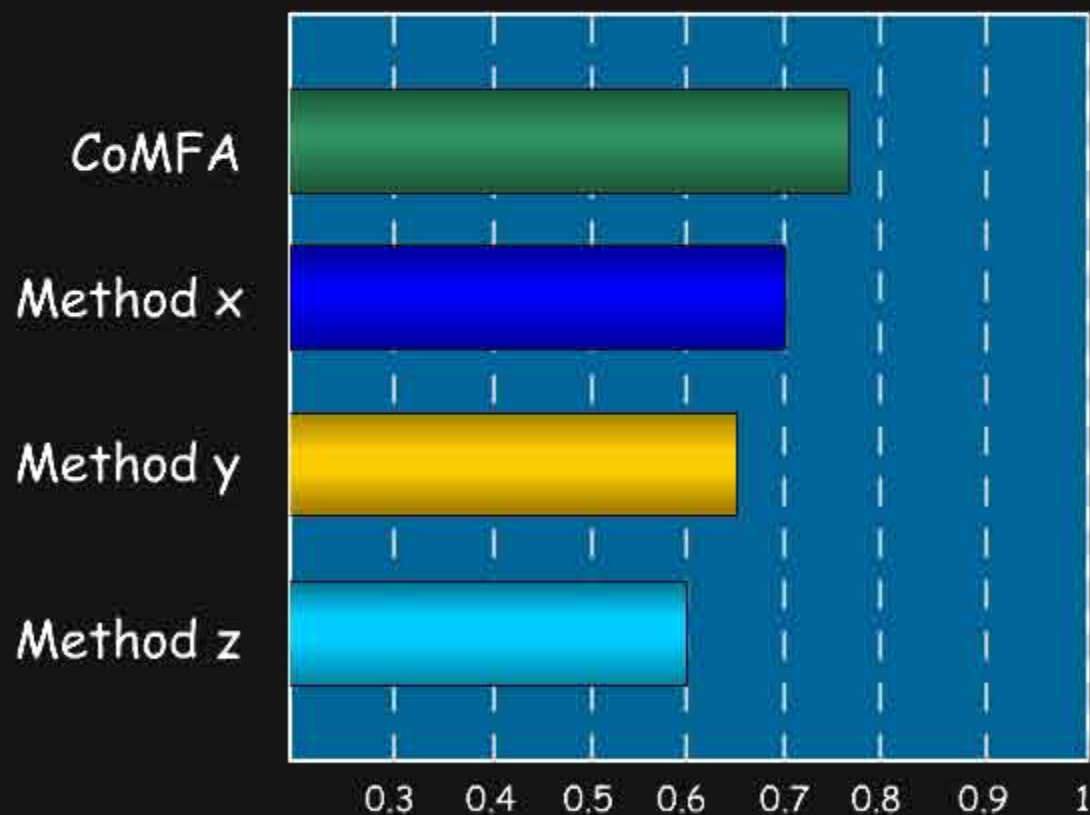
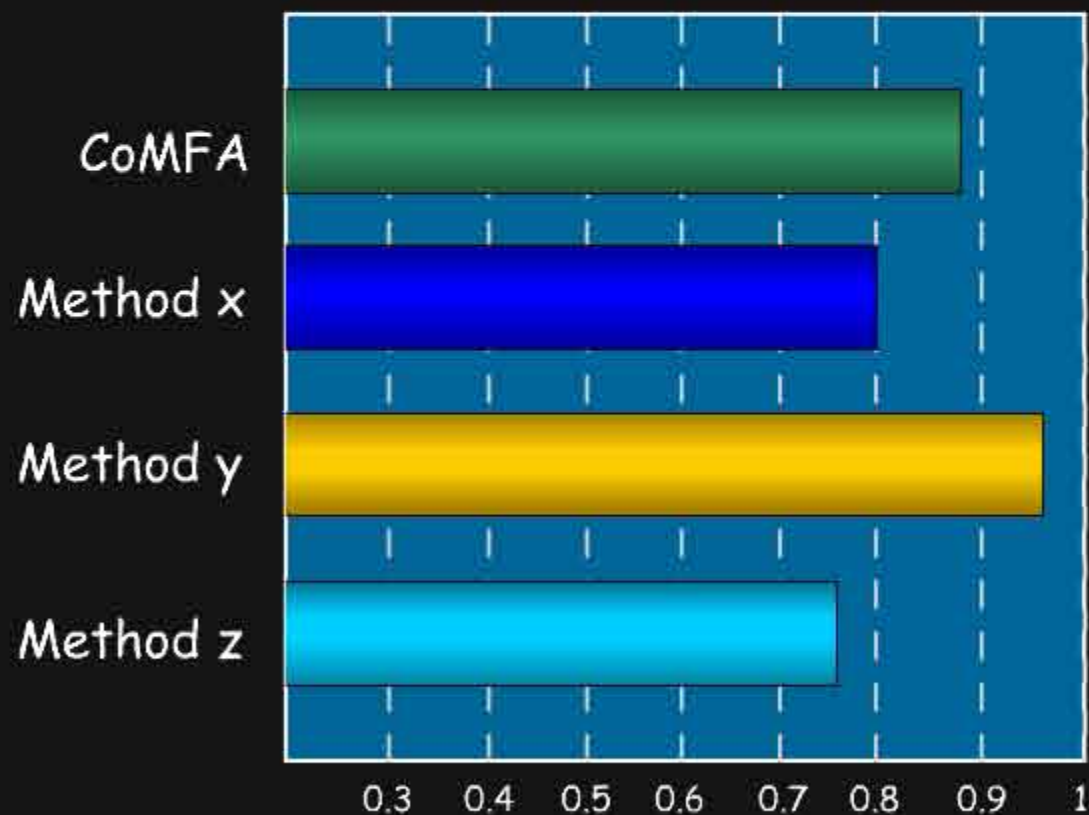


F2.5.10 A Benchmark Set for 3D-QSAR

The introduction of CoMFA triggered the development of an entire generation of new 3D-QSAR methods, which will be presented in the next section. The steroid data set used for illustrating CoMFA has now become a standard set for the validation of any novel 3D-QSAR approach.

r^2

Q^2



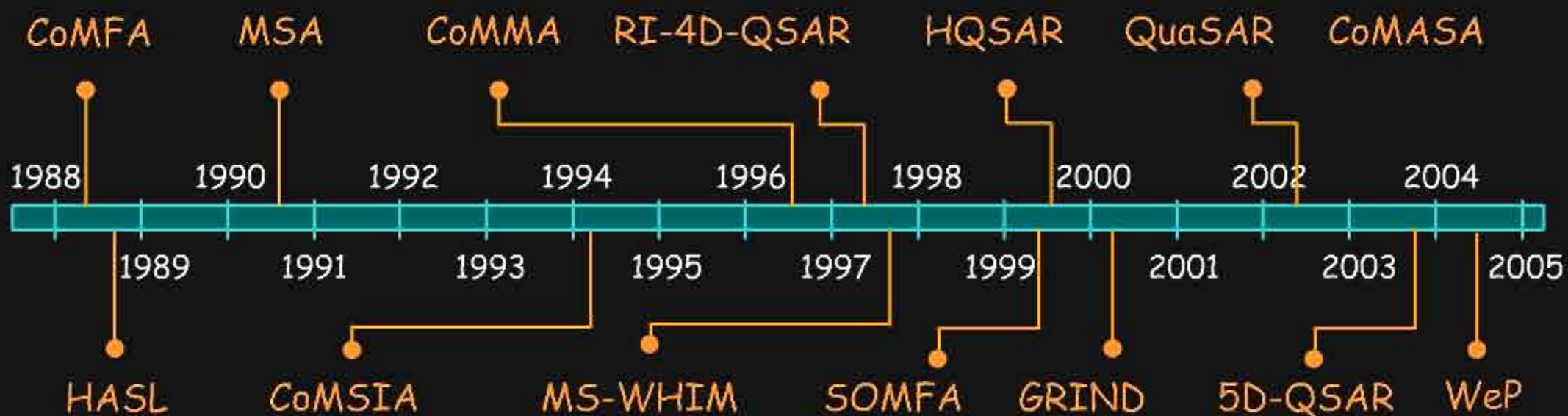


The topic Other 3D-QSAR Methods contains the following 13 pages:

- 3D-QSAR Programs
- Best Method?
- CoMFA
- HASL
- CoMSIA
- CoMMA
- MS-WHIM
- SOMFA
- HQSAR
- GRIND
- Quasar
- CoMASA
- WeP

F2.6.1 3D-QSAR Programs

After the introduction of CoMFA, other 3D-QSAR methods were developed by different groups. Some of these programs are mentioned below, listed by first date of publication. In the following pages the specificity of some of them is briefly outlined.



F2.6.2 Best Method?

It is not unusual to see QSAR practitioners simultaneously using several methods, comparing the corresponding models with the statistics associated to them, for purposes of deciding which method is the most suitable for their particular application. An objective comparison among methods is still not possible, in particular because of the stability problem mentioned previously. Many research groups have their favorite method and keep one or two data sets to validate new ones they develop.

Method	r^2	Q^2	F
CoMFA	0.85	0.65	132
CoMSIA	0.81	0.70	143
HQSAR	0.91	0.69	117
.....

F2.6.3 CoMFA

The CoMFA approach (comparative molecular field analysis, developed by R. Cramer et al.) uses steric and electrostatic fields to capture the properties of the space around a molecule. The probe atom is a carbon with a charge of +1. CoMFA paved the way for the development of several generations of new 3D-QSAR methods.

CoMFA

- First 3D-QSAR approach
- Two probes: steric and electrostatic
- Stability problems (small changes in the alignment)

F2.6.4 HASL

In HASL (hypothetical active site lattice, developed by A. Doweyko et al.) a molecule is converted into a set of regularly-spaced points (lattice) defined by Cartesian coordinates (x,y,z) and atom type. In this way individual molecules can be compared to one another, and their lattices merged to form a composite description called the HASL. The individual property of each molecule is distributed among the points in the HASL in such a way that the sum of the partial activity values associated with a set of points belonging to each molecule is equal to the total known value for that molecule. The result is a model of the receptor site consisting of points in space capable of predicting the activities of new molecules.

HASL

- Molecules converted to a "lattice" set of atoms ("lattice")
- Composite description as the merger of all lattices
- Activity = sum of activities of associated points

F2.6.5 CoMSIA

The CoMSIA approach (comparative molecular similarity index analysis, developed by G. Klebe et al.) uses molecular potentials that are smoothed with Gaussian functions to eliminate singularities at atomic nuclei (as observed in CoMFA). This has been proven to reduce the sensitivity to small changes in the alignment of compounds or the orientation of the grid. In addition to the steric and electrostatic fields CoMSIA includes hydrogen bonding (donor and acceptor) and hydrophobic fields.

CoMSIA

- Potential smoothed with Gaussian function
- Sensitivity to alignment reduced as compared to CoMFA
- H-bond (donor and acceptor) and hydrophobic fields

F2.6.6 CoMMA

CoMMA (comparative molecular moment analysis, developed by D. Silverman et al.) is a method based on descriptors such as the moments of inertia, dipole and quadrupole moments, that do not require molecular superposition or alignment for the assignment of molecular similarity. Initially developed for properties where the zero-order moment of the expansion vanishes, the approach has been extended to cases where this is not true, and higher levels of moment expansions of property fields are then considered (first-order, second-order moments).

CoMMA

- Moment descriptors
- e.g. inertia, dipole, quadrupole
- higher levels of moment expansions

F2.6.7 MS-WHIM

MS-WHIM (molecular surface weighted holistic invariant molecular, developed by G. Bravi et al.) is a method based on a small number (e.g. 12) of statistical parameters derived from molecular surface properties and calculated within different weighting schemes. The MS-WHIM fields contain structural information in terms of size, shape, symmetry and atom distribution that are invariant with respect to the coordinate system.

MS-WHIM

- Small number of statistical parameters
- Derived from molecular surface properties
- Invariant with respect of coordinate system

F2.6.8 SOMFA

SOMFA (self-organizing molecular field analysis, developed by D. Robinson et al.) is an approach where the fields are calculated as in CoMFA however their values are multiplied by the activity of the molecule expressed on a scale where the most active molecules have positive values and the least active molecules have negative values. Therefore the most active and least active molecules have higher values than the less interesting ones, which are close to the mean activity. This form of descriptor filtering intends to increase the quality of the correlation.

SOMFA

- Scaling of fields
 - most active molecule: positive values
 - least active molecule: negative values
- Higher values for most active and inactive compounds
- Less interesting compounds close to mean activity

F2.6.9 HQSAR

In the HQSAR method (Hologram-QSAR, developed by T. Heritage et al.) the structure of a molecule is encoded as a single string of binary numbers. The different molecular substructures are expressed in fingerprints that are then hashed into hologram bins and used as descriptors in QSAR models. A molecular hologram contains all the possible molecular fragments within a molecule and implicitly encodes 3D structural information. The method is somewhat related to the early Free-Wilson approach in which activities are correlated with the presence of various functional groups.

HQSAR

- Structure encoded in binary string
- Fingerprints hashed into hologram bins
- Related to Free-Wilson approach

F2.6.10 GRIND

GRIND (GRid-INdependent descriptors, developed by G. Cruciani et al.) is a method producing descriptors that are calculated from the 3D structure of the molecules, but they do not depend on their position or orientation in the space. In the first step several fields are calculated and transformed (simplified), and in the second step the results are encoded into alignment-independent variables based on auto-correlation theory. The results can be visualized with 2D or 3D diagrams.

GRIND

- Descriptors derived from 3D structures
- Invariant with respect of coordinate system
- Usage of auto-correlation theory

F2.6.11 Quasar

The Quasar approach (quasi-atomistic SAR, developed by A. Vedani et al.) is based on the construction of an envelope of pseudo-atoms that represent the surface of a putative binding site. Several fields are calculated for each of these pseudo-atoms; they include hydrophobic neutral, salt bridge positive, salt bridge negative, H-bond donor, H-bond-acceptor, hydrophobic positive, hydrophobic negative, H-bond flip/flop, solvent fields. Induced fit of the receptor to the different compounds is taken into account by allowing the mean envelope to adapt its shape to the individual molecules.

Quasar

- Envelope of pseudo-atoms
- Fields calculated for all pseudo-atoms
- Many fields are calculated
- Induced fit of receptor taken into consideration

F2.6.12 CoMASA

CoMASA (comparative molecular active site analysis, developed by T. Kotani et al.) is a method replacing the 3D regular grid points of the lattice by a small number of "representative points" selected by cluster analysis (tree-variable multivariate). No alignment of the molecules is needed and the method is very rapid (e.g. for the 31 steroid benchmark, only 92 representative points are sufficient, instead of the 7200 grid points normally used in CoMSIA).

CoMASA

- Regular lattice replaced by small number of "representative points" generated by cluster analyses
- Alignment of molecules not necessary
- Very rapid

F2.6.13 WeP

WeP (weighted probes, developed by W. Shin et al.) is based on the idea that certain regions of the receptor surface contribute, to varying extents, to the differences in the activities of the ligands, whereas other regions do not. The probes, placed around the surface of a superimposed set of ligands, are associated with fractional weights, which are then optimized with a genetic algorithm. A pseudo receptor is generated, which consists of the surviving probes with nonzero weight values.

WeP

- Fractional weights associated to probes
- Weights optimized by genetic algorithm
- Pseudo-receptor is generated



The topic Conclusion contains the following page:

- Conclusion

F2.7.1 Conclusion

3D-QSAR proves to be very useful for the optimization of a reference chemical structure by improving the binding affinities with an unknown receptor. The method generates data of great practical value such as correlation functions and 3D visualizations of favorable and unfavorable interactions with the receptor. 3D-QSAR is a tool that holds a great potential and is expanding rapidly towards a broader field of application.

