INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Fast, Efficient Generation of High-Quality Atomic Charges

Araz Jakalian

A Thesis

in

The Department

of

Chemistry and Biochemistry

Presented in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy at
Concordia University
Montréal, Québec, Canada

July 2000

ABSTRACT

Fast, Efficient Generation of High-Quality Atomic Charges

Araz Jakalian, Ph.D.
Concordia University, 2000

The novel AM1-BCC charge model quickly and efficiently generates high-quality atomic charges for organic molecules suitable for computer simulations in the solution phase. The concept of the AM1-BCC charge model is to produce atomic charges that emulate the HF/6-31G* electrostatic potential (ESP). Underlying electronic structure features including formal charge and electron density delocalization are first captured by AM1 atomic charges; bond charge corrections (BCCs) are then simply added to these AM1 atomic charges to produce the AM1-BCC charges. The BCCs have been determined such that, when added to the AM1 atomic charges, the resulting AM1-BCC atomic charges emulate the HF/6-31G* ESP. The BCCs were parameterized against a training set of >2700 molecules sampling most organic functional groups and their combinations, as well as an extensive variety of cyclic and fused bicyclic heteroaryl systems. The resulting BCC parameters allow the AM1-BCC charging scheme to handle virtually all organic compounds in The Merck Index and the NCI Database. The AM1-BCC charge model reproduces *ab initio* dimer energies of a diverse set of molecules with an average error of 0.9 kcal/mol and experimental relative free energies of solvation of a diverse set of compounds with an average error 0.7 kcal/mol.

Acknowledgments

I would like to acknowledge the trust, support, and guidance of Christopher Bayly and David Jack throughout my learning process of scientific research.

I would like to thank Christine and Lara for their patience, my parents Ashod and Chaké for the confidence they instill in me, and my brothers Sarmen and Chant for the strength they give me.

I also wish to acknowledge the financial support of the National Science and Engineering Council (NSERC) of Canada and Merck Frosst Canada Inc. for an Industrial Postgraduate Scholarship (IPS) as well as Concordia University for an External Grant Holder Doctoral Scholarship.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

SPECIAL SYMBOLS

AM1	the AM1 semi-empirical method
MNDO	the MNDO semi-empirical method
PM3	the PM3 semi-empirical method
BCC	bond charge correction
AM1-BCC	the AM1-BCC charge model
HF	Hartree Fock
6-31G*	the 6-31G* basis set
ESP	electrostatic potential
NCI	National Cancer Institute
Ψ	molecular wavefunction
ψ	molecular orbital
$c_{\mu i}$	molecular coefficient
ϕ_{μ}	atomic orbital
STO	Slater-type orbital
GTO	Gaussian-type orbital
SCF	self-consistent-field
MP	Møller-Plesset
K	number of basis functions
NDDO	neglect of diatomic differential overlap
FF	force field
MM	molecular mechanics
DNA	deoxyribonucleic acid

AMBER	assisted model building with energy refinement
q_i	atomic charge on center i
K_r	force constant for bond
K_θ	force constant for angle
r_{eq}	equilibrium bond length
θ_{eq}	equilibrium angle
E_{total}	total force field energy
E_{bond}	force field bond energy
E_{angle}	force field angle energy
$E_{dihedral}$	force field dihedral energy
E_{vdW}	force field vdW energy
E_{elec}	force field electrostatic energy
V_n	magnitude of the torsion energy barrier
A_{ij}	van der Waals repulsion term between atoms i and j
B_{ij}	van der Waals dispersion term between atoms i and j
R_{ij}	distance between atoms i and j
ε_{ij}	Lennard-Jones well depth for interaction between atoms i and j
R_{ij}^{*}	Lennard-Jones radius for interaction between atoms i and j
ε	dielectric constant
$V(\mathbf{r})$	ESP at point $\mathbf{r}$

Z_i	nuclear charge on atom I
$\mathbf{R}_i$	and position of nucleus I
$\rho(\mathbf{r}')$	electron density at position $\mathbf{r}'$
$P_{\mu\nu}$	density matrix elements obtained from the wavefunction
EM	energy minimization
MD	molecular dynamics
fs	femtosecond
K	kelvin
NVT	constant number of particles, volume, and temperature
ΔG	Gibbs free energy difference between two states
R	gas constant
T	temperature
ΔH	difference between two Hamiltonians
λ	changing parameter in free energy calculations
QM	quantum mechanical
SPC, TIP3P, TIP4P	rigid, non-polarizable water models
PPC	polarizable point-charge model
BCI	bond charge increment
MMFF	molecular mechanics force field
CM1, CM2	charge model 1 and charge model 2
Q^{AM1}	AM1 atomic charges on a molecule
Q^{BCC}	correction term that adjusts the AM1 charges to produce the AM1-BCC charges

$Q^{AM1-BCC}$	AM1-BCC atomic charges on a molecule
q_j^{AM1}	AM1 atomic charge on atom j
q_j^{BCC}	net correction applied to the AM1 charge on atom j
$q_j^{AM1-BCC}$	AM1-BCC charge on atom j
EHT	extended Hückel theory
B_α	bond charge
T	the bond connectivity template matrix
NMA	N-methyl acetamide
V_l^{calc}	the calculated electrostatic potential
V_l^{diff}	difference between the QM and calculated electrostatic potentials
V_l^{corr}	ESP generated by the bond charge corrections
χ^2	fitting objective function
V_l^{QM}	the QM electrostatic potential
RESP	restrained electrostatic potential
RMS	root-mean square
μ	dipole moment
D	debye
TS1, **TS2**, **TS3**	training sets 1, 2, and 3, respectively
ps	primary scaffold
ss	secondary scaffold
pf	primary functional group

sf	secondary functional group
ACC	atomic charge compensation
PATTY	programmable atom-typer
MOPAC	program for semi-empirical calculations
GAUSSIAN	program for molecular orbital calculations
RFES	relative free energies of solvation
GAMESS	program for molecular orbital calculations
PMF	potential of mean force
CC	correlation coefficient

1 Introduction

1.1 Chemistry and Computation

As in any discipline that uses modeling to solve problems, from meteorology to the study of evolution, the goal remains the same: to design a mathematical model, i.e. an equation with adjustable parameters, that will firstly reproduce the known results, secondly, predict results yet unknown, and thirdly, provide insight into the problem which otherwise would be difficult to attain.

As chemists we often build and use models to predict the outcome of an organic reaction, a molecular spectrum, or a binding constant even before an experiment has begun (e.g. Hammett constants, Hückel theory, Born solvation model, etc.). We use theory, experience, and rule-of-thumb approaches to predict chemical phenomena. Ultimately the experiment is physically performed and the results are compared to the model.

There is a field of chemistry that has evolved from these simple predicting methods into an accurate prediction and understanding tool. Using computers as the main instrument, the field of *computational chemistry* uses models at various levels of sophistication or approximation (e.g. *ab initio*, semi-empirical, and molecular mechanics) to predict and gain insight into chemical phenomena.

Ab initio methods use the fewest approximations to model the electronic structure of molecules and produce reliable electronic properties, but they are computer resource intensive and consequently are limited to modeling small molecules. Semi-empirical methods are less reliable than *ab initio* methods but require only a fraction of the computer resources. These two methods require only the nuclear coordinates of the atoms, and the net charge and the multiplicity of the molecule to solve the wavefunction of the molecule. Properties such as molecular geometry, charge distribution, bonding topology, and bond orders can be extracted from the wavefunction. Molecular mechanics (MM) however, uses the most approximations and therefore can model large systems (e.g. proteins, and DNA strands in solution). Unlike, the *ab initio* and semi-empirical methods, MM does not solve the wavefunction of the molecule and requires, besides the nuclear coordinates, knowledge of the charge distribution, the bonding topology, and bond orders of the molecule to calculate an energy and often times the interaction parameters are obtained from *ab initio* or semi-empirical calculations.

These energy-based methods (described in more detail in the next sections) will be used in this work to model the electrostatic interactions of organic molecules in solution. Electrostatic interactions play an important role in drug-receptor binding, host-guest complexations, and solvation. Of particular interest to this work is the solution phase electrostatic potential (ESP) of organic molecules. This is an important property because as two molecules approach, e.g. a ligand and protein, they are "seen" and interact by their ESPs. The solution phase ESP differs from the gas phase ESP because it is polarized due to non-covalent "pushing" and "pulling" of electron clouds by short-range polar

interactions (e.g. hydrogen-bonding). Therefore, to properly model solution phase phenomena of organic molecules, an accurate description of the polarized ESPs is necessary. The ESP can be obtained from the electron density, an observable property that can be obtained from X-ray diffraction techniques, but for our purposes it is more convenient to calculate the ESP from *ab initio* methods.

The goal of this work is to emulate this *ab initio* ESP in a fast and efficient manner using point charges centered on the atoms in a molecule for use in molecular mechanics modeling of organic molecules in solution.

1.1.1 Ab initio Methods

Ab initio molecular orbital methods[1] seek an approximate solution to the Schrödinger equation:

$$H\psi = E\psi \tag{1.1}$$

where H is the Hamiltonian that operates on the wavefunction ψ, and E is the energy of the system. The square of the wavefunction multiplied by a small volume gives the probability of finding the electron in that volume. This is an observable quantity directly related to the electron density as observed through X-ray diffraction techniques. In fact, eq. 1.1 is a *set* of equations where ψ_n and E_n are the wavefunctions (or eigenfunctions) and energies (or eigenvalues) corresponding to the quantum energy level n. For example, the wavefunctions of the hydrogen atom are the functions that describe the 1s, 2s, 2p, etc. orbitals. The Hamiltonian operator represents the potential and kinetic energy terms

(kinetic energy of the nucleus and the electron; and the potential energy contributions due to nucleus-electron, nucleus-nucleus, and electron-electron interactions). However, the Schrödinger equation is too complicated to solve except for the hydrogen atom, He^{+1} ion, Li^{+2} ion, etc. In order to simplify the Schrödinger equation and facilitate a solution, the Born-Oppenheimer approximation is used, i.e. the nucleus may be regarded as fixed in space since the motion of the electron is much faster than the motion of the nucleus and therefore the description of the two can be separated. This simplifies the Hamiltonian by removing the nuclear kinetic energy term and making the nuclear-electron potential energy term a constant. However, even with this approximation, eq. 1.1 still cannot be solved for multielectron atoms, therefore two more approximations are used: 1) the Hartree-Fock approximation and 2) the linear combination of atomic orbitals (LCAO) approximation.

The Hartree-Fock (HF) approximation replaces the multi-electron wavefunction by the product of one-electron functions. Each electron is treated separately and is regarded as moving under the influence of an average potential energy representing the nuclei and the other electrons. The wavefunction is written as a single determinant with one-electron functions as elements. The one-electron functions are called spin orbitals because they are the product of a space component, called molecular orbitals, and a spin component (α or β). The single determinant of spin-orbitals guarantees that Pauli's exclusion principle is satisfied. To solve the wavefunction, one must solve for the spin-orbitals. However, the molecular orbitals are made to be linear combinations of *known*

one-electron functions (i.e. atomic orbitals) or linear combination of atomic orbitals (LCAO):

$$\psi_i = \sum_{\mu=1}^{K} c_{\mu i}\phi_\mu \qquad (1.2)$$

where ψ_i is the i'th molecular orbital, $c_{\mu i}$ are the molecular coefficients, ϕ_μ are the atomic orbitals (also called basis functions), and the summation runs over the K atomic orbitals which comprise the *basis set*. That is, a basis set consists of the set of basis functions that describe the electrons in an atom. Slater-type orbitals (STOs) can be used as basis functions since they contain a spherical component that enables them to embed the angular properties from the solution of the hydrogen atom wavefunction (e.g. s, p, d, etc.). However, STOs are not suitable for fast calculations and thus their use has been limited. Gaussian-type orbitals (GTOs) which contain a cartesian component (not spherical) are used to approximate the shape of STO functions by summing a number of gaussian functions (also called gaussian primitives) with different exponents and coefficients. In order to reduce computation time, linear combinations of gaussian primitives (called contractions) can be used as a basis function. Several introductory texts[1, 2] and reviews[3, 4] on basis sets are available, but for our purposes, each inner-shell atomic orbital will be described by six gaussian primitives contracted to one (e.g. the 1s orbital of carbon will be described by six s-type gaussian primitives contracted to one). Each valence-shell atomic orbital will be described by two basis functions, one containing three gaussian primitives and the other containing one uncontracted gaussian

primitive (e.g. the $2p_x$ orbital of carbon will be described by four p-type gaussian primitives of which three are contracted into one and the other is uncontracted). This can be summarized by writing the above basis set for an atom in Pople's notation: 6-31G. Addition of functions with higher angular quantum numbers (i.e. d-type polarization functions) on heavy centers (i.e. non-hydrogens) can be written as 6-31G(d) (or 6-31G*), while d-type polarization functions on heavy centers plus p-type polarization functions on hydrogens is written as 6-31G(d,p) (or 6-31G**). For example, the 6-31G* basis set for the carbon atom will have one basis function for the 1s orbital, two basis functions each for the 2s, $2p_x$, $2p_y$, and $2p_z$ orbitals, and six d-type basis functions for polarization.

To solve the wavefunction, one must now find the optimum set of molecular coefficients, $c_{\mu i}$, that minimize the energy according to the variational theorem, i.e.

$$\frac{\partial E}{\partial c_{\mu i}} = 0. \tag{1.3}$$

However, E cannot be calculated until the $c_{\mu i}$ coefficients have been determined, subsequently this problem cannot be solved in closed form. Therefore, an iterative approach is taken where a starting set of $c_{\mu i}$ coefficients is guessed and the energy is minimized until a convergence criterion is reached. This is called the self-consistent-field (SCF) approach.

The description of the wavefunction as a single determinant is an important deficiency with HF theory, i.e. the HF wavefunction represents only a single electronic configuration of the molecule. Consequently, a description of the correlation between the motions of electrons cannot be accounted for and hence, HF models overestimate electron-electron repulsion. A solution is to build a wavefunction that incorporates allowed excited states of electronic configurations (each represented by a single determinant) with the ground state. This is often done with Møller-Plesset (MP) perturbation theory where the electronic Hamiltonian is described by a perturbation of the HF wavefunction. Truncation of the perturbation equation after the second order energy term is referred to as MP2, after the third order energy term is referred to as MP3, and so on.

Because *ab initio* methods require a large number of calculations, proportional to K^4 where K is the number of basis functions, they are confined to the treatment of small molecules. However, they provide a very good description of electronic properties such as dipole moments, electron densities, electrostatic potentials, etc.

1.1.2 Semi-empirical

Semi-empirical molecular orbital methods such as MNDO[5, 6], AM1[7], and PM3[8, 9] also solve the Schrödinger equation but with many more approximations; therefore the resulting wavefunction is considered to be less accurate than *ab initio* methods. Firstly, only valence electrons are considered in the calculations, the inner-shell electrons are grouped into a frozen core. Secondly, the basis functions (i.e. the atomic orbitals) are restricted to s-type and p-type, e.g. 1s, 2s, $2p_x$, $2p_y$, $2p_z$. Thirdly, the overlap integral of

basis functions located on different atoms is set to zero. This last approximation, called neglect of diatomic differential overlap (NDDO)[10], is the most important in reducing computations from K^4 to K^2, where K is the number of basis functions. Compared to *ab initio* methods, larger molecules can be treated. Although semi-empirical methods are not usually as accurate as *ab initio* methods for electronic properties, they do however provide reliable results for heats of formation and generally capture the essence of the electronic structure of a molecule in a fraction of the time.

1.1.3 Molecular Mechanics

Molecular mechanics (MM) methods[11], often called the ball-and-spring model, describe a molecule as a mechanical or classical object that obeys the Newtonian laws of motion. Electrons are not explicitly included in the description of the molecule and thus, atoms are treated as balls and bonds as springs, electronic effects such as repulsion, dispersion, and electrostatics are approximated by simple functions that are parameterized to reproduce experimental or *ab initio* quantities. These approximations allow MM methods to model systems with thousands of atoms e.g. DNA strands and proteins surrounded by water molecules. The forces acting on the nuclei are approximated by the average potential field of the electrons. Different MM models approximate this "force field" (FF) in different ways. The total FF energy of a molecular system is given by the sum of the individual energy components, e.g.:

$$E_{total} = E_{bond} + E_{angle} + E_{torsion} + E_{vdW} + E_{elec} \tag{1.4}$$

where E_{total} is the total energy of the system and individual terms represent bond, angle, torsion, van der Waals (vdW), and electrostatic energy terms. However, eq. 1.4 is considered minimalistic in that these are the minimum energy terms needed to describe conformational energies and intermolecular interactions of organic molecules. The functional form of eq. 1.4 is described in more detail in the next section.

1.2 Components of Force Fields

Force fields (FFs) have been developed for small molecules[12], proteins and DNA[13], inorganic molecules[14], etc. FFs are mathematical models with adjustable parameters. Organic/biological-based molecular modeling FFs, e.g. that of Cornell *et al.*[15] (implemented in the AMBER5.0[16] program), commonly use an effective two-body additive potential energy function, as shown in eq. 1.5 (below), to describe structural features of a molecule.

$$E_{total} = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left[1 + \cos\left(n\phi - \gamma\right)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \tag{1.5}$$

The many-body non-covalent interactions in eq. 1.5 are truncated at the two-body level with the three-body, four-body, etc. non-covalent interaction terms embedded implicitly in the parameters of the potential energy function, therefore eq. 1.5 is termed an "effective two-body additive potential energy function". Implicit polarization can be embedded in eq. 1.5 through the charge parameters q_i. The parameters in eq. 1.5 (e.g.

K_r, r_{eq}, K_θ, etc.) are determined by fitting them to experimental or high-level *ab initio* data. The energy terms in eq. 1.5 account for bond vibrations, angle bending, torsion twisting, electron-electron repulsion, dispersion, and electrostatics, respectively. A graphical representation of these terms is shown in Figure 1.1.

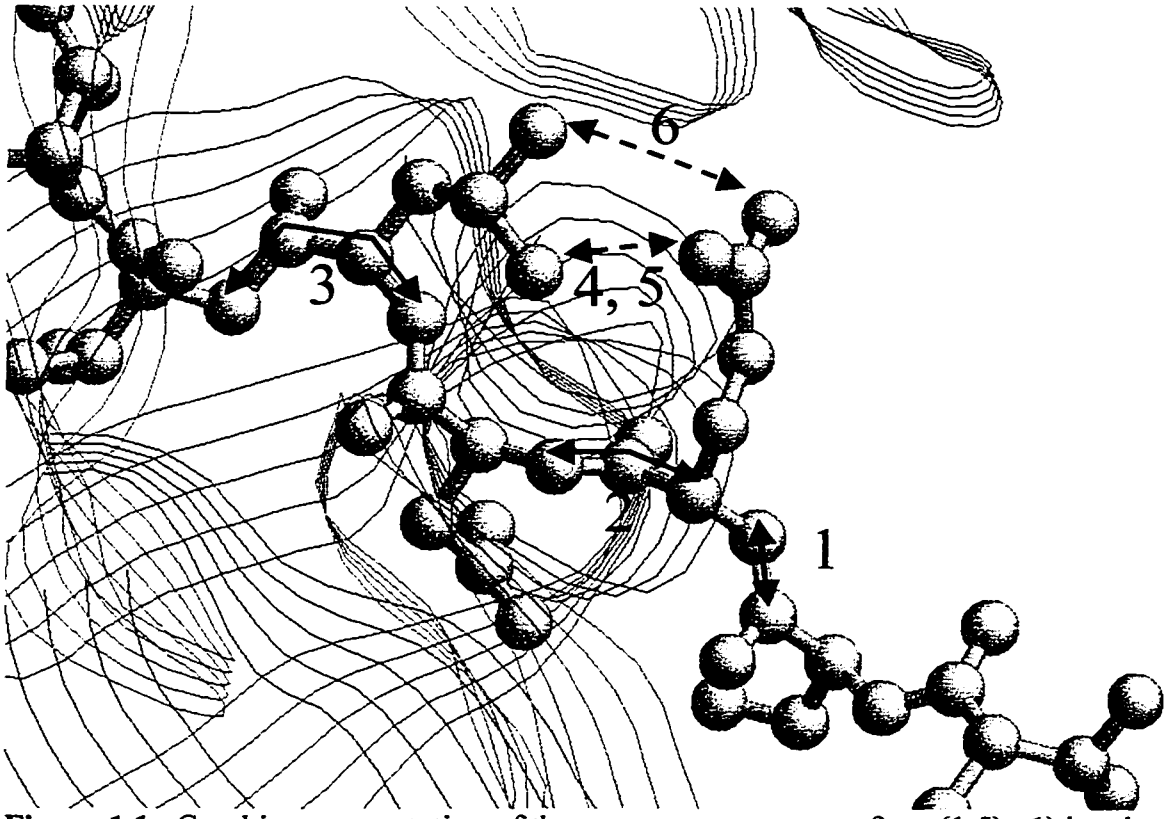


Figure 1.1. Graphic representation of the energy components of eq. (1.5). **1**) bond vibrations, **2**) angle bending, **3**) torsion twisting, **4**) electron-electron repulsion, **5**) dispersion, **6**) electrostatics.

1.2.1 Covalent Terms

The first three terms in eq. 1.5 describe bond vibrations, angle bending, and torsion twisting, respectively. For bond vibrations (Figure 1.1-**1**), Hooke's law provides

a good description of the stretching and compressing of bonds near the equilibrium bond length, where the total bond energy of a system is given by:

$$E_{bond} = \sum_{bonds} K_r (r - r_{eq})^2 \tag{1.6}$$

where r_{eq} and r represent an equilibrium bond length and instantaneous bond length, respectively, K_r is the force constant of the bond and the summation runs over all the bonds in the system. Hooke's law is also used to describe angle bending (see Figure 1.1-**2**), where the total angle energy of a system is given by:

$$E_{angle} = \sum_{angle} K_\theta (\theta - \theta_{eq})^2 \tag{1.7}$$

where θ_{eq} and θ represent an equilibrium angle and instantaneous angle, respectively, K_θ is the force constant of the angle and the summation runs over all the angles in the system. The torsion term (see Figure 1.1-**3**) is described by a Fourier series given by:

$$E_{torsion} = \sum_{torsion} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \tag{1.8}$$

where V_n is the magnitude of the energy barrier of the torsion, γ is the phase offset, n is the periodicity of the torsion, ϕ is the actual angle, and the summation runs over all torsions in the system. Standard texts on these topics are available.[17]

FFs that aim to accurately reproduce vibrational spectra often incorporate anharmonic effects through inclusion of cross-terms such as bond-bond, bond-angle, etc.

1.2.2 Non-Covalent Terms

Non-covalent forces govern molecular interactions and determine e.g. host-guest and ligand-protein complexations. A proper description of these forces requires electron-electron repulsion (the Pauli exclusion principle), dispersion (London[18] forces), and electrostatic interaction terms. The first two are usually grouped as the van der Waals interactions (see Figure 1.1-**4** and 1.1-**5**) and are approximated by a Lennard-Jones potential:

$$E_{vdW} = \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right] \tag{1.9}$$

where R_{ij} represents the distance between non-bonded atoms i and j, and A_{ij} and B_{ij} are the repulsion and dispersion coefficients, respectively and are given by:

$$A_{ij} = \varepsilon_{ij}(R_{ij}^{\cdot})^{12} \tag{1.10}$$

and

$$B_{ij} = 2\varepsilon_{ij}(R_{ij}^{\cdot})^6 \tag{1.11}$$

where the well depth ε_{ij} is given by

$$\varepsilon_{ij} = \sqrt{(\varepsilon_i \varepsilon_j)} \tag{1.12}$$

where ε_i and ε_j are the van der Waals well depths for atoms i and j (in kcal/mol) and the equilibrium distance R_{ij}' is given by

$$R_{ij}^{\cdot} = R_i + R_j \tag{1.13}$$

where R_i and R_j are the van der Waals radii of atoms i and j (in units of Å). Jorgensen's group has determined these parameters for atoms in organic molecules in condensed phase systems by adjusting them to reproduce solution phase properties of liquids such as densities and heats of vaporization[19]. Fortunately, these parameters have proven to be transferable to most organic molecules and thus far, do not need to be recalculated.

A point charge model, usually centered on the atoms, approximates the electrostatic contribution to non-covalent interactions:

$$E_{elec} = \sum_{i<j} \frac{q_i q_j}{\varepsilon R_{ij}} \tag{1.14}$$

where q_i and q_j are the atomic charges located on atoms i and j (there must be at least three bonds separating the atoms if they reside on the same molecule), ε is the dielectric constant of the medium, R_{ij} is the distance between atoms i and j, and the summation runs over all eligible pairs of atoms. These "atomic-charges" are constructed to mimic certain properties of the continuous electron distribution of molecules, e.g. the electrostatic potential (ESP)[20], the molecular charge density[21], or the dipole and higher electric multipole moments[22]. The ESP of a molecule is often chosen for this purpose because it is a quantum mechanical property and is partially derived from the electron density, an experimentally observable quantity. Also, as two molecules approach, e.g. a ligand and protein, they are "seen" and interact by their electrostatic potentials. Furthermore, it has been shown in the literature that atomic charges that reproduce the ESP perform well in reproducing relative free energies of solvation of organic molecules.[23,24] Unlike the vdW parameters, atomic charges have not proven to be transferable and unfortunately, each molecule requires its own set of atomic charges.

1.3 The Electrostatic Potential (ESP)

The net electric potential "felt" by a positive probe at a point **r** around a molecule, due to the positive nuclear and negative electric contributions, is called the electrostatic potential (ESP) and is given by:

$$V(\mathbf{r}) = \sum_{i=1}^{N} \frac{Z_i}{|\mathbf{r} - \mathbf{R}_i|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \tag{1.15}$$

where $V(\mathbf{r})$ is the ESP at point $\mathbf{r}$, Z_i and $\mathbf{R}_i$ are the nuclear charge and position of nucleus i, respectively, $\rho(\mathbf{r}')$ is the electron density at position $\mathbf{r}'$, and N is the number of atoms in the molecule. Although the electron density is an experimentally observable quantity, by X-ray diffraction techniques, for practical purposes it is extracted from an *ab initio* wavefunction (e.g. HF/6-31G*) and can be written as:

$$V(\mathbf{r}) = \sum_{i=1}^{N} \frac{Z_i}{|\mathbf{r} - \mathbf{R}_i|} - \sum_{\mu}^{K} \sum_{\nu}^{K} P_{\mu\nu} \int \frac{\phi_\mu \phi_\nu}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \tag{1.16}$$

where $P_{\mu\nu}$ are the density matrix elements obtained from the wavefunction (see Appendix A), and ϕ_μ and ϕ_ν are the set of K basis functions used.

The ESP of a molecule can also be calculated by a classical approach by summing the point charge contributions of all the atoms in the molecule:

$$V(\mathbf{r}) = \sum_{i=1}^{N} \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|} \tag{1.17}$$

where q_i and $\mathbf{r}_i$ are the partial charge and position of atom i.

1.4 Application of Force Fields

1.4.1 Energy Minimization & Molecular Dynamics

A FF can be combined with energy minimization methods to search the configurational space of molecules, with molecular dynamics (MD) techniques to sample phase space[25] (i.e. momentum and coordinate spaces) or with Monte Carlo techniques to sample configurational space.

Energy minimization (EM) involves finding *the* set of variables (e.g. the x, y, and z coordinates of the atoms in the system) that minimize the potential energy function (e.g. eq. 1.5). EM techniques do not include time or temperature since they simply locate minima over a multidimensional (e.g. x, y, and z of each atom) surface. In this work, EM techniques are used to i) remove close contacts between atoms before a MD run is performed so that high potential energy volumes "hot-spots" are removed and ii) find the optimum geometry of hydrogen-bonded dimer systems.

Molecular dynamics aims to mimic "reality" by using Newton's second law of motion ($F = ma$) to propagate the coordinates of atoms through time. Initially, the force on each atom is calculated by taking the negative of the derivative of the potential energy function (e.g. eq. 1.5) with respect to the coordinates. Knowing the mass of each atom, the acceleration of each atom can be calculated. Given a time step (for total times usually on the order of 10^{-15} seconds, i.e. 1 fs), the new set of coordinates of each atom is

calculated and the process is repeated thousands of times (usually in the range of picoseconds to nanoseconds). The temperature of the system can be controlled through the kinetic energy term. The propagation of the coordinates through time, i.e. the trajectory, can be saved, replayed, and analyzed. Also, thermodynamic properties can be accumulated along the trajectory as ensemble averages and related to macroscopic properties, e.g. internal energy, heat capacity, pressure, etc. In this work, MD is used to i) dissipate "hot-spots" by running a simulation at a temperature of 1K; ii) raise the temperature of a system from 1K to 300K; and iii) sample phase space during a free energy perturbation calculation.

1.4.2 Free Energy Calculations

Predicting the outcome of a chemical process, such as the hydration of a molecule or the binding of a ligand to a protein, requires calculating the free energy of that process. The free energy is considered to be a very important thermodynamic quantity because it tells us whether a chemical process will occur (i.e. will the molecule be solvated? Will the ligand bind to the protein? How tightly?).

Under conditions of constant number of particles, volume, and temperature (NVT), the free energy is expressed as the Helmholtz function while under conditions of constant number of particles, pressure, and temperature, the free energy is given by the Gibbs function. The statistical mechanics definition of free energy is given by the Boltzmann weighted partition function. The partition function is expressed as the integral of the configurational energy in phase space (i.e. coordinate and momentum

space). Therefore, to obtain the free energy one must calculate the partition function; to calculate the partition function, one must accumulate an ensemble of configurations over phase space. Unfortunately, this is a difficult quantity to calculate because an analytical function can only be written for the simplest of systems. Consequently, molecular dynamics simulations are used to accumulate the free energy as a time average, assuming the ergodic hypothesis[26], over the trajectory of a simulation by sampling the configurational space. However, it is much easier to calculate a free energy difference between two *similar* systems A and B (e.g. methanol and methanthiol) rather that an absolute free energy because regions of phase space of both systems will cancel. Rigorous free energy perturbation calculations are based on the following equation:

$$\Delta G = G_B - G_A = -RT \ln\left\langle e^{-\Delta H/RT} \right\rangle_A \tag{1.18}$$

where $\Delta H = H_B - H_A$ and $\langle\ \rangle_A$ signifies an ensemble average over system A. H_A and H_B are the Hamiltonians of systems A and B, respectively. It has been shown in the literature that for similar systems the kinetic energy term of the Hamiltonians can be ignored[27] and hence H_A and H_B can be approximated as the force field energies E_A and E_B (see eq. 1.5). Equation. 1.18 can be implemented by "mutating" system A into system B in the gas phase and in solution; i.e. changing the FF parameters that describe methanol into the FF parameters that describe methanthiol (see Figure 1.2). Since free energy is a state function, the relative free energy of solvation between system A and B can be calculated as $\Delta\Delta G = \Delta G_{sol}(B) - \Delta G_{sol}(A) = \Delta G_2 - \Delta G_1$.

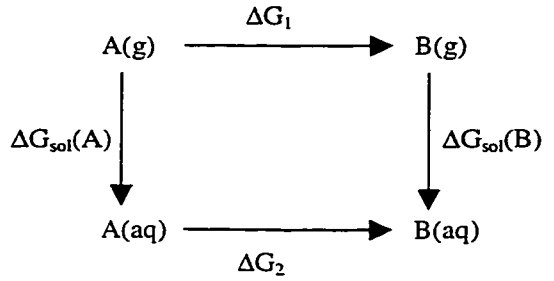


Figure 1.2. Thermodynamic cycle for calculating relative free energies of solvation. System *A* (e.g. methanol) is "mutated" into system *B* (e.g. methanthiol) in the gas phase and in solution, i.e. ΔG_1 and ΔG_2, respectively. The relative free energy of solvation between system *A* and *B* can be calculated as $\Delta\Delta G = \Delta G_{sol}(B) - \Delta G_{sol}(A) = \Delta G_2 - \Delta G_1$.

The changing parameters are coupled to a variable λ that can vary from 0 ($H=H_A$) to 1 ($H=H_B$) in small enough increments (windows) in order to ensure that the phase spaces of both systems overlap. Therefore the Hamiltonian as a function of λ can be written as:

$$H(\lambda) = \lambda H_B + (1-\lambda) H_A \tag{1.19}$$

Equation 1.18 now becomes:

$$\Delta G = G_B - G_A = \sum_{\lambda=0}^{\lambda=1} - RT \ln\left\langle e^{-\Delta H/RT} \right\rangle_{\lambda} \tag{1.20}$$

where $\Delta H' = H_{\lambda+\Delta\lambda} - H_{\lambda}$. This is called the *free energy perturbation* method. An alternative to free energy perturbation is the *slow growth* method where, at each step of the simulation, the Hamiltonian changes by an infinitesimal amount:

$$\Delta G = \sum_{\lambda=0}^{\lambda=0} (H_{n+1} - H_n) \tag{1.21}$$

where, for a given λ, H_n is the Hamiltonian at step n, and H_{n+1} is the Hamiltonian at the next step. Thermodynamic integration is yet another alternative to rigorous free energy perturbation calculations.[28] An introduction and review on free energy calculations is available.[28, 29]

1.5 Review of Atomic Charges

In this section atomic charges are introduced and various models are reviewed. A simple Coulombic electrostatic model is commonly used in FFs to calculate the electrostatic energy between point charges. These point charges are usually located at the center of the atom and are therefore referred to as "atomic charges". The electrostatic energy between two atomic charges can be calculated using eq. 1.14. An atomic charge, in effect, places onto the atom's nucleus the amount of charge associated with that atom in a molecule according to some charge partitioning scheme. Although "atomic charge" is a useful concept it is not a quantum mechanical (QM) property in that an atomic charge QM operator does not exist. Electrostatic potential-fit charges from the HF/6-31G* wavefunction (i.e. atomic charges that emulate the HF/6-31G* ESP) have been shown to reproduce relative free energies of solvation of organic molecules[24] and are therefore

adequate for simulations of solution phase systems. This wavefunction overestimates the polarity of molecules by approximately 10 to 15%[30] which offers a fortuitous "implicit" polarization to compensate for the fact that a two-body additive FF, by construction, does not include polarization. Since the polarity of widely used water models, e.g. SPC[31], TIP3P[32] and TIP4P[32], is also overestimated, solute molecules solvated by these water models must include implicit polarization in order to maintain a proper electrostatic balance between the solute and solvent molecules.[15] These water models are rigid and nonpolarizable and therefore the effects of electronic polarization are not explicitly included. The polarizable point-charge (PPC) model[33] for water overcomes this problem by allowing the three atom-centered point charges to be dependent on the electric field. The point-charges were parameterized from *ab initio* results of the water molecule in an applied electric field. Structural, thermodynamic, dielectric properties, and the self-diffusion coefficient at temperatures between 263 to 573 K are in good agreement with experimental results. There is no unique way to assign atomic charges; various methods have been developed and are reviewed in the following paragraphs.

The electron density analysis method developed by Bader[34], called "atoms in molecules"[35] assigns atomic charges by first defining a volume for each atom and then numerically integrating the electron density within that volume. This partitioning of electron density is based on *gradient paths*, which are the steepest ascents in the electron density of a molecule. These gradient paths terminate at locations called *critical points* usually located at atomic centers or bond centers, i.e. where the electron density is minimal between the bonded atoms. Starting at critical points and following the gradient

paths results in a three-dimensional partitioning of the molecular electron density. It was found that the resulting atomic charges were relatively independent of the basis set used and were in agreement with experimentally determined bond dipoles of methane and ethyne.[36]

The Merz-Singh-Kollman[37, 38] and CHELPG[39] ESP-fitting schemes fit atomic charges to reproduce the QM ESP calculated at grid points located around a molecule. These schemes differ only in the number and placement of the grid points around the molecule and therefore generally produce similar atomic charges. Although these methods produce charges suitable for condensed phase simulations, they are subject to the problems inherent in current ESP-fit methods: i) charges are conformer dependent[40]; ii) charges on and near buried atoms (e.g. methyl carbon atoms) are numerically unstable, i.e. their magnitude can vary widely while barely perturbing the quality of the fit[41, 42]; iii) large molecules must be treated as a superposition of fragments due to computer resource limitations; and iv) charges are not easily transferable between common functional groups in related molecules. In addition, the costs of the CPU intensive calculation of the ESP at the *ab initio* level pose a significant barrier to the routine use of ESP charges, especially in pharmaceutical applications where high throughput has become a major issue. These shortcomings serve as the motivation for this work.

Several methods have been developed to reduce ESP-fit problems. For example, Reynolds *et al.*[30b] found that the conformational dependencies of ESP-fit charges can be decreased by simultaneously fitting charges to the ESP of multiple conformations of a

molecule. Bayly *et al.*[41] found that numerical instabilities in the fitting process can be further diminished by restraining unstable charges to lower magnitudes (the RESP model[41]), which also reduces the transferability problem. Imposing symmetry on equivalent atomic centers during the fitting process improves the quality-of-fit relative to simply averaging those charges, but the error thus introduced into the polar areas of the electrostatic potential significantly decreases the performance of these charges in condensed phase simulations.[41] Francl *et al.*[42] used singular value decomposition to determine the rank of the least-squares matrix, selecting a subset of atoms for which statistically valid charges can be assigned based on that rank estimate and then refitting the rest. This method assumes that instabilities are associated with particular atoms. Although these tactics address some problems in direct ESP-fitting of atomic charges, the human effort increases: e.g. multiple conformations of a molecule must be constructed and the ESP calculated for each one; molecules must be treated individually to determine buried centers and assign atomic equivalencies.

Scaling charges from semi-empirical ESPs to mimic the HF/6-31G* ESP-fit charges have been proposed by Besler *et al.*[37] and Alemán *et al.*[43] These methods take advantage of the greater speed and capacity of semi-empirical calculations to treat even large systems quickly and efficiently. The most obvious limitation, however, is that charged molecules become scaled to have non-integral charges. Furthermore, these methods were only developed by fitting to the numerically unstable ESP-fit charges; the resulting ESPs generated by these scaled charges have not been validated by direct comparison to the *ab initio* ESPs.

The bond charge increment (BCI) approach of the popular MMFF[44] model is bond topology based, i.e. atomic charges are assigned according to bond-types, and consequently can generate atomic charges from pre-calculated BCI parameters without even a semi-empirical calculation (see Figure 1.3). BCIs are classified according to the MMFF[44] atom-types forming the bond and fitted statistically to ESP-fit dipole moments, scaled QM interaction energies, and hydrogen bond geometries.[45] Summing bond charge increments from connecting atoms that reflect the polarity of each bond-type generates charge on an atom, while maintaining total molecular charge.

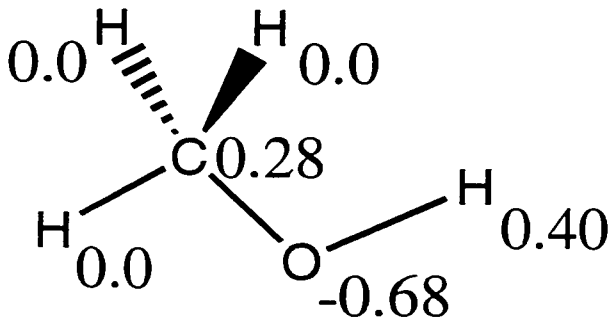


Figure 1.3. a) molecule of methanol. The MMFF charges are shown next to each atom and b) the bond charge increments (BCIs) needed to charge methanol. For each atom in the molecule, the BCIs are added from all the bonded neighbours. The subscripts refer to the bond-types composed of atom-type *I*, the bond order, and atom-type *J*.

The resulting BCIs can then be transferred to any desired compound. A total of 82 atom-types[44], a relatively large number, is required by MMFF to allow the BCIs to adequately represent the diverse spectrum of organic chemical functionalities. Since the BCI formalism is by construction incapable of introducing a net charge, molecules bearing a

net charge (i.e. ions) pose a problem since the net charge must also be correctly partitioned amongst the various atoms. Keeping the speed, efficiency and transferability of BCIs, Bush *et al.*[46] developed a way to fit BCIs directly to the HF/6-31G* ESPs of a training set of molecules, an approach that forms the background for this work. The authors found that, compared to the direct ESP-fit of individual charges, BCIs greatly improved the stability of the matrix equations for each separate molecule. A further improvement in stability comes from fitting BCIs simultaneously across a training set of compounds, which greatly reduces the degrees of freedom from many charges down to fewer BCIs. However, Bush *et al.*[46] concluded that the MMFF classification, even with a fairly large number of bond-types is not flexible enough to adequately describe π-delocalized systems.

The CM1[47] and more recent CM2[48] charge models (CMs) use semi-empirical charges to obtain "crude" charges that are corrected to reproduce desired properties. Charge-based corrections are partitioned across bonds to correct for deficiencies encountered in the simple atomic population charges derived from a semi-empirical wavefunction. In CM1 and CM2, bond-based corrections depend upon an analysis of the semi-empirical wavefunction to generate bond orders and the parameterization scheme is directed towards reproducing the experimental molecular dipole moments of the molecules in the training set. Jorgensen's group tested the CM1 model, with the PM3 (CM1P) semi-empirical methods, by predicting logP(octanol/water partition coefficient) values from solution phase simulations for a diverse set of organic molecules including 125 drugs and related heterocycles[49] and concluded that:

- CM1A and CM1P yield excellent results for gas phase dipole moments but are inappropriate for solution phase simulations and therefore the charges on neutral molecules were scaled by a factor of 1.3.
- There is a problem with secondary and tertiary amines, which are not predicted to be hydrophobic enough.
- Nitro groups are too hydrophilic.
- Unconjugated amines need corrections.
- Carboxylic acids are predicted to be too hydrophilic.
- ESP-fit charges are preferred over CM1P but are not available for all combinations of functional groups.

Jorgensen's group had previously tested the CM1 model with the AM1 methodology (CM1A) but preferred the CM1P method because of better representation of nitrogen containing functional groups, in particular the amides.[50]

To address the above issues, the AM1-BCC charge model is presented. This model marries complementary features of already existing and readily available methods, namely AM1 atomic charges, the BCI approach, and ESP methodologies. AM1 atomic charges[51] are Coulson-type "population" quantities based directly on the occupancies of the atomic orbitals[52]. They are not meant to reproduce the ESP or even the multipole moments of the subject molecule, therefore they perform poorly in solution phase simulations[38] compared to ESP-derived charges. On the other hand they can be calculated very quickly and they capture underlying features of the electron distribution of a molecule, including net charge and π -delocalization. Bond charge corrections

(BCCs), which have been parameterized using standard least-squares fitting procedures[37] to reproduce the difference between the HF/6-31G* ESP and the ESP generated by the AM1 atomic charges of a training set of molecules, are then added to the AM1 atomic charges in the same fashion as the BCI approach to emulate the HF/6-31G* ESP. The parameterization algorithm used here is formally identical to that developed by Bush *et al.*[46] in the consensus fitting of BCIs to HF/6-31G* ESPs. The AM1-BCC model uses much fewer atom-types than the BCI method since it takes advantage of the AM1 atomic charges to express subtle chemical variations of electron distribution. This further reduces the degrees of freedom in the parameterization compared to the BCI consensus fitting, and so the same or even a greater increase in numerical stability is expected.

2 The Concept of Bond Charge Corrections

The concept of the AM1-BCC methodology[53] is to correct an initial set of charges, the AM1 charges, that do not fully emulate the HF/6-31G* electrostatic potential (ESP) into a new set of charges that do, namely the AM1-BCC charges, using simple additive correction terms:

$$Q^{AM1} + Q^{BCC} = Q^{AM1-BCC} \tag{2.1}$$

The first term, Q^{AM1}, represents the AM1 atomic charges[51] on a molecule. They are relatively quick to calculate and capture fundamental features of the electronic structure of a molecule (e.g. formal charge and delocalization) but poorly replicate the HF/6-31G* ESP. The second term, Q^{BCC}, is a correction term that adjusts the AM1 charges to produce the AM1-BCC charges, $Q^{AM1-BCC}$ which emulate the HF/6-31G* ESP. These charge corrections are applied to the atoms according to their immediate valence bonded environment in a molecule and are therefore called bond charge corrections or BCCs. Equation 2.1 can be written for the individual atoms in a molecule:

$$q_j^{AM1} + q_j^{BCC} = q_j^{AM1-BCC} \tag{2.2}$$

where q_j^{AM1} is the AM1 atomic charge on atom j, q_j^{BCC} is the net correction applied to the AM1 charge on atom j, and $q_j^{AM1-BCC}$ is the AM1-BCC charge on atom j. Thus, AM1-

BCC atomic charges of *ab initio* ESP-fit quality can be generated by simply performing a semi-empirical calculation to obtain the AM1 atomic charges followed by addition of BCCs (the HF/6-31G* ESP does not need to be calculated by the user).

2.1 Theory

2.1.1 AM1 Atomic Charges

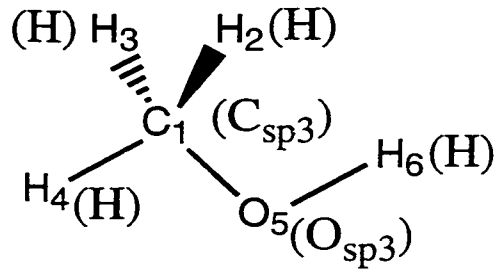
Several pre-charge models were initially considered: charges from extended Hückel theory (EHT)[54] and semi-empirical (MNDO[5, 6], AM1[7], PM3[8, 9]) atomic charges. Charges from EHT did not perform well in reproducing the QM ESP within a BCC model and were consequently abandoned. The performance of all semi-empirical atomic charges was nearly identical; AM1 atomic charges were ultimately chosen for this work because of their availability, robustness and their wide acceptance within the scientific community. AM1 atomic charges are the standard Coulson-type population charges produced by default by MOPAC-6 and are derived from the Coulson density matrix (see Appendix A).

2.1.2 Bond Charge Corrections

The second term in eq. 2.2, q_j^{BCC}, is a correction term that adjusts the AM1 atomic charge on atom j to reproduce the HF/6-31G* ESP; it is defined as:

$$q_j^{BCC} = \sum_{\alpha=1}^{nb} T_{j\alpha} B_\alpha \qquad (2.3)$$

where the summation runs over the total number of bond-types (*nb*) in a molecule, and B_α is the bond charge correction (BCC) for bond-type α. The bond-type α is determined by the atom-types *J* and *I* of atoms *j* and *i* and the bond order connecting the two atoms, shown for methanol in Figure 2.1*a* and 2.1*b*. Conversion from a bond charge (B_α) to an atom-centered charge (q_j^{BCC}) can be accomplished through the use of the bond connectivity template matrix **T** (referred to as the **T** matrix hereafter). The indices *j* and α of the **T** matrix in eq. 2.3 denote an atom and a bond-type, respectively. The **T** matrix for methanol is shown in Figure 2.1*c*. Rows correspond to the atoms in the molecule, i.e. row 1 represents the C_1 atom, row 2 represents the H_2 atom, and so forth. Columns represent bond-types, i.e. column 1 represents bond-type $C_{sp3} - single - H$, an sp^3 carbon atom single-bonded to a hydrogen atom, column 2 represents bond-type $C_{sp3} - single - O_{sp3}$, an sp^3 carbon atom single-bonded to an sp^3 oxygen atom, and so forth. In this example the bond charge correction for the carbon-oxygen bond-type in methanol is $B_{(Csp3\text{-}single\text{-}Osp3)}$=0.0835. This means that 0.0835 charge units are added to the AM1 atomic charge on the sp^3 hybridized carbon atom and -0.0835 charge units to the AM1 atomic charge on the sp^3 hybridized oxygen atom. By construction, the forward atom in a bond (i.e. atom *j* in the format *i-j*, or O_{sp3} in this case) receives $-B_\alpha$ and the backward atom (i.e. atom *i* in the format *i-j*, or C_{sp3} in this case) receives B_α. Consequently, the charges on the molecule are added in a neutral manner, i.e. $B_{\alpha(J,I)} = -B_{\alpha(I,J)}$. Note that the bond charge correction is set to be zero when two atom-types in a bond are identical, e.g. $B_{(Csp3\text{-}single\text{-}Csp3)}$=0.



a

$$\mathbf{T} = \begin{array}{c} \\ C_1 \\ H_2 \\ H_3 \\ H_4 \\ O_5 \\ H_6 \end{array} \begin{array}{c} \begin{array}{ccc} B_1 & B_2 & B_3 \end{array} \\ \left[\begin{array}{ccc} 3 & 1 & 0 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{array}\right] \end{array}$$

c

$$B_1 = B_{(Csp3\text{-}single\text{-}H)} = 0.0274$$
$$B_2 = B_{(Csp3\text{-}single\text{-}Osp3)} = 0.0835$$
$$B_3 = B_{(Osp3\text{-}single\text{-}H)} = -0.2142$$

b

Figure 2.1. *a*) The methanol molecule. The atom-type *J* of each atom is shown in parentheses. *b*) The bond charge corrections, B_α, needed for methanol, the bond-types α are shown in parentheses. *c*) Definition of the T matrix for methanol. The rows in the matrix correspond to the atoms in the molecule, i.e. row 1 represents the C_1 atom, row 2 represents the H_2 atom, and so forth. The columns in the matrix represent bond-types (α), i.e. column 1 represents a $C_{sp3} - single - H$ bond-type, column 2 represents a $C_{sp3} - single - O_{sp3}$ bond-type, and column 3 represents an $O_{sp3} - single - H$ bond-type.

For example, the process of generating the AM1-BCC charges for N-methyl acetamide (NMA) is shown in Figure 2.2. The set of three numbers adjacent to the atoms represent, in descending order, the AM1 atomic charges, the net bond charge corrections for that atom, and the AM1-BCC charges. Table 2.1 lists the bond-types and the BCC values needed to generate the AM1-BCC charges for NMA. The AM1 atomic charge of each atom is corrected by applying the proper BCC values from Table 2.1.

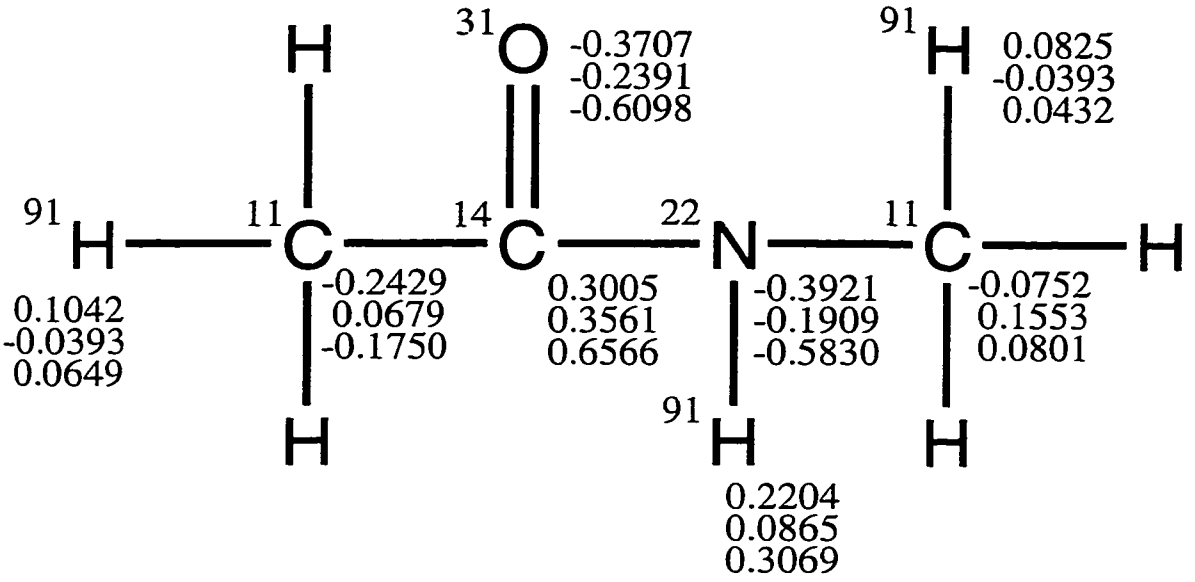


Figure 2.2. N-methyl acetamide. A set of three numbers is shown next to each atom. Row 1 shows the AM1 atomic charges. Row 2 shows the sum of the BCCs for that atom and row 3 shows the AM1-BCC charges. The atom-types are displayed on the upper left side of each atom. The unlabeled hydrogen atoms on the methyl groups are equivalent in atom-type and charge to the labeled hydrogen atoms on their respective methyl groups.

TABLE 2.1.
The Bond-types and Bond Charge Corrections Needed to Charge NMA[a].

Bond-types[b]	BCC[c]
11 01 91	0.0393
11 01 14	-0.0500
14 02 31	0.2391
14 01 22	0.0670
22 01 91	-0.0865
11 01 22	0.0374

a See figure 2.2.

b A bond-type (e.g. 11 01 91) is composed of atom-type *I* (e.g. 11), a bond order (e.g. 01), and atom-type *J* (e.g. 91), respectively (see figures 5.4 and 5.5 for the atom- and bond-typing definitions).

c The BCC values are added to the AM1 atomic charge of atom i and subtracted from the AM1 atomic charge of atom j (see Table 5.1 for a full listing).

2.2 Novelty and Added Value

The novelty of the AM1-BCC concept is in the *combination* of the AM1 wavefunction to generate AM1 atomic charges as an initial guess to the HF/6-31G* ESP *and* the use of ESP-fit BCCs to correct the AM1 atomic charges. The concept can be considered a hybrid between the BCI work of Bush *et al.*[46] and the CM1 charge model of Truhlar's group[47]. Similarly to the CM1 work, the AM1-BCC model calculates AM1 atomic charges as an initial guess; but contrary to CM1, the AM1-BCC model corrects the AM1 atomic charges using simple bond order assignments (e.g. single, double, triple, etc.) to reproduce an ESP whereas CM1 uses fractional bond orders to reproduce a dipole moment. The consensus BCI (bond charge increments) method[46] simply fits BCIs to the ESP; an initial set of pre-charges (e.g. AM1 atomic charges) are not present, therefore the BCIs must bear the total weight of reproducing the ESP taking into account formal charges, π-delocalization, and nearby σ-donating or withdrawing groups.

Employing the AM1 wavefunction to capture subtle chemical variations in the electron distribution of a molecule and the underlying features such as formal charges and delocalization is especially useful in densely functionalized molecules such as molecules with complicated electron distributions such as DNA bases or "drug-like" molecules. By taking advantage of the AM1 atomic charges, the AM1-BCC model should need fewer atom-types than the BCI method (based on the MMFF atom- and bond-typing schemes). This further reduces the degrees of freedom in the parameterization compared to the BCI consensus fitting[46], and so the same or even a greater increase in numerical stability is expected.

The bond-based corrections of the CM1 model depend upon an analysis of the semi-empirical wavefunction to generate bond orders, whereas AM1-BCC uses a simple formal bond order derived from the bond topology of the molecule (e.g. single, double, etc.). In CM1 and CM2 the parameterization scheme is directed towards reproducing the experimental molecular dipole moments of the molecules in the training set; this emphasizes the far field aspects of the ESP. In AM1-BCC, fitting directly to the short-range sampling of the ESP captures features of higher-order electric multipoles important to characterize short-range strong hydrogen bonding. Fitting directly to the HF/6-31G* ESP as in AM1-BCC uses much more information per molecule, as well as embedding the necessary over-polarization of the *ab initio* wave function. As noted by Jorgensen's group, the CM1P charges were inappropriate for solution phase simulations and needed to be scaled by 1.3.[49] Fitting to the ESP also allows the bond charge corrections the opportunity to make up for any systematic intrinsic deficiency in the semi-empirical wavefunction vis-a-vis the *ab initio* 6-31G* wavefunction.

3 Parameterization of the Bond Charge Corrections

3.1 General Approach

The general approach used to parameterize the BCCs was partly based on standard ESP-fitting methods and partly on the BCI work of Bush *et al.*[46] A least-squares fitting approach was used to determine the optimal BCCs that correct AM1 atomic charges to reproduce the HF/6-31G* ESP for a training set of molecules. In order to fit to BCCs, and not to atomic centers, the concept of the template matrix was incorporated into the fitting process.[46]

3.2 Theory

The initial steps of BCC parameterization are similar to standard ESP-fitting protocols.[37] Initially, the *ab initio* or quantum mechanical (QM) ESP, V_l^{QM}, of a molecule is calculated at the HF/6-31G* level of theory at a set of grid points l around the molecule. The grid points are located outside the van der Waals (vdW) surface at 1.4 to 2.0 times the vdW radius of each atom[55] (see appendix B for the list of van der Waals radii used). A face centered cubic grid with 0.5Å spacing was used.

Using any set of atomic charges, the ESP may be calculated for the same set of grid points as for V_l^{QM}; this calculated ESP, V_l^{calc}, is given by the simple Coulombic equation,

$$V_l^{calc} = \sum_{j=1}^{N} \frac{q_j}{r_{lj}} \tag{3.1}$$

where q_j is the charge centered on atom j, r_{lj} is the distance between grid point l and atom j and N is the total number of atoms. In our charge model however, q_j is a sum of two terms: the AM1 charges and the BCCs (see section 2.1.2). Substituting q_j in eq. 3.1 with the definitions found in eqs. 2.2 and 2.3 V_l^{calc} can be written as follows:

$$V_l^{calc} = \sum_{j=1}^{N} \frac{q_j^{AM1}}{r_{lj}} + \sum_{j=1}^{N} \sum_{\alpha=1}^{\gamma} \frac{T_{j\alpha} B_\alpha}{r_{lj}} \tag{3.2}$$

Two quantities are now defined: V_l^{diff}, the difference between the QM ESP and the ESP generated by the AM1 charges:

$$V_l^{diff} = V_l^{QM} - \sum_{j=1}^{N} \frac{q_j^{AM1}}{r_{lj}} \tag{3.3}$$

and V_l^{corr}, the ESP generated by the BCCs:

$$V_l^{corr} = \sum_{j=1}^{N} \sum_{\alpha=1}^{\gamma} \frac{T_{j\alpha} B_\alpha}{r_{lj}} \tag{3.4}$$

This potential difference, V_l^{diff}, can be thought of as an ESP correction at each grid point *l*. The B_α parameters are fitted to an ESP difference and are therefore corrections to the AM1 charges, not full atomic charges *per se*. The objective function, χ^2, is constructed which summarizes the difference between V_l^{diff} and V_l^{corr}:

$$\chi^2 = \sum_{l=1}^{M} \left(V_l^{QM} - V_l^{calc}\right)^2 = \sum_{l=1}^{M} \left(V_l^{diff} - V_l^{corr}\right)^2 = \sum_{l=1}^{M} \left(V_l^{diff} - \sum_{j=1}^{N} \sum_{\alpha=1}^{\gamma} \frac{T_{j\alpha} B_\alpha}{r_{lj}} \right)^2 \tag{3.5}$$

The summation in eq. (3.5) runs over the grid points *M*. The derivative of the objective function with respect to each BCC parameter B_β is set to zero in order to produce the best overall fit to V_l^{diff}

$$\frac{\partial \chi^2}{\partial B_\beta} = \sum_{l=1}^{M} \sum_{k,j=1}^{N} \sum_{\alpha=1}^{\gamma} \frac{T_{\beta k} T_{j\alpha} B_\alpha}{r_{kl} r_{lj}} - \sum_{l=1}^{M} \sum_{k=1}^{N} \frac{T_{\beta k} V_l^{diff}}{r_{kl}} = 0 \tag{3.6}$$

Unlike fitting charges to an ESP, this method does not require a Lagrangian constraint to conserve net charge. The AM1 charges take care of any formal charges and the B_α have no effect on total molecular charge (see section 2.1.2). Eq. 3.6 can also be expressed in matrix form as

$$\mathbf{Ax} = \mathbf{b} \tag{3.7}$$

where

$$A_{\beta\alpha} = \sum_{l=1}^{M} \sum_{k,j=1}^{N} \sum_{\beta=1}^{\gamma} \frac{T_{\beta k} T_{j\alpha}}{r_{kl} r_{lj}} \tag{3.8}$$

,

$$b_{\beta} = \sum_{l=1}^{M} \sum_{k=1}^{N} \frac{T_{\beta k} V_l^{diff}}{r_{kl}} \tag{3.9}$$

and

$$x_{\alpha} = B_{\alpha} \tag{3.10}$$

These linear equations can be solved by standard methods (e.g. Gauss-Jordan elimination, LU decomposition, etc.).[56] Simultaneously fitting the ESPs of multiple molecules or conformations can be accomplished by summing the $A_{\beta\alpha}$ matrices and b_{β} vectors of each molecule (or each conformation) according to parameter types.

The parameterization process is qualitatively demonstrated in Figure 3.1. The QM (V_l^{QM}) and AM1 ESPs for imidazole (see Figure 4.1 **36** for orientation) are shown in Figures 3.1a and 3.1b respectively. The difference between them (V_l^{diff}) is depicted in Figure 3.1c. The atom charges implied by the fitted B_{α} are added to the AM1 atomic charges to give total atomic charges. These in turn give rise to the AM1-BCC ESP (V_l^{calc}), which is very similar to the QM ESP (cf. Figure 3.1a and 3.1d).

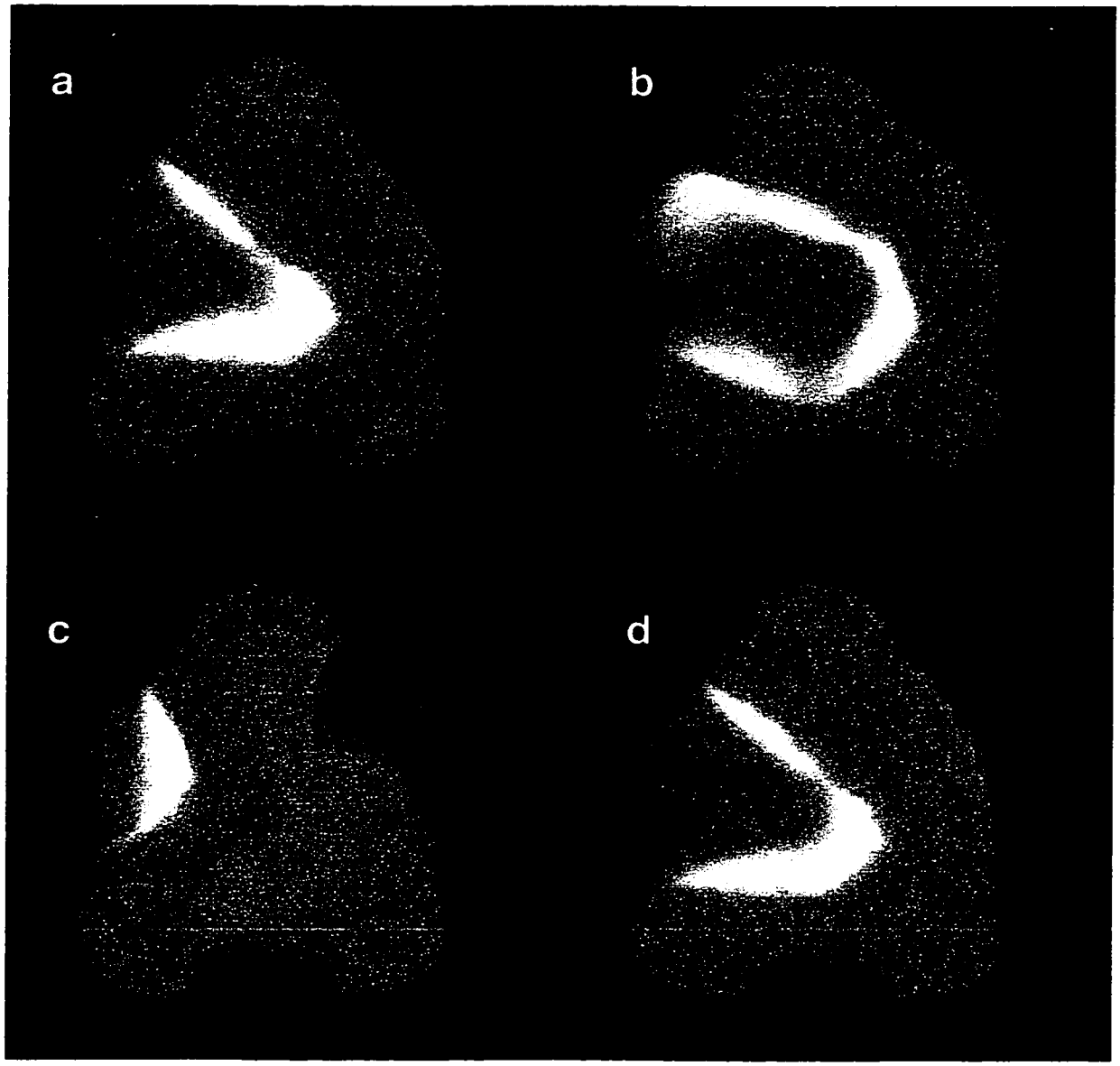


Figure 3.1. The ESP of imidazole (orientation shown in Figure 4.1 **36**) showing negative potentials in red and positive potentials in blue: *a*) QM ESP (V_l^{QM}), *b*) the ESP generated by AM1 atomic charges, *c*) the difference between a and b, i.e. V_l^{diff} and *d*) the ESP generated by the AM1-BCC charges (V_l^{calc}). Note that the full color spectrum has been mapped to the range of the ESP in each case. The actual range differs between the various ESPs shown

4 Preliminary Study for Proof-of-Concept

4.1 General Considerations & Criteria for Proof-of-Concept

The idea behind the AM1-BCC charge model is to correct an initial set of AM1 atomic charges to emulate the HF/6-31G* ESP of molecules. In order to verify that this concept actually worked, a preliminary study was performed. A training set of 45 molecules was created; their geometries optimized with MMFF[44] implemented in the program OPTIMOL[44]; atom- and bond-types were assigned; the HF/6-31G* ESP and the AM1 atomic charges for each molecule in the training set were then calculated; the ESP due to the AM1 atomic charges was generated and subtracted from the HF/6-31G* ESP to obtain the difference potential; the BCCs were then fitted to this difference potential using the formalism outlined in Chapter 3. Then the quality of AM1-BCC charges was assessed by examining how well the AM1-BCC charges reproduced the HF/6-31G* ESP. The AM1-BCC dipole moments were also compared to the HF/6-31G* ESP-fit dipole moments.

4.2 Criteria for Proof-of-Concept

4.2.1 Quality of the ESP

For a given molecule, the ESP arising from a given charge model (e.g AM1, AM1-BCC, MMFF, or RESP) can be evaluated over the same set of grid points used in the evaluation of the QM ESP. The traditional "RMS" metric[57] is used as the basis of comparison between the HF/6-31G* ESP (V_i^{QM}) and the various charge models (V_i^{calc}):

$$RMS = \sqrt{\frac{\sum_{i=1}^{M} \left(V_i^{QM} - V_i^{calc}\right)^2}{N}} \tag{4.1}$$

where N is the number of atoms in the molecule and the summation runs over the total number of grid points M. The ESP of a molecule having a RMS value less than 0.05 a.u. is considered to be of high quality.

4.2.2 Dipole Moments

The net dipole moment of a neutral molecule can be calculated by summing the charge contributions from the x, y, and z coordinates:

$$\mu = \sqrt{\left(\sum_{i=1}^{\#atoms} q_i x_i\right)^2 + \left(\sum_{i=1}^{\#atoms} q_i y_i\right)^2 + \left(\sum_{i=1}^{\#atoms} q_i z_i\right)^2} \tag{4.2}$$

where μ is the dipole moment is units of debye (D), q_i is the charge on atom i and x_i, y_i, and z_i are the cartesian coordinates of atom i. The dipole moment of a molecule differing by less than 0.5 D from that of the RESP model is considered good. This criterion assures higher-quality dipole moments than AM1 dipole moments.

4.3 Training Sets and Atom/Bond-types

The training set chosen to test the concept of the AM1-BCC charge model was limited to a relatively small set of 45 molecules, consisting of three subsets chosen to address specific issues (shown in Figure 4.1). The first component, **TS1**, contained 15

common non-aromatic oxygen-, nitrogen- and sulfur-containing organic functionalities representing simple, non-delocalized molecules, in particular exercising the $C_{sp3}-single-O_{sp3}$, $C_{sp2}^{C=O}-double-O_{sp2}$, and *O-H* BCCs in a variety of chemical contexts to test their robustness. **TS2**, the second subset, contained 15 mono-, di- and tri-substituted benzene molecules, and was used to test the ability of our charge model to express the delocalized charge density with electron-withdrawing and -donating groups. The third subset, **TS3**, comprised 15 principally heteroaromatic molecules representing pharmaceutically relevant fragments. The geometries of all molecules were optimized with MMFF using the program OPTIMOL. AM1 atomic charges on symmetry equivalent centers (symmetry here relates to bond connectivity) were not made equivalent during the parameterization of the BCCs. The ESPs were generated at the HF level of theory with the 6-31G* basis set using GAUSSIAN-92 (using **prop=(grid,field)** and **nosymm** keywords which allows for the ESP and field properties to be calculated at grid points around the molecule and specifies that any inherent symmetry of the molecule will not be utilized in the calculation). This training set was only a small subset of that required for global parameterization; while some BCCs were represented in only a few contexts, others were sampled more widely and within many contexts. Overall, this set represented a diverse cross-section of organic functionalities and thus sufficed for proof-of-concept.

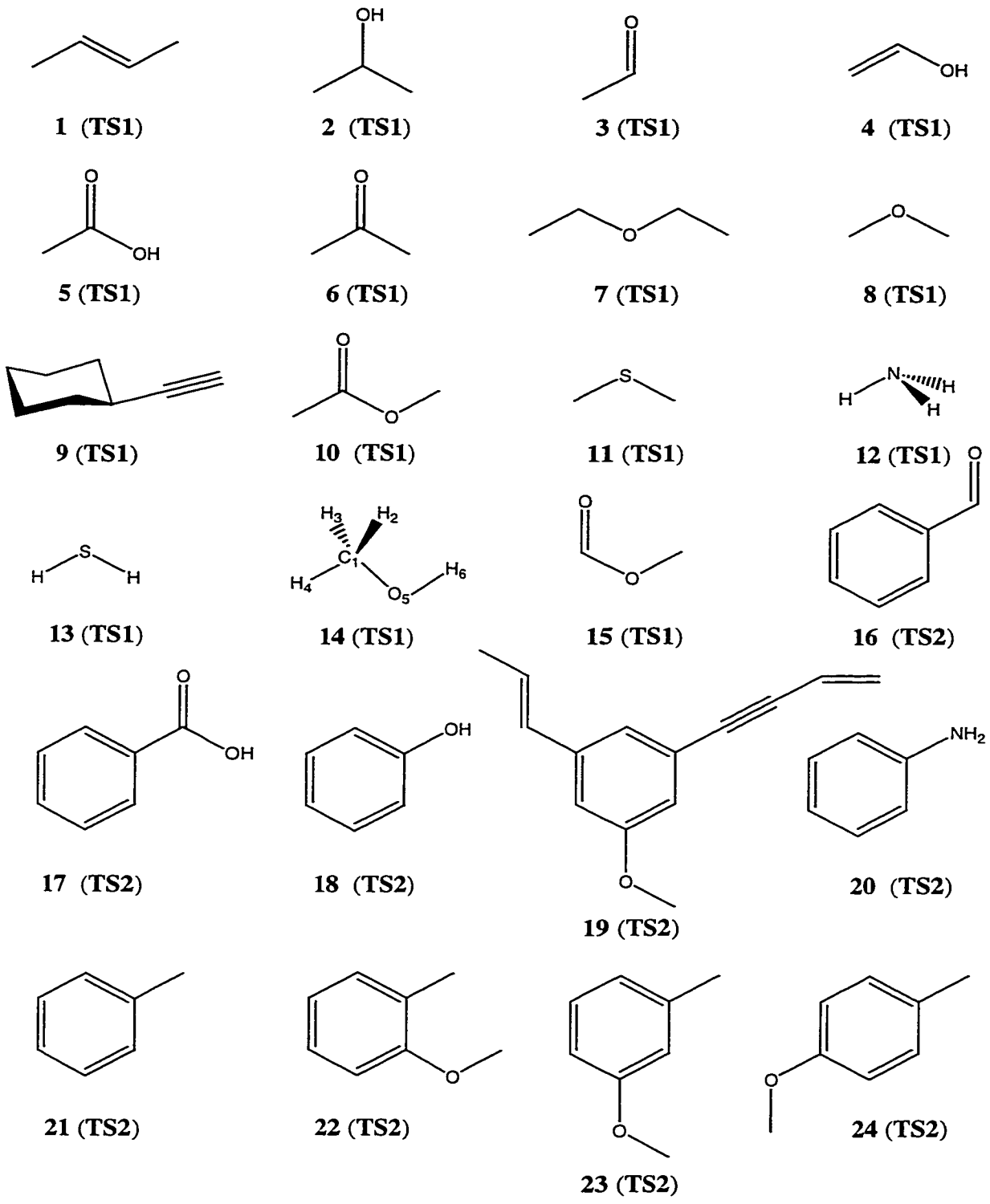


Figure 4.1. Structures for the molecules in **TS1**, **TS2**, and **TS3**; atom-numbering schemes are given for methanol **14**, imidazole **36** and indole **41**. (Continues...).

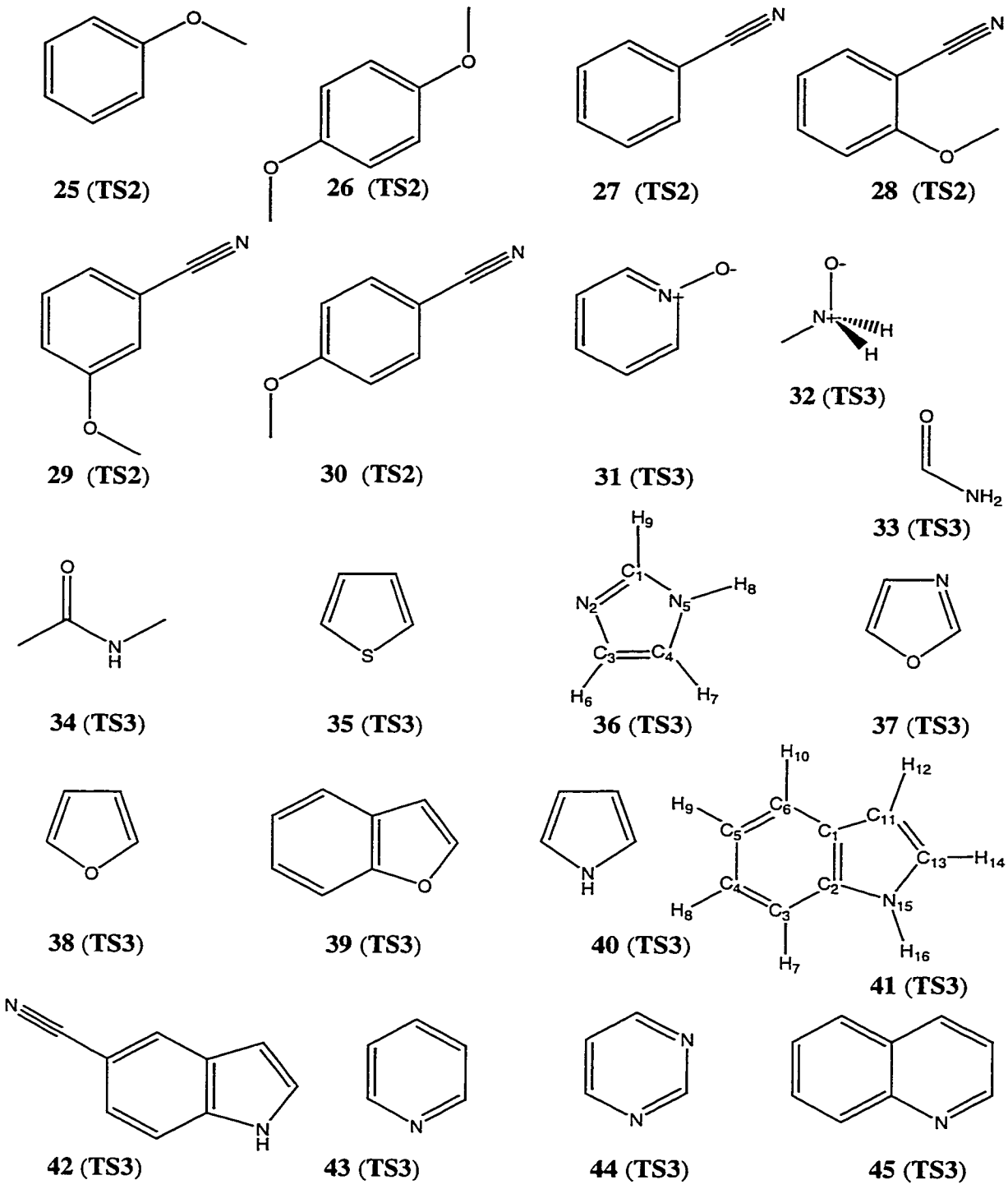


Figure 4.1. (Continuing from previous page)

The simple, parsimonious, and robust atom- and bond-typing scheme chosen allowed efficient atom- and bond-type assignments and reduced the degrees of freedom in the fit, thus increasing numerical stability. A single atom-type was used for each of the sp^3 and sp hybridized carbon, and carbon sp^2 centers had three types: aromatic, non-aromatic, and double-bonded to an electrophilic atom (as in carbonyl). Hydrogen was a single atom-type. Oxygen had only sp^3 and sp^2 types, but nitrogen was found to require slightly more subtle distinctions ultimately comprised of six types: sp^3, sp^3(+) (i.e. a cation), sp, and for sp^2 types: aromatic with lone pair, aromatic with substituent, and non-aromatic. A single sp^3 sulfur atom-type sufficed since only divalent sulfur was in the training set. Bond-types were restricted to formal single, double, and triple, plus aromatic.

4.4 Results and Discussion

All molecules in the training set were used simultaneously to parameterize the BCCs as described in Chapter 3. The performance of the various charge models is evaluated by comparing the RMS deviation of the atomic charge-derived ESP from the QM ESP. This is shown in Figures 4.2, 4.3 and 4.4 for **TS1**, **TS2** and **TS3** respectively. The same general trend is observed in all subsets: the ESP produced by the AM1 atomic charges give the greatest deviation from the QM ESP while the best fit is given by the ESP arising from the RESP charges. The RESP charges are by construction the charges that best fit the QM ESP for an individual molecule because they are directly derived for each molecule from the QM ESP, whereas the AM1-BCC charges reflect a consensus over the entire training set.

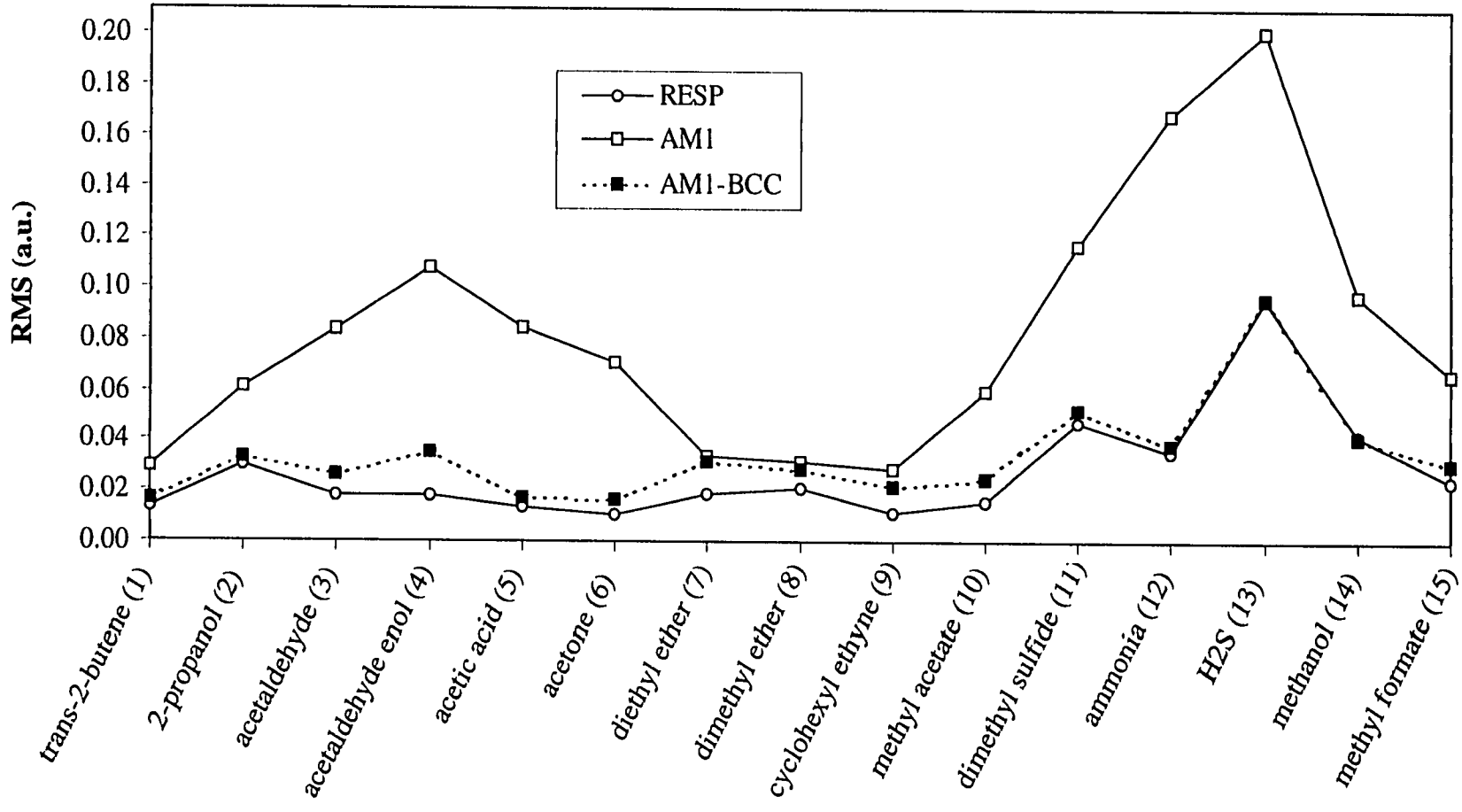


Figure 4.2. RMS deviations from the QM ESP of the ESPs generated by the RESP, AM1 and AM1-BCC charge models for molecules in **TS1**. The large RMS deviation values for **13** (H_2S) are caused because of the inability of an atom-centered charge model to account for the presence of lone-pairs on the sulfur atom. The lines connecting the data points are included as a visual aid and do not imply a systematic variation.

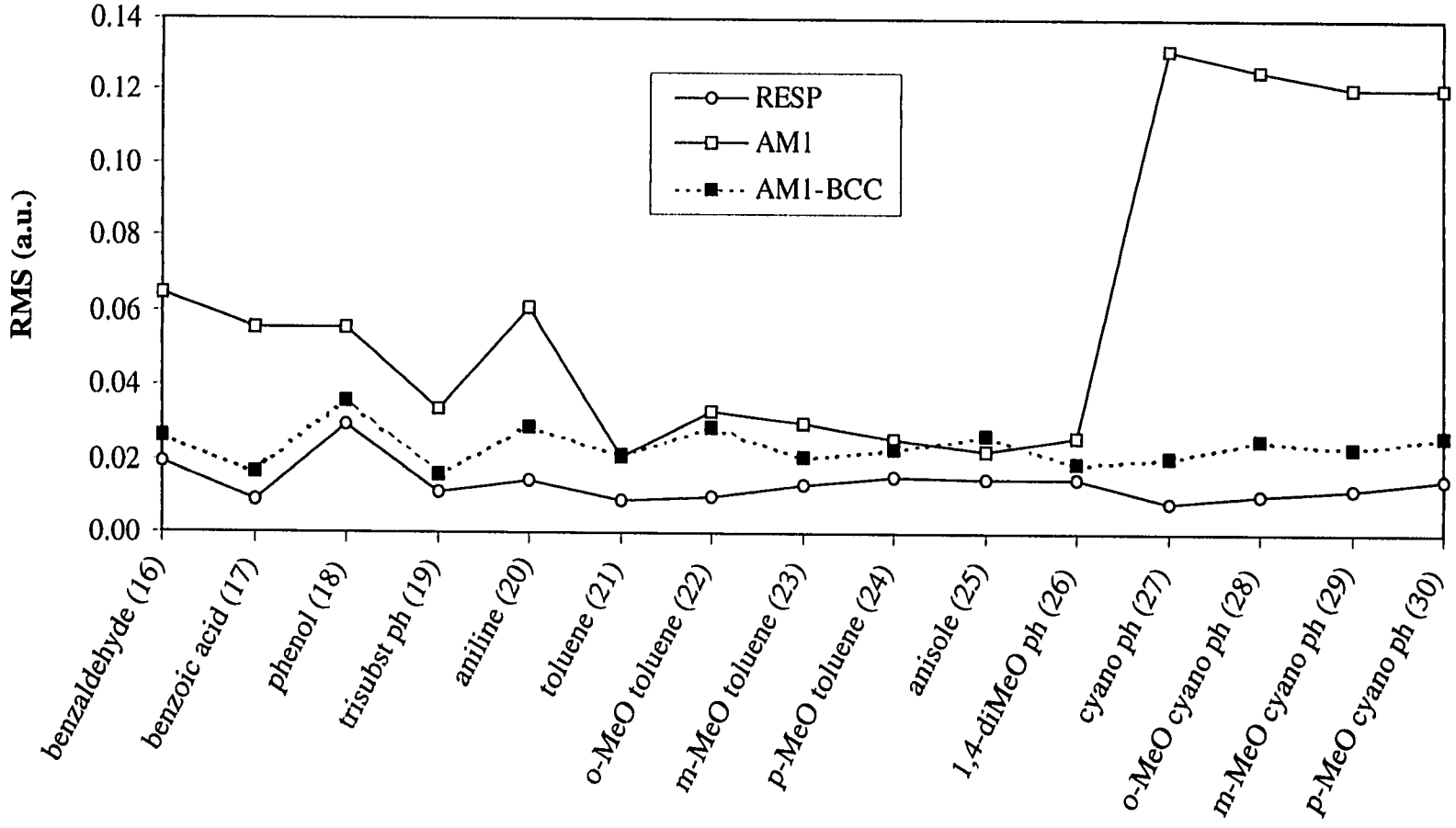


Figure 4.3. RMS deviations from the QM ESP of the ESPs generated by the RESP, AM1 and AM1-BCC charge models for molecules in **TS2**. The large RMS deviation values for **27**, **28**, **29** and **30** (the cyano-containing molecules) are completely corrected by the AM1-BCC charges. The lines connecting the data points are included as a visual aid and do not imply a systematic variation.

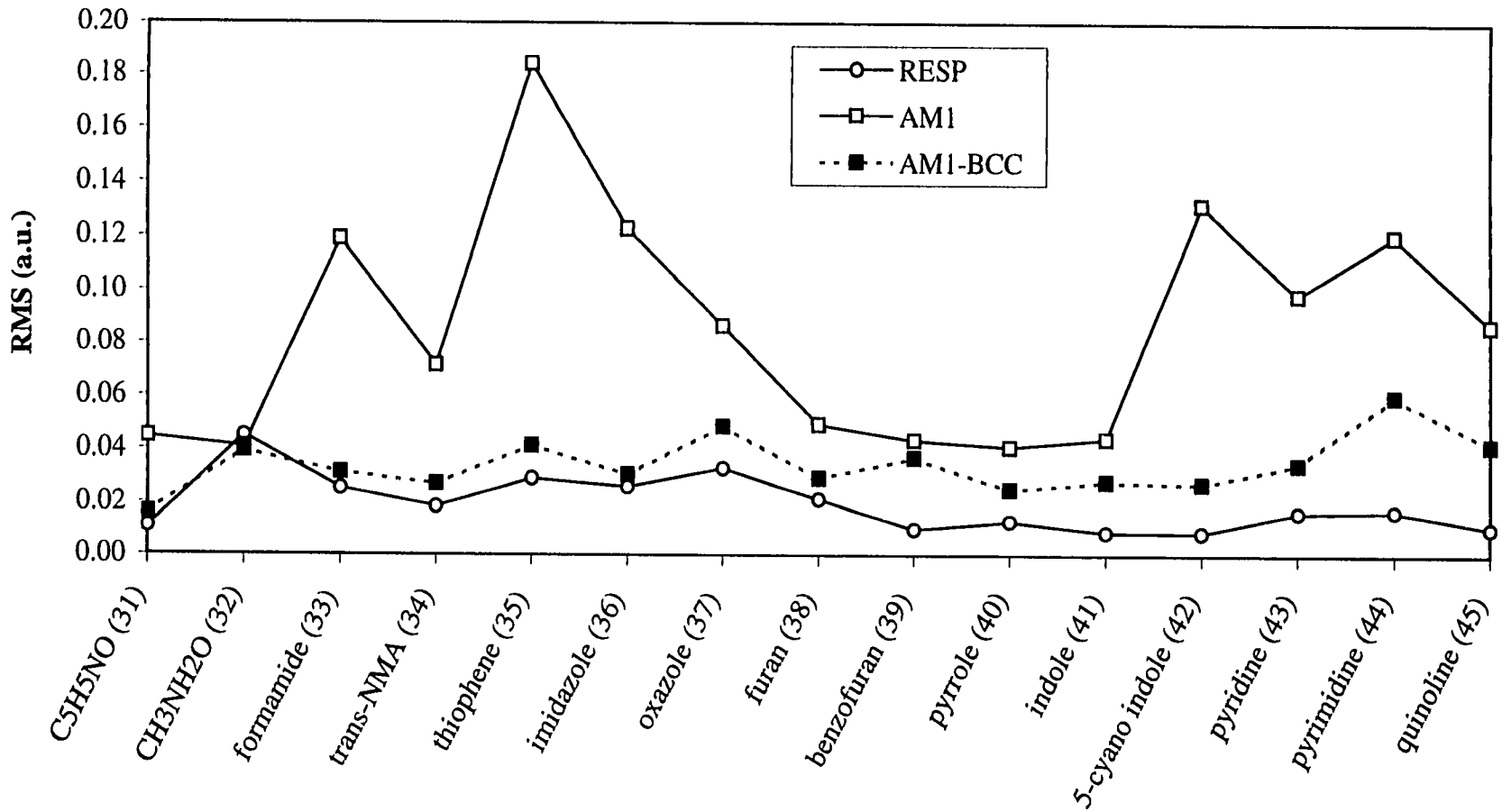


Figure 4.4. RMS deviations from the QM ESP of the ESPs generated by the RESP, AM1 and AM1-BCC charge models for molecules in **TS3**. The large RMS deviation value for thiophene **35** is due to the inability of an atom-centered charge model to account for the presence of lone-pairs on the sulfur atom. The lines connecting the data points are included as a visual aid and do not imply a systematic variation.

Nonetheless there is a dramatic overall improvement in the quality-of-fit when the bond charge corrections (BCCs) are added to the AM1 atomic charges. The greatest improvement is observed for the cyano-containing molecules **27-30** in **TS2** (Figure 4.3). The AM1 atomic charges perform poorly giving RMS deviations above 0.12 a.u., but these drop to values below 0.03 a.u. for the AM1-BCC model. This poor performance of the AM1 atomic charges might be due to AM1's inability to capture the non-local electron distribution of the characteristically strong electron-withdrawing effect of cyano functionalities. In some cases the improvements are very small, e.g. diethyl ether **7** and dimethyl ether **8**, and cyclohexyl ethyne **9** in **TS1**; toluene **21**, *p*-methoxytoluene **24** in **TS2**; and methyl N-oxide **32** in **TS3**. The only molecule that does not benefit from the BCCs is anisole **25** in **TS2** for which the AM1 atomic charges fortuitously perform very well. The RMS of anisole suffers very slightly when corrected with the BCCs because the correction terms are a consensus over the entire training set and do not represent *the* correction terms needed specifically for anisole. The average RMS potential deviations for the molecules of each subset are shown in Table 4.1.

TABLE 4.1.
Average RMS Deviation from the QM ESP[a] for RESP, AM1, and AM1-BCC Charge models.

	RMS Deviation from QM ESP (a.u.)		
Training Subset	RESP	AM1	AM1-BCC
TS1	0.0277	0.0826	0.0339
TS2	0.0139	0.0618	0.0240
TS3	0.0195	0.0856	0.0345
Average	0.0203	0.0766	0.0308

[a] The ESP was calculated at the HF/6-31G* level of theory.

The RMS deviation of the potential generated by AM1 atomic charges alone is lowered substantially when the BCCs are applied, to a value that approaches that of the RESP model (0.0203). The overall behaviour of the AM1-BCC model is consistently similar to the RESP model in terms of the RMS deviation, compared to the erratic and occasionally very poor behaviour of the uncorrected AM1 atomic charges (cf. Figures 4.2, 4.3 and 4.4).

Of the resulting BCCs, a subset is given in Table 4.2; the first eight entries are the only ones required to generate charges for methanol **14**, imidazole **36**, and indole **41**. Each of these BCCs occurs in multiple contexts in the training set, as shown in column five of Table 4.2; hence their performance addresses the generality of the atom- and bond-typing methods.

TABLE 4.2.
Bond Charge Corrections (BCCs) for Methanol, Imidazole, Indole and Test Molecules.

Bond-type α[a]				
Atom-type *J*	Bond Order	Atom-type *I*	B_α[b]	Num. Of Occur.[c]
C_{sp3}	*single*	H	0.0274	124
C_{sp3}	*single*	O_{sp3}	0.0835	18
O_{sp3}	*single*	H	-0.2142	6
C_{sp2}^{arom}	*single*	H	0.0100	122
C_{sp2}^{arom}	*aromatic*	C_{sp2}^{arom}	0.0000	144
NH_{sp2}^{arom}	*single*	H	-0.0882	4
C_{sp2}^{arom}	*aromatic*	NH_{sp2}^{arom}	-0.0110	8
C_{sp2}^{arom}	*aromatic*	N_{sp2}^{arom}	0.1707	12
C_{sp3}	*single*	C_{sp3}	0.0000	10
C_{sp3}	*single*	$C_{sp2}^{C=O}$	-0.0799	6
C_{sp3}	*single*	C_{sp2}^{arom}	-0.0164	4
$C_{sp2}^{C=O}$	*single*	C_{sp2}^{arom}	0.0405	2
C_{sp2}^{arom}	*single*	O_{sp3}	0.0494	11
$C_{sp2}^{C=O}$	*single*	O_{sp3}	0.1176	4
$C_{sp2}^{C=O}$	*double*	O_{sp2}	0.2501	9

[a] Each bond-type is composed of atom-type *J*, a bond order, and atom-type *I*.
[b] The value of B_α is added to the AM1 atomic charge of atom *j* and subtracted from that of atom *i*.
[c] The number of occurrences of a bond-type in the training set.

Table 4.3 shows the charge comparison for methanol **14**, a simple polar compound able to donate and accept hydrogen bonds. The methyl carbon charges differ even in sign between the AM1 and RESP models, and the AM1 oxygen charge is only half that of the RESP oxygen. Addition of the BCCs onto the AM1 atomic charges

produces hydroxyl charges with RESP-like values; the methyl carbon atom adopts a lower magnitude charge. The RMS deviation drops by more than half and is comparable to the RESP value. The dipole moment increases from the AM1 value of 1.31 to 2.14 debye, approaching the 2.22 debye for the RESP model.

TABLE 4.3.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Methanol for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	0.1406	-0.0655	0.1003
$H_{2,3,4}$	0.0314	0.0656	0.0382
O_5	-0.6555	-0.3301	-0.6278
H_6	0.4207	0.1987	0.4129
RMS (a.u.)	0.0418	0.0969	0.0407
Dipole moment (D)	2.2174	1.3053	2.1395

[a] Subscript refers to atom numbering for methanol (**14**) in Figure 4.1.

In the case of imidazole **36** (Table 4.4), improvement in the RMS deviation is more pronounced than for methanol with the RMS deviation decreasing fourfold compared with simple AM1 pre-charges. This improvement is demonstrated in Figure 3.1 where the AM1-BCC ESP (Figure 3.1d) is nearly identical to the QM ESP (Figure 3.1a). The dipole moment increases from 2.32 debye for the AM1 model to 4.06 debye for the AM1-BCC model, very close to the RESP value of 3.99 debye. Interestingly, while the character of the C_1 and C_3 atoms change from negative charges for the AM1 model to approximately neutral for the AM1-BCC model, they remain markedly different from the positive charges of the RESP model. That is, compared to the RESP charges optimized to this individual molecule, the consensus BCC parameters produce less charge separation along the bonds from C_1 and C_3; this might be expected of a consensus

method, and may reflect the greater stability of the procedure. The decrease in accuracy of the ESP is minimal: the RMS deviation for AM1-BCC is only 0.005 atomic units greater than for the optimal RESP model.

TABLE 4.4.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Imidazole for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	0.1465	-0.1312	0.0385
N_2	-0.5286	-0.1455	-0.4869
C_3	0.1313	-0.1653	0.0154
C_4	-0.3067	-0.2063	-0.2073
N_5	-0.2366	-0.1551	-0.2214
H_6	0.1183	0.1785	0.1685
H_7	0.1976	0.1738	0.1638
H_8	0.3377	0.2525	0.3407
H_9	0.1405	0.1987	0.1887
RMS (a.u.)	0.0259	0.1227	0.0306
Dipole moment (D)	3.9850	2.3237	4.0581

[a] Subscripts refer to atom numbering for imidazole (**36**) in Figure 4.1.

In the case of indole **41** (Table 4.5) the improvements are less pronounced than in the previous two cases; the RMS error of fit is more than tripled compared to RESP. Most carbon atoms are bonded to similar carbon atom-types $C_{sp2}^{arom} - arom - C_{sp2}^{arom}$ which have a zero correction term by definition because both atom-types are the same. Consequently, most AM1 atomic charges are only altered by $C_{sp2}^{arom} - arom - H$ BCCs. As a result, the C_1 atom (see Figure 4.1 **41** for the atom-numbering scheme) cannot receive any correction terms since it is bonded to three identical carbon atom-types and therefore the AM1-BCC charge on this atom has the same value as in the AM1 model. In principle this could be alleviated by differentiating subclasses of sp^2 hybridized aromatic

carbon atom-types, but this would compromise the simplicity of our model. Although the improvement in the RMS deviation from the QM ESP is moderate since most charges are similar to the AM1 atomic charges, the overall quality-of-fit of the AM1-BCC model for indole is good (cf. Figure 4.3). The molecular dipole moment is underestimated, albeit markedly improved over the uncorrected AM1 model.

TABLE 4.5.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Indole for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	0.3136	-0.0808	-0.0808
C_2	0.2670	-0.0015	-0.0125
C_3	-0.3535	-0.1478	-0.1378
C_4	-0.1305	-0.1117	-0.1017
C_5	-0.1995	-0.1592	-0.1492
C_6	-0.3213	-0.0809	-0.0709
H_7	0.1867	0.1299	0.1199
H_8	0.1508	0.1278	0.1178
H_9	0.1583	0.1288	0.1188
H_{10}	0.1887	0.1315	0.1215
C_{11}	-0.5415	-0.2081	-0.1981
H_{12}	0.2309	0.1518	0.1418
C_{13}	0.0024	-0.0820	-0.0830
H_{14}	0.1822	0.1652	0.1552
N_{15}	-0.5239	-0.2078	-0.2741
H_{16}	0.3896	0.2448	0.3330
RMS (a.u.)	0.0079	0.0440	0.0280
Dipole moment (D)	2.1245	1.3875	1.8997

[a] Subscript refers to atom numbering for indole (**41**) in Figure 4.1.

The BCCs in Table 4.2 were used to charge three test molecules not in the training set: D-glucose, aspirin, and eriodictyol (Figures 4.5*a*, *b*, and *c*, respectively), the AM1-BCC ESP was then compared to the 6-31G* ESP as well as that produced by RESP charges. These molecules were chosen because they are rich in the *C-O*, *C=O*, and *O-H*

functionalities tested most diversely in the training set, and yet they were still tractable for the calculation of the 6-31G* wavefunction and electrostatic potential.

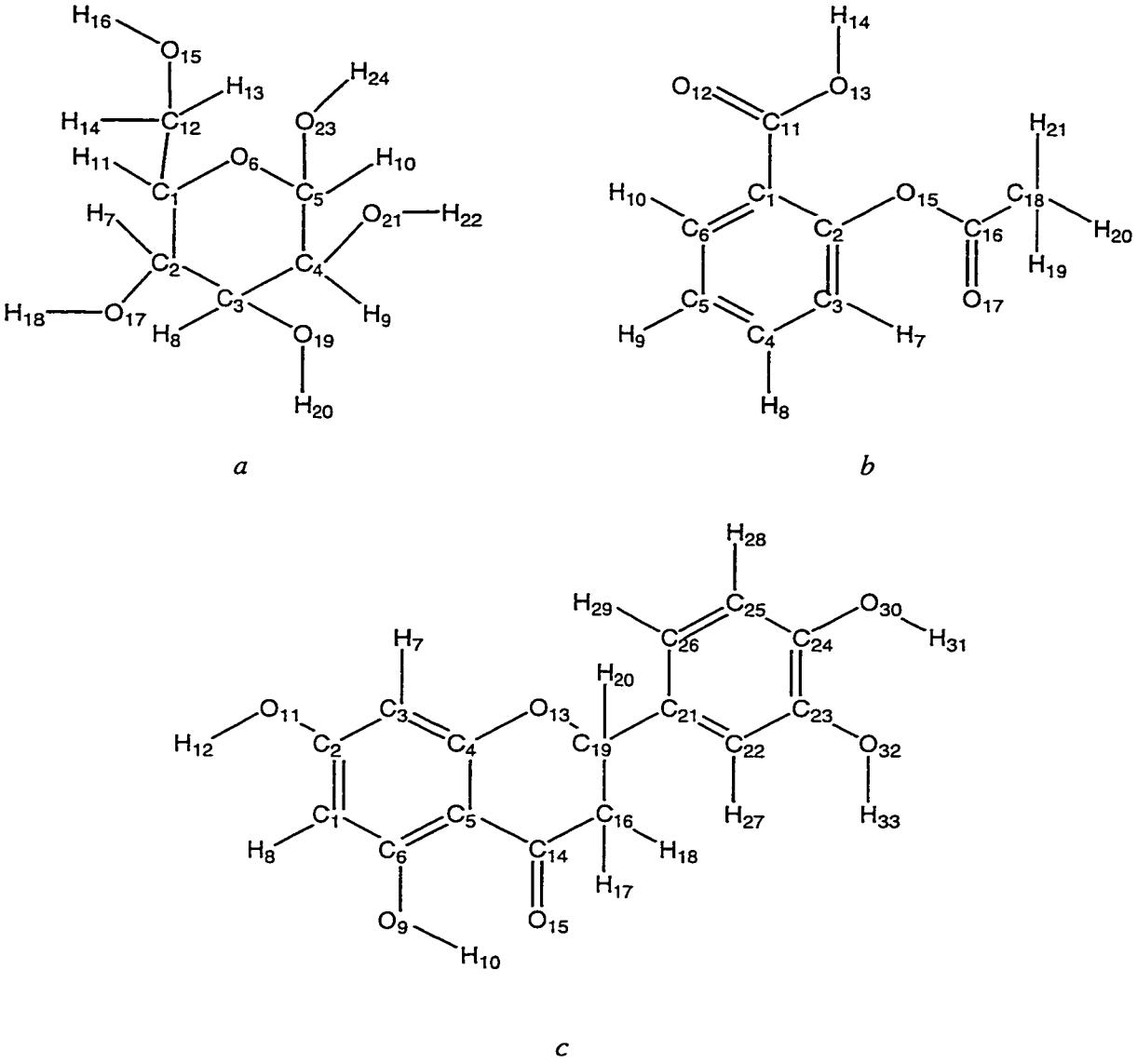


Figure 4.5. Structure and atom-numbering schemes for the test set molecules: *a*) D-glucose, *b*) aspirin, and *c*) eriodictyol.

AM1 calculations were carried out and the BCCs in Table 4.2 were applied using equations 2.2 and 2.3 to arrive at a set of AM1-BCC charges for these molecules. This

AM1-BCC procedure took only 3 seconds for the set of three molecules. By contrast, approximately 100 minutes for the set of three molecules were needed for 6-31G* calculations required to generate the QM ESPs and then approximately 5 human minutes for the set of three molecules to assign equivalencies and restraints needed to fit RESP charges. The QM ESP calculations were run on a CrayYMP supercomputer. Tables 4.6, 4.7, and 4.8 give the charges for D-glucose, aspirin, and eriodictyol, respectively, and compare the RMS deviations from the QM ESP as in Tables 4.3 to 4.5. As with the training set, the AM1-BCC charges produce ESPs and dipole moments that are far more RESP-like than do the AM1 atomic charges (although the AM1-BCC dipole for aspirin is slightly exaggerated). In particular, the improvements in the RMS deviation were very similar to those observed for the training set (cf. Tables 4.1, and 4.3 to 4.5). While this is not a comprehensive test of the AM1-BCC model, it demonstrates that the model can be applied to organic molecules to arrive at charges of quality comparable with 6-31G* ESP-fit charges, and of greater transferability, with far less computational cost and human effort.

TABLE 4.6.
Charges, RMS Deviations from the QM ESP and Dipole Moments for D-Glucose for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	-0.0312	0.0140	0.1249
C_2	0.0580	-0.0295	0.0814
C_3	0.0367	-0.0266	0.0843
C_4	0.0605	-0.0120	0.0989
C_5	0.1743	0.1224	0.3168
O_6	-0.3237	-0.2948	-0.4617
H_7	0.1028	0.0852	0.0578
H_8	0.1345	0.0909	0.0634
H_9	0.1873	0.1459	0.1185
H_{10}	0.1793	0.1194	0.0920
H_{11}	0.1691	0.1458	0.1184
C_{12}	0.2253	-0.0399	0.0984
$H_{13,14}$	0.0524	0.0948	0.0674
O_{15}	-0.6970	-0.3264	-0.6240
H_{16}	0.4431	0.2186	0.4328
O_{17}	-0.6157	-0.3243	-0.6219
H_{18}	0.4102	0.2104	0.4247
O_{19}	-0.6784	-0.3212	-0.6188
H_{20}	0.4345	0.2101	0.4243
O_{21}	-0.6568	-0.3317	-0.6294
H_{22}	0.4647	0.2327	0.4469
O_{23}	-0.6400	-0.3007	-0.5984
H_{24}	0.4589	0.2219	0.4361
RMS (a.u.)	0.0151	0.1109	0.0263
Dipole moment (D)	5.5880	3.2202	5.2887

[a] Subscript refers to atom numbering for D-glucose as shown in Figure 4.5*a*.

TABLE 4.7.

Charges, RMS Deviations from the QM ESP and Dipole Moments for Aspirin for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	-0.0036	-0.1163	-0.1569
C_2	0.2696	0.0988	0.1482
C_3	-0.2128	-0.1425	-0.1325
C_4	-0.1551	-0.0894	-0.0794
C_5	-0.1238	-0.1442	-0.1342
C_6	-0.2414	-0.0684	-0.0584
H_7	0.1832	0.1488	0.1388
H_8	0.1689	0.1404	0.1305
H_9	0.1504	0.1412	0.1312
H_{10}	0.1998	0.1553	0.1453
C_{11}	0.7064	0.3671	0.7753
O_{12}	-0.5755	-0.3440	-0.5941
O_{13}	-0.5993	-0.2947	-0.6266
H_{14}	0.4479	0.2402	0.4545
O_{15}	-0.4214	-0.2102	-0.3773
C_{16}	0.7796	0.3169	0.7646
O_{17}	-0.5420	-0.3101	-0.5602
C_{18}	-0.4466	-0.2262	-0.2238
$H_{19,20,21}$	0.1386	0.1124	0.0850
RMS (a.u.)	0.0108	0.0652	0.0261
Dipole moment (D)	1.9486	2.0204	2.1336

[a] Subscript refers to atom numbering for aspirin as shown in Figure 4.5*b*.

TABLE 4.8.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Eriodictyol for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	-0.3265	-0.3056	-0.2956
C_2	0.2688	0.1767	0.2261
C_3	-0.1891	-0.2398	-0.2298
C_4	0.0084	0.1843	0.2337
C_5	-0.0471	-0.3495	-0.3901
C_6	0.2116	0.2239	0.2733
H_7	0.1751	0.1679	0.1579
H_8	0.1900	0.1547	0.1447
O_9	-0.5587	-0.2556	-0.5193
H_{10}	0.4624	0.2639	0.4781
O_{11}	-0.5588	-0.2405	-0.5041
H_{12}	0.4386	0.2607	0.4449
O_{13}	-0.2666	-0.1867	-0.3196
C_{14}	0.5201	0.2907	0.6612
O_{15}	-0.5735	-0.3254	-0.5754
C_{16}	-0.2822	-0.2270	-0.2521
$H_{17, 18}$	0.1366	0.1323	0.1049
C_{19}	0.2264	0.0841	0.1786
H_{20}	0.0963	0.1019	0.0744
C_{21}	0.5928	-0.0886	-0.0722
C_{22}	-0.5958	-0.1020	-0.0920
C_{23}	0.4643	0.0665	0.1160
C_{24}	-0.0232	0.0094	0.0589
C_{25}	0.4643	-0.1814	-0.1714
C_{26}	-0.5958	-0.1132	-0.1032
H_{27}	-0.3494	0.1615	0.1515
H_{28}	-0.0754	0.1384	0.1284
H_{29}	0.2201	0.1357	0.1257
O_{30}	-0.6483	-0.2694	-0.5330
H_{31}	0.4710	0.2360	0.4502
O_{32}	-0.6274	-0.2452	-0.5088
H_{33}	0.4692	0.2389	0.4532
RMS (a.u.)	0.0252	0.0803	0.0338
Dipole moment (D)	4.0337	2.2686	4.0399

[a] Subscript refers to atom numbering for eriodictyol as shown in Figure 4.5*c*.

Consistent behaviour of the same functional group in different bonded contexts is important in parameterizing a force field, especially in order to minimize problems with 1,4 electrostatic torsional terms. Table 4.9 compares this consistency between the AM1-BCC and RESP models for the $C_{sp3} - single - O_{sp3}$ bond-type found in ether functional groups.

TABLE 4.9.
Charges on C_{sp3} and O_{sp3} Atom-types in the C_{sp3}-*single*-O_{sp3} Bond-Type in Methyl Ether Functional Groups.

Molecule	Charge on C_{sp3}		Charge on O_{sp3}	
(cf. Figure 4.1)	RESP	AM1-BCC	RESP	AM1-BCC
Dimethyl ether **8**	-0.0598	0.0947	-0.3376	-0.4386
Methyl acetate **10**	-0.0164	0.0941	-0.4397	-0.4604
Methyl formate **15**	-0.0416	0.0879	-0.4157	-0.4615
Anisole **25**	0.0706	0.0956	-0.3627	-0.3373
o-MeO toluene **22**	-0.0814	0.0984	-0.2556	-0.3407
m-MeO toluene **23**	0.0350	0.0960	-0.3355	-0.3377
p-MeO toluene **24**	0.0385	0.0957	-0.3590	-0.3378
o-MeO cyano ph **28**	-0.0172	0.0947	-0.2847	-0.3276
m-MeO cyano ph **29**	0.0509	0.0940	-0.3513	-0.3326
p-MeO cyano ph **30**	0.0688	0.0875	-0.3553	-0.3360
1,4-diMeo ph **26**	0.0329	0.0965	-0.3467	-0.3383
Tri-sub ph **19**	-0.0421	0.0950	-0.3264	-0.3366
Minimum	-0.0814	0.0875	-0.4397	-0.4615
Maximum	0.0706	0.0984	-0.2556	-0.3276

Although RESP was developed to reduce numerical instabilities during the ESP-fitting process,[41] examination of Table 4.9 suggests that they are still present in the RESP model to some degree. The RESP-fit charges on the carbon atoms fluctuate between –0.0814 and 0.0706 while the AM1-BCC charges on the same atoms only vary between 0.0875 and 0.0984. Although the fluctuations in charge on the O_{sp3} atom-types are

larger (because of the different bonded contexts), the AM1-BCC charges are much more stable within each molecular grouping. The small variation in charge on the methyl carbon atom, one of the most unstable centers in ESP-fit methods (i.e. large fluctuations in charge produce small variations in the RMS), demonstrates the numerical stability of the AM1-BCC model over ESP-fit methodologies. Particularly interesting are the chemically non-intuitive decreases in RESP charges on the oxygen and methyl carbon of the *ortho*-methoxy substituted toluene **22** and benzonitrile **28** compared to the related *meta* and *para* substitution patterns. The adjacent *ortho* substituents on the phenyl ring occlude the surface of the methoxy from the grid-point sampling for the QM ESP, burying the methoxy oxygen and carbon atoms even more than usual, thus exacerbating the numerical instability of the charges fitted to these centers. To confirm that the charge decreases could be attributed to numerical instability, the RESP charge fitting for **22** and **28** was repeated holding the oxygen charge constant at the *meta*-substitution values of –0.3355 and –0.3513, respectively (cf. Table 4.9); the RMS deviation increased by only 0.0004 for **22** and 0.001 for **28**. Therefore the large variations in charge are not needed to reproduce the ESP with high accuracy. This unnecessary variation of the RESP charges would degrade transferability of the ether torsion parameters due to the inconsistency of the 1-4 electrostatic terms in different molecules. In contrast, the AM1-BCC charges show no unusual behaviour for these centers and yet give a quality-of-fit comparable to the other isomer substitution patterns (cf. Table 4.9 and Figure 4.3). This consistent behaviour of the AM1-BCC charges would benefit the force field parameterization of torsion terms of the ether functional groups.

4.5 Conclusions

The AM1-BCC charge model satisfies the criteria for proof-of-concept. It performs well in reproducing the QM ESP of a variety of polar, non-polar and aromatic molecules with an average RMS value over the three training sets of 0.0308 a.u., although in a few cases the RMS was higher than 0.05 a.u. (e.g. dimethyl sulfide and pyrimidine). Generally, they show dramatic improvement over the AM1 atomic charges, approaching RESP quality, although in some challenging cases (such as indole) the improvement is moderate. The new charge model also corrects the erratic behaviour of the AM1 atomic charges in their fit to the QM ESP. The AM1-BCC model tends to produce charges of lower magnitude than the RESP model but with similar molecular dipole moments; this may reflect a greater stability of the consensus fit. The AM1-BCC dipole moments were within 0.3 D of the RESP dipole moments for the six molecules examined, well within the tolerance of 0.5 D. The problem of numerical instability on buried centers is greatly reduced provided that the training set includes sufficient examples of the same bond parameter types in well-exposed contexts, as previously discussed and as demonstrated here with the ether bond-type. With the ability of the AM1-BCC model to emulate the HF/6-31G* ESP, this method offers an effective alternative to human and computer resource-intensive HF/6-31G* ESP-fitting methods for generating atomic charges. Based on the encouraging results of this preliminary study, global parameterization of the AM1-BCC charge model was undertaken.

5 Global Parameterization of BCCs

5.1 General Considerations & Issues from Proof-of-Concept Chapter

The concept of the AM1-BCC charge model was validated in Chapter 4. The objective of this chapter is to generate a complete set of BCCs that would enable this charge model to be used on any organic molecule. This requires development and construction of a training set that adequately samples bond-types relevant to organic chemistry. The completeness of this training set is cross-examined with The Merck Index[58] (an encyclopedia of > 10000 chemicals, drugs, and biological molecules) and the National Cancer Institute (NCI) structure database of ca. 250000 molecules.[59]

The preliminary study of proof-of-concept not only proved the validity of the idea, it also revealed problem areas to be anticipated in the global parameterization effort as well as areas in need of improvement. One such improvement was the replacement of the MMFF geometry optimization method, before the AM1 charges are calculated, with the AM1 geometry optimization method. Optimization of the molecule and calculation of the AM1 atomic charges are now performed in a single step: the molecule is optimized at the AM1 level and the AM1 atomic charges are obtained at the end of the optimization steps. This improves efficiency by reducing the number of steps needed to charge a molecule and simplifies the overall methodology.

The small training set of 45 molecules in the proof-of-concept study demonstrated that some atom–types required sub-classification. For example, sp^2 hybridized carbon

atom-types needed to be classified into three categories (see section 4.3). Similar trends were also observed with the nitrogen atom-types. Therefore, with a global training set, the need for further sub-classifications is anticipated.

The preliminary study also revealed deficiencies in the proposed charge model. The RMS deviations from the QM ESPs of pyridine, pyrimidine, and quinoline were observed to be high, especially for pyrimidine (see Figure 4.4). Both pyridine and quinoline contain a nitrogen atom with a lone pair of electrons within an aromatic ring while pyrimidine contains two such nitrogens. The high RMS values for these molecules arise because of insufficient degrees of freedom needed to describe the lone pair regions of these molecules. These considerations also indicated the need for increased degrees of freedom through sub-classification of atom-types.

With these issues in mind, global parameterization of the BCCs was started. A training set of molecules was developed and constructed along with an atom- and bond-typing scheme, the *ab initio* ESPs and the AM1 atomic charges for each molecule in the training set were generated, the differences between the *ab initio* and AM1 ESPs were calculated, and then the BCCs were fitted to these difference ESPs.

5.2 The Training Set of Molecules

The training set presented here is composed of 2755 organic molecules that sample organic bond-types composed of H, C, N, O, F, P, S, Cl, Br, and I atoms. Although Si containing bonds are also sampled, no effort was made to ensure that all Si bond-types were present in the training set. In order to improve the statistical significance of the

BCCs, each bond-type was sampled across a minimum of four different chemical environments.

To ensure a diverse and systematic sampling of common organic bond-types, a scaffold/functional group scheme was developed. A set of molecular scaffolds (i.e. molecular fragments that contain an empty valence) were created onto which can be inserted a variety of functional groups (i.e. as with the scaffolds, these functional groups also contain an empty valence) to create a large number of diverse organic molecules. Primary functional (pf) groups (depicted in Figure 5.1) were systematically added to primary scaffolds (ps) (depicted Figure 5.2.). For example, functional group pf01 is added to scaffold ps08 to build a molecule of acetone, functional group pf07 is added to scaffold ps06 to build a molecule of phenol, and so on. Similarly, a set of secondary functional groups (sf) and scaffolds (ss) (Figures 5.1 and 5.2, respectively) are used to sample an extended range of functionalities (whereas the primary set of scaffolds/functional groups sampled basic molecular fragments, i.e. common to amino acids, the secondary set is more diverse). The primary functional groups are added to the secondary scaffolds in the same manner, in addition, the secondary functional groups are also added to the primary scaffolds to increase sampling of various bond-types. This scaffold/functional group scheme generated a total of 390 molecules after elimination of a few multiple occurrences (e.g. ps06_pf01 and ps07_pf02).

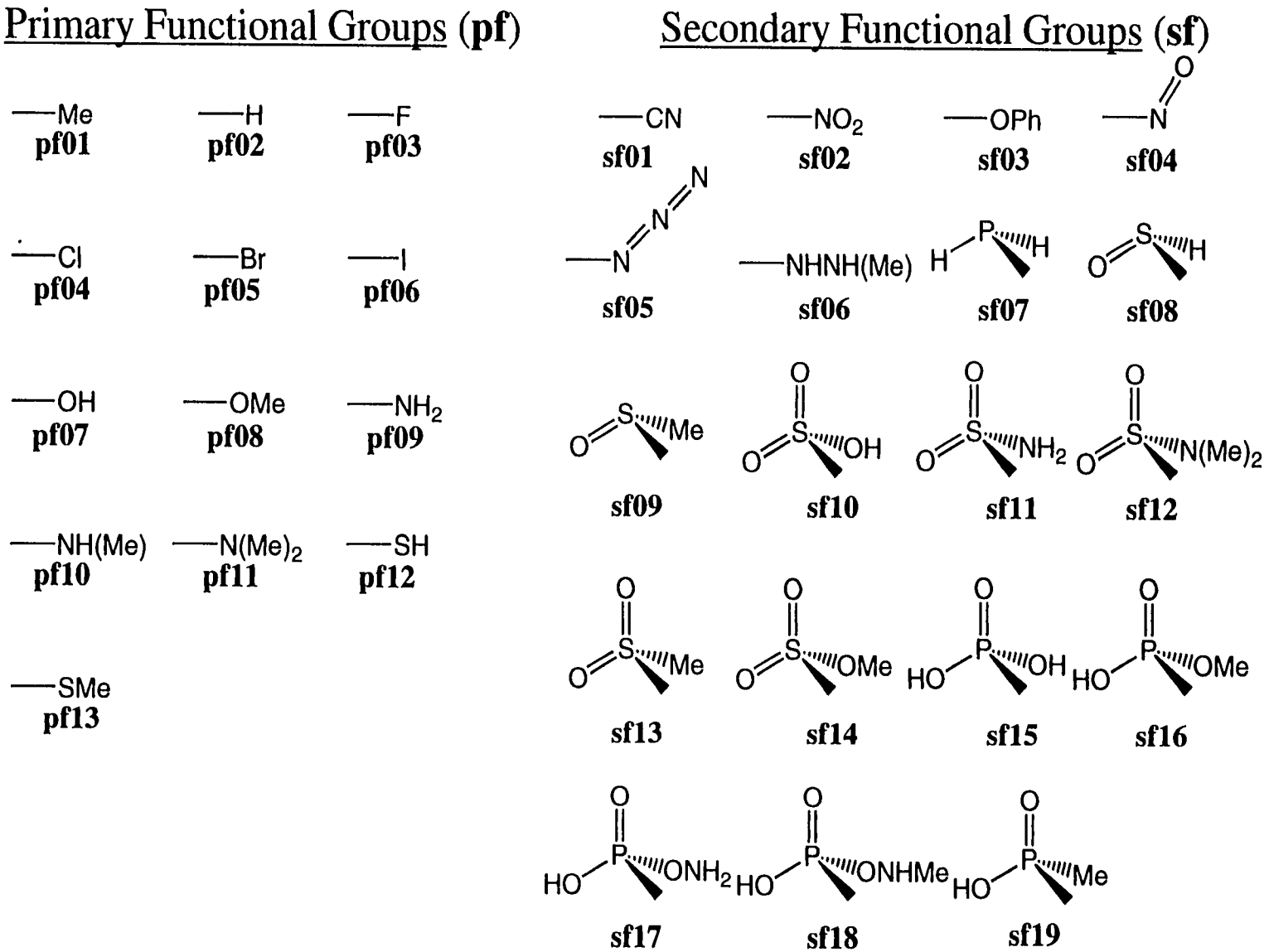
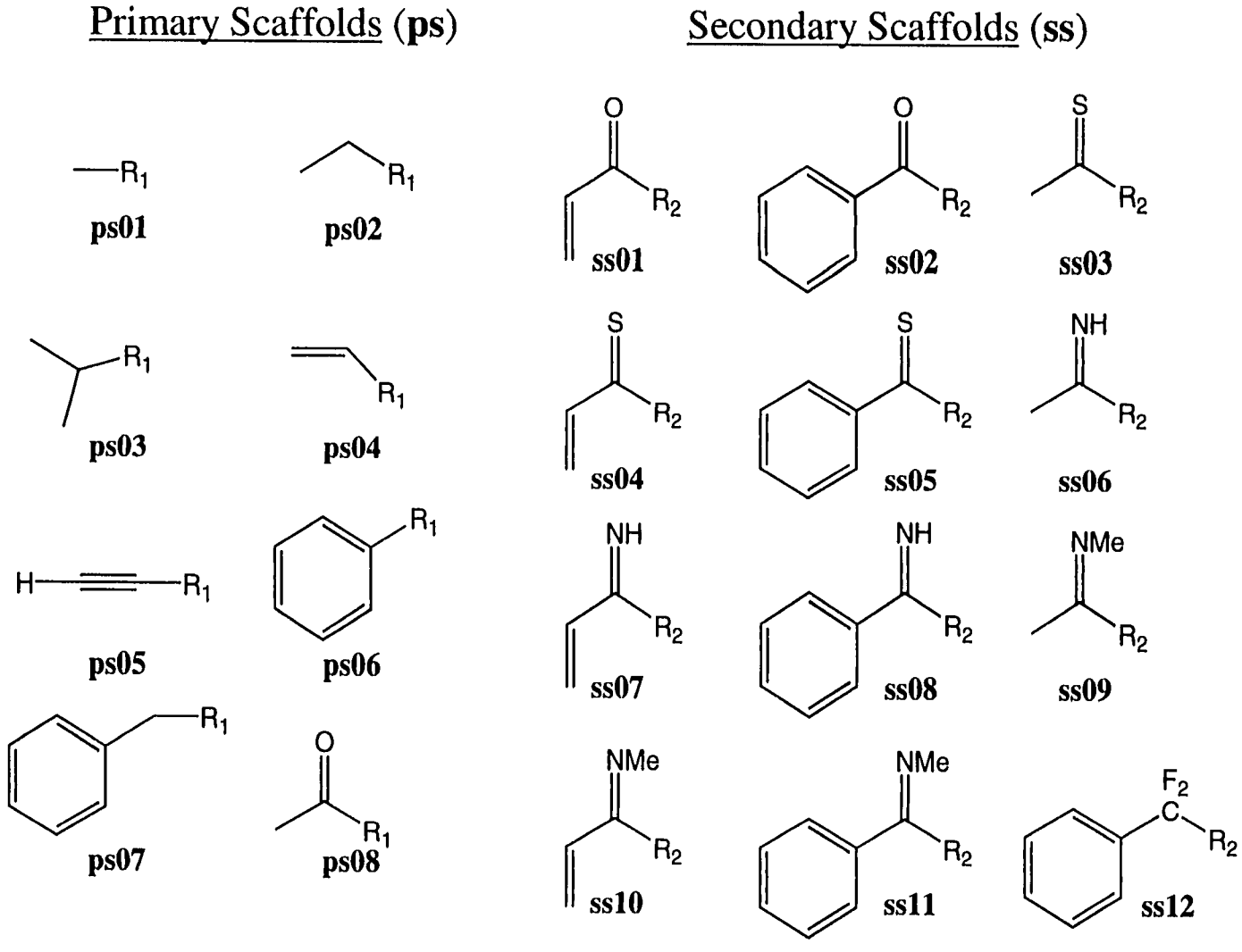


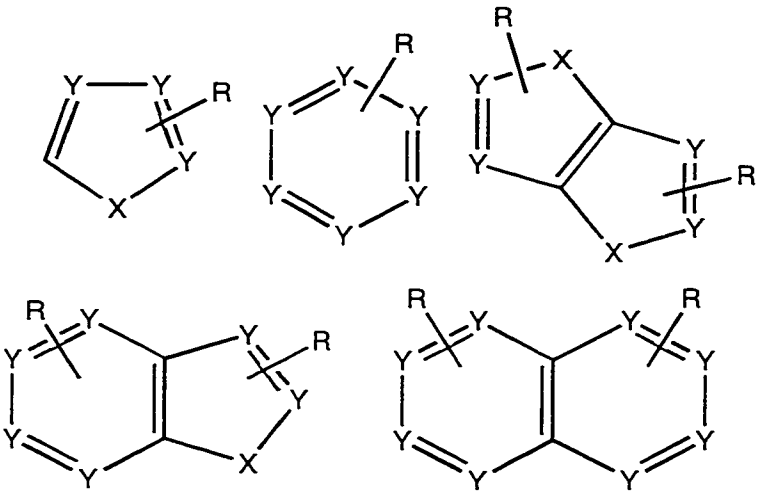
Figure 5.1. Primary and secondary functional groups. See appendix C for details of structures.



R_1 = primary functional group (**pf**), secondary functional group (**sf**) R_2 = primary functional group (**pf**)

Figure 5.2. Primary and secondary scaffolds. See appendix C for details of structures.

Although the above scheme sampled many bond-types, additional molecules were needed to sample cyclic and fused bicyclic heteroaryl bond-types. The building schemes for these molecules are shown in Figure 5.3, which generated a subset of 723 molecules. To further improve sampling and extend the variety of bond-types, the MMFF94 validation suite of 761 molecules was obtained from the Computational Chemistry List[60] (CCL) and incorporated into the training set. To complete the training set (see below for criteria of completeness), approximately 900 extra molecules were created. The complete training set of 2755 molecules is listed in appendix C.



X = O, S, N
Y = N, C
R = H, Me, Ph, NH_2, NHMe, NMe_2

Figure 5.3. Construction of cyclic and fused bicyclic heteroaryl systems. See appendix C for details of structures.

The robustness of the bond-typing scheme and the completeness of the training set was examined by using the current bond-typing schemes to assign bond-types to our training set, The Merck Index[61] of > 10000 molecules, and the NCI database[62] of ca.

250000 molecules. The bond-types generated from the training set were compared to the bond-types generated from The Merck Index and the NCI database. If bond-types were present in either the Index or the database and not in the training set, then a minimum of four diverse molecules were created to sample the missing bond-type. Consequently, the 354 BCCs that were derived from the training set of 2755 molecules using the current bond-typing scheme are able to charge any molecule in either Index or database of over 260000 molecules except for molecules containing boron or molecules that contain covalently bound metal atoms.

5.3 Atom-Types and Bond-Types

Keeping the atom- and bond-typing scheme simple, robust, and parsimonious was a goal throughout the development of the atom- and bond-types to simplify the fitting the process, decrease the degrees of freedom, and provide a robust method for users to bond-type any organic molecule. The development of the atom-types and bond orders began with a simple classification: the atom-type of an atom in a molecule was the same as the element name of that atom i.e. all carbon atoms were *C* atom-types, all oxygen atoms were *O* atom-types, etc. Along with the 11 atom-types (*C*, *N*, *O*, *F*, *S*, *P*, *Cl*, *Br*, *I*, *Si*, and *H*), seven bond orders were used: single, double, triple, aromatic single, aromatic double, dative, and delocalized, however there are six *unique* bond orders because the BCC values of bond-types containing aromatic single and aromatic double bond orders are equivalenced. Both bond orders are kept to facilitate labeling aromatic molecules.

Preliminary studies indicated that an atom-type classification according to hybridization states was required, and in some cases a further level of sub-classification according to bonded environments (within the same hybridization state). Carbon atom-types were classified as C_4, $C_3^{=C}$, $C_3^{=N,P}$, $C_3^{=O,S}$, $C_{1,2}$, C_{ar}, and C_{ar}^{lp} where the subscripts refer to the number of bonded neighbours (e.g. a carbon atom in a molecule of methanol would be classified as C_4), $C_{1,2}$ refers to a carbon atom with one or two bonded neighbours (e.g. isonitriles and ketenes, respectively), and C_{ar} signifies an aromatic carbon atom. The superscripts refer to bonded environments (i.e. $=C$ signifies double bonded to a carbon atom; $=N,P$ signifies double bonded to a nitrogen or phosphorus atom; $=O,S$ signifies double bonded to an oxygen or sulfur atom; and *lp* signifies adjacent to an aromatic nitrogen or oxygen atom bearing a lone pair of electrons). The reason for two aromatic carbon types was because $C_{ar}-H$ bond-types differed in their BCC value if they were adjacent to an aromatic atom containing a lone pair (i.e. N, O). For example, the BCC values for the $C_{ar}-H$ bond-types in benzene and pyridine were 0.0128 and 0.1366, respectively. This phenomenon is also observed in ESP-fit methods, for example, the H_6 hydrogen atom in imidazole (see Figure 4.1 **36**) has a smaller charge than the H_7 hydrogen atom (see the RESP column in Table 4.4). This phenomenon is termed as "atomic charge compensation" (ACC). The H_6 hydrogen atom in imidazole is compensating for the lack of a point charge (or bond-type in the case of BCC fitting) needed to describe the nitrogen lone pair[63] in the plane of the molecule. In general, ACC occurs when a limited number of point charges are used to describe an ESP that is varying too quickly over the ESP surface (e.g. in this case the ESP due to the lone pair of

electrons). In this work, ACC is allowed by differentiating two aromatic carbon atom-types: C_{arom} and C^{lp}_{arom}. Although aromatic sulfur and phosphorus atoms also have lone pairs, it was found that the BCC value of the bond-type $C_{ar}-H$ did not differ significantly when adjacent to an aromatic sulfur or phosphorus atom containing a lone pair, therefore ACC was not required for $C_{ar}-H$ bond-types adjacent to these atoms.

Nitrogen atom-types were classified as $N_{2^-.3.4^+}$, N_3^{deloc}, N_3^{hdeloc}, N_2, and $N_{1.2^-}$, where the subscripts represent the number of bonded neighbours and the plus or minus signs represent a charged atom (e.g. $N_{2^-.3.4^+}$ may represent either an anion nitrogen atom with two bonded neighbours, an amine nitrogen atom with three bonded neighbours, or a cation nitrogen atom with four bonded neighbours). The superscripts *deloc* and *hdeloc* signify delocalized and highly-delocalized, respectively. The sub-classification of the N_3 into N_3^{deloc} and N_3^{hdeloc} was needed to describe the different levels of delocalization of the lone pair of electrons of a nitrogen atom in conjugated systems. These degrees of delocalization were established in consideration of the different BCC corrections needed for partial delocalization (*deloc*) of the lone pair of electrons on an amide nitrogen vs. the complete delocalization (*hdeloc*) of the lone pair of electrons on the nitrogen atom in an aromatic ring (e.g. pyrrole).

Oxygen atom-types were classified as $O_{1.2}$, $O_1^{ester.acid}$, and O_1^{lact} where the superscripts *ester* and *acid* denote an ester and acid functional groups and *lact* signifies either a lactone or lactam. It was found that different carbonyl oxygen atom-types were

needed for *cis*- and *trans*-esters. Thus, as with carbon and nitrogen, a sub-classification beyond hybridization states was needed, in this case to separate *trans*-esters and amides from *cis*-esters and carboxylic acids. The difference might be caused by ACC or the different electronic characters of the delocalized lone pairs of electrons between the *cis* and *trans* conformations.

Phosphorus atom-types were classified simply as $P_{2,3}$, and $P_{3,4}^{=}$ where the superscript = signifies a double bond. Sulfur atom-types were classified as $S_{1,2}$, S_3, and S_4. It was not necessary to further divide the monovalent atoms (i.e. the halogens and hydrogen) because the sub-classifications of the C, N, O, S, and P atom-types provided enough degrees of freedom to fit the BCCs to the data.

The process of assigning bond-types in a molecule begins with assigning aromaticity to atoms using Fig. 5.4a, starting with case 1 and progressing through case 5 allowing redefinition by each case. Note that case 5 demonstrates that a five-membered unsaturated ring is considered "aromatic" if it is not fused to a six-membered "aromatic" ring. The five-membered ring in these types of systems are not considered to be aromatic within the AM1-BCC method (e.g. the three atoms in the five-membered ring of indole are not considered to be aromatic). Bond orders are then assigned using Fig. 5.4b followed by the atom-types using Fig. 5.5. The bond-types can then be constructed by combining atom-type *I*, the bond order between atoms *i* and *j*, and atom-type *J*. The BCC values for the bond-types are listed in Table 5.1. This atom- and bond-typing scheme is robust and general and can be used to unambiguously assign atom-types in an automated

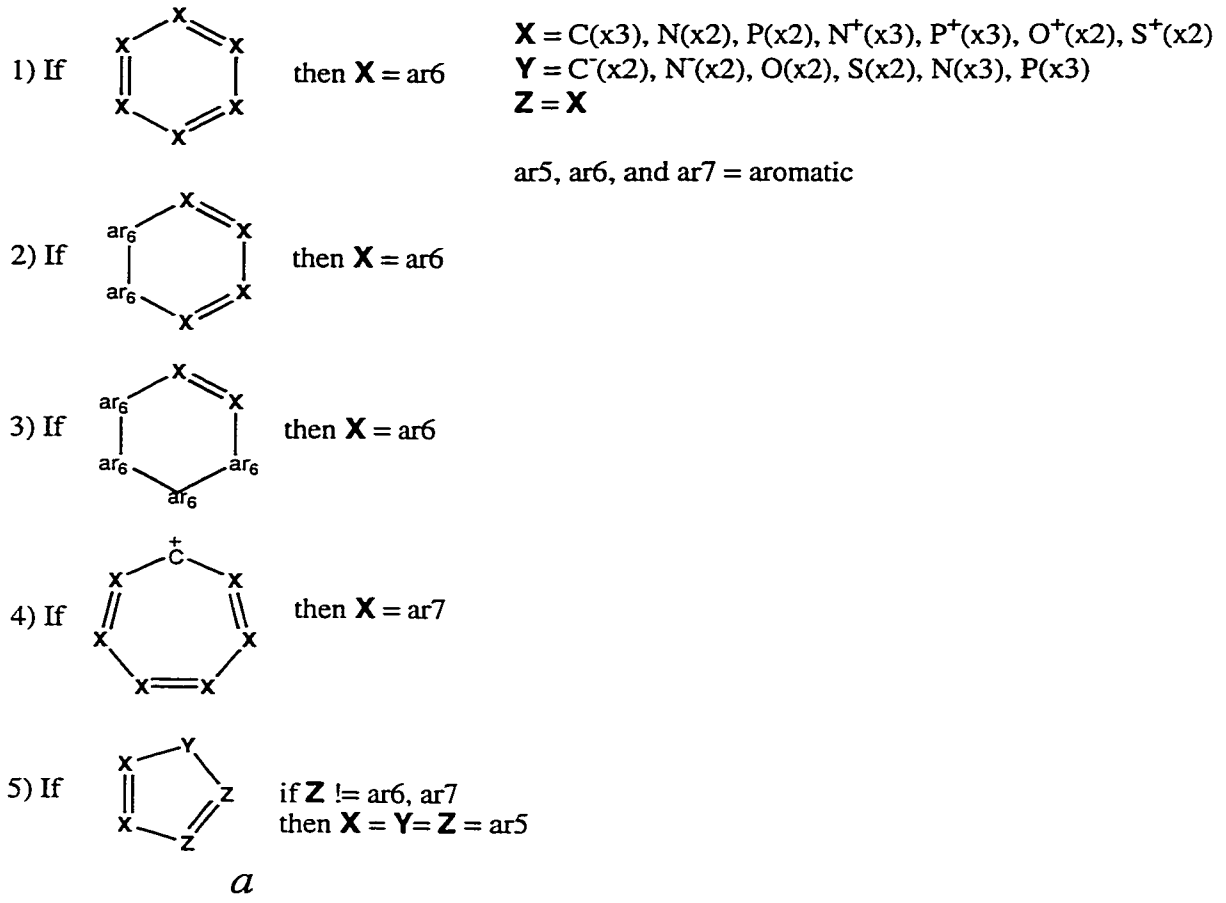
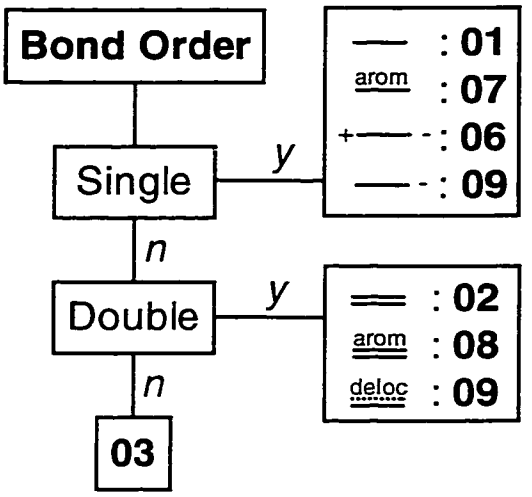


Figure 5.4. a) Definition of aromaticity. An atom is considered "aromatic" if one of the 5 cases is true. The last case (i.e. case 5) illustrates that a five-membered unsaturated ring is considered "aromatic" if it is not fused to a six-membered "aromatic" ring. The symbol "!=" signifies "not equal".



b

Figure 5.4. (continued) b) Definition of the bond order codes: 01 signifies single bond, 02 signifies double bond, 03 signifies triple bond, 06 signifies dative bond (e.g. N-oxide), 07 signifies aromatic single bond, 08 signifies aromatic double bond, 09 signifies single bond with charge (e.g. methoxide, sulfoxide, etc.) or delocalized bond (e.g. nitro or carboxy and their sulfur analogues).

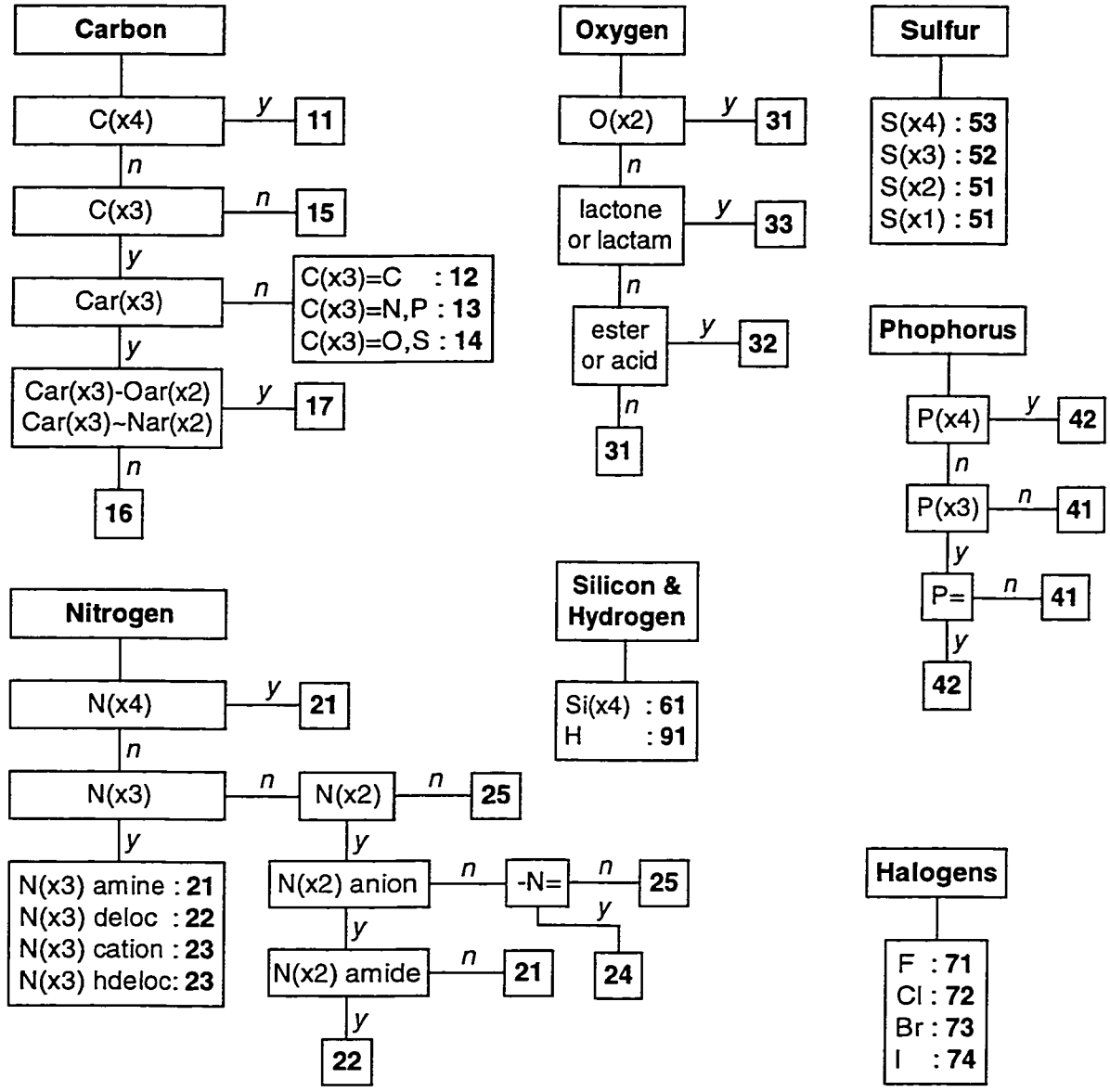
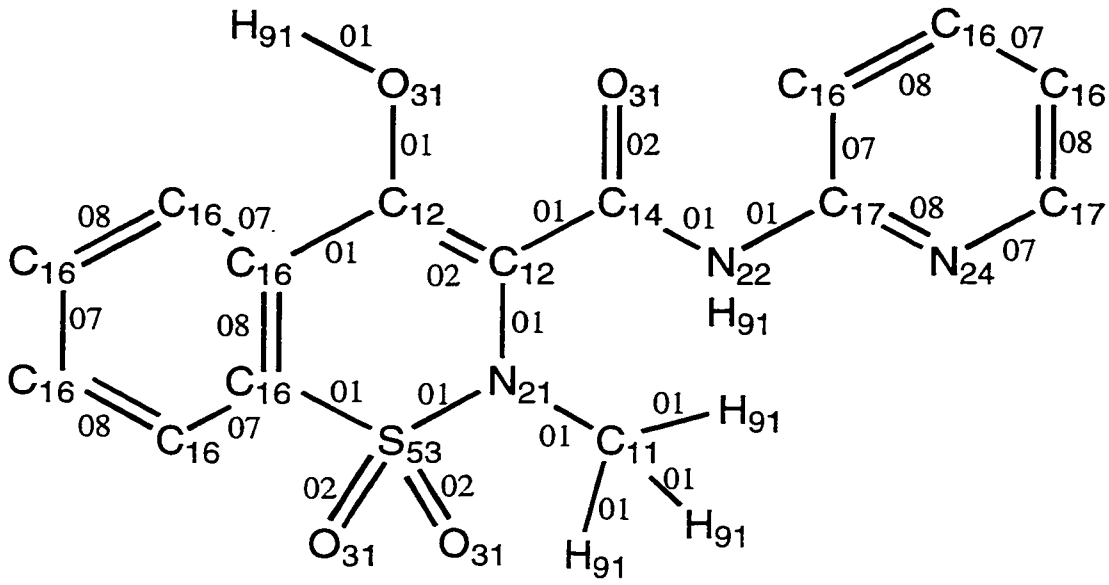


Figure 5.5. The atom-typing definitions. The atom-type of an atom in a molecule is determined by starting at the top of each atom-tree and answering the questions in the boxes until the atom code is reached. Definitions: "," signifies or; "~" signifies aromatic single or double bond; "-" single bond; "=" signifies double bond; "x" signifies number of bonded neighbours (e.g. N(x3) signifies a nitrogen atom with three bonded neighbours); "ar" signifies aromatic; "Oar" signifies oxygen atom in an aromatic ring bearing lone pairs; "Nar" signifies nitrogen atom in aromatic ring bearing a lone pairs (e.g. in pyridine); "deloc" signifies delocalized lone pair (e.g. amide nitrogen atoms); "hdeloc" signifies highly delocalized lone pair (e.g. the nitrogen atom in indole, pyrrole, or nitro group); "n" indicates no; "y" indicates yes.

fashion, e.g. using PATTY[64], a programmable atom-typer. There are a total of 26 atom-types and 6 unique bond orders.

An example of this atom- and bond-typing process is shown for piroxicam, a non-steroidal anti-inflammatory drug (Figure 5.6a). The atoms in the two outer 6-membered cyclic rings are assigned to be aromatic while the atoms in the middle 6-membered cyclic ring and all other atoms in the molecule are assigned to be non-aromatic according to Fig. 5.4a. The aromatic bond order codes (07 and 08 for aromatic single and double, respectively) are assigned to the two outer aromatic rings while non-aromatic bond order codes (01 and 02 for single and double, respectively) are assigned to the non-aromatic bonds according to Figure 5.4b. The atom-types are then determined using Fig. 5.5. Using the amide carbon atom in piroxicam (Figure 5.6a) as an example, the questions in the boxes beginning at the top of the "Carbon" tree in Figure 5.5 are answered. Does the amide carbon have four bonded neighbors? Answering no leads to the next question and the process is repeated until an answer is found. In this case, the amide carbon atom-type is *C(x3)=O,S*, coded **14**. This process is repeated for all the atoms in the molecule. The atom-types and bond order codes are shown for piroxicam in Figure 5.6a and the bond-types are shown in Figure 5.6b. The bond-types are formed for each bond in the molecule by combining atom-type *I*, the bond order between atoms *i* and *j*, and atom-type *J*. The BCC values for each bond-type in Figure 5.6b can be found in Table 5.1.



a

11 01 21	12 01 21	16 01 53
11 01 91	12 01 31	16 01 91
12 01 14	12 01 91	17 01 22
12 01 16	14 01 22	17 01 91
21 01 53	31 02 53	16 08 16
31 01 91	16 07 16	16 08 17
12 02 12	16 07 17	17 08 24
14 02 31	17 07 24	

b

Figure 5.6. a) Atom typing of piroxicam, a non-steroidal anti-inflammatory drug. The atom labels represent the atom-types and the bond labels represent the bond orders (see figures 5.4 and 5.5 for atom- and bond order definitions). b) The bond-types needed to charge piroxicam. The first pair of digits represents atom-type *I*, the second pair represent the bond order, and the third pair represent atom-type *J*. The bond charge corrections (BCCs) for each bond-type can be found in Table 5.1.

5.4 The Bond Charge Corrections (BCCs)

Applying the above atom- and bond-typing schemes (i.e. Figures 5.4 and 5.5) to The Merck Index and the NCI database showed that only 354 BCCs would cover the desired chemical variability. These BCCs were fitted to the ESP of a training set of 2755 organic molecules (for fitting methodology see chapter 3). The geometry of each molecule was initially optimized at the AM1 level with the MOPAC-6 program (vs. MMFF optimization in the proof-of-concept chapter). Note that it is important to avoid over-polarizing atomic centers when optimizing a molecular geometry because the AM1 atomic charges are sensitive to polarization. The term "over-polarization" in this context refers to the charge separation that occurs during non-covalent interactions between highly polar atoms (e.g. a carbonyl oxygen and a hydrogen). The major source of over-polarization in a vacuum phase geometry optimization occurs from the formation of internal hydrogen-bonds. Those atoms involved (the hydrogen-bond donors and acceptors) will have different AM1 atomic charges than the same atoms not involved in hydrogen-bonds. Since the BCCs are fit to non-polarized molecules, they cannot properly correct AM1 atomic charges that are polarized. The QM ESP of each molecule was then evaluated at grid points around the molecule using GAUSSIAN-92 and the difference between the QM and AM1 ESPs was calculated at the same set of grid point as described in Chapter 3. Out of the 354 BCCs, 29 of them are zero by construction because of equivalent atom-types in the bond (e.g. $C_4 - C_4$, bond code 11 01 11), and 16 of them were equivalenced during the fit to ensure symmetry in aromatic molecule (e.g. C_{ar}^{lp} *arom* N_2 and C_{ar}^{lp} *arom* N_2, bond codes 17 07 24 and 17 08 24, respectively). Therefore, 309 unique BCCs were fitted to the 44194 bonds in the training set of 2755

molecules. Table 5.1 presents the bond-types, the BCC values and the number of times they occur in the training set. The first pair of digits in the bond-types represents atom-type *I*, the second pair represent the bond order, and the third pair represent atom-type *J*. (see Figures 5.4 and 5.5 for atom- and bond-typing definitions).

The resulting set of BCCs was subjected to a validation set of dimer and free energy of solvation calculations that resulted in *a posteriori* adjustment of five BCCs to improve the relative free energies of solvation of amines, nitros, and some hydrocarbons. These adjustments will be discussed in detail in section 6.3.1.

Table 5.1.

Table of Bond Codes, Number of Occurrences in the Training Set and the BCC Values.

	Type	Occur.	BCC		Type	Occur.	BCC		Type	Occur.	BCC
1	11 01 11	1876	0.0000	46	12 01 74	4	0.2724	91	15 01 21	10	0.0554
2	11 01 12	262	0.0050	47	12 01 91	1093	0.0000	92	15 01 22	4	0.0047
3	11 01 13	132	-0.0762	48	13 01 13	26	0.0000	93	15 01 23	5	-0.0650
4	11 01 14	372	-0.0486	49	13 01 14	26	0.0611	94	15 01 24	4	0.0524
5	11 01 15	103	-0.0344	50	13 01 15	5	0.0305	95	15 01 25	4	0.0810
6	11 01 16	238	-0.0022	51	13 01 16	81	0.0833	96	15 01 31	5	0.0307
7	11 01 17	85	-0.0936	52	13 01 17	15	-0.0299	97	15 01 41	5	0.3982
8	11 01 21	860	0.1582	53	13 01 21	114	0.1528	98	15 01 42	5	0.3709
9	11 01 22	164	0.0380	54	13 01 22	13	0.0060	99	15 01 51	4	0.2242
10	11 01 23	244	-0.0186	55	13 01 23	68	0.0382	100	15 01 52	5	0.4584
11	11 01 24	117	0.1260	56	13 01 24	27	0.2191	101	15 01 53	5	0.4020
12	11 01 25	12	0.1822	57	13 01 25	4	0.2384	102	15 01 61	7	0.2357
13	11 01 31	750	0.0725	58	13 01 31	48	0.1342	103	15 01 71	4	0.0041
14	11 01 41	96	0.2583	59	13 01 41	6	0.3519	104	15 01 72	4	0.0917
15	11 01 42	131	0.4387	60	13 01 42	5	0.5282	105	15 01 73	4	0.2348
16	11 01 51	173	0.1826	61	13 01 51	41	0.2249	106	15 01 74	4	0.2892
17	11 01 52	82	0.4269	62	13 01 52	5	0.4293	107	15 01 91	32	0.0577
18	11 01 53	110	0.4612	63	13 01 53	7	0.5077	108	16 01 16	35	0.0000
19	11 01 61	117	0.1552	64	13 01 71	7	0.1063	109	16 01 17	44	-0.1074
20	11 01 71	123	0.0770	65	13 01 72	14	0.1128	110	16 01 21	160	0.0820
21	11 01 72	147	0.0740	66	13 01 73	6	0.1262	111	16 01 22	17	-0.0076
22	11 01 73	36	0.1276	67	13 01 74	4	0.2765	112	16 01 23	140	-0.0455
23	11 01 74	4	0.3014	68	13 01 91	211	0.1267	113	16 01 24	73	0.1400
24	11 01 91	10885	0.0403	69	14 01 14	21	0.0000	114	16 01 25	5	0.1853
25	12 01 12	116	0.0000	70	14 01 15	4	-0.0317	115	16 01 31	105	0.0456
26	12 01 13	142	-0.0861	71	14 01 16	86	0.0201	116	16 01 41	23	0.3264
27	12 01 14	187	-0.0217	72	14 01 17	15	-0.0900	117	16 01 42	20	0.3998
28	12 01 15	51	0.0345	73	14 01 22	352	0.0668	118	16 01 51	48	0.2270
29	12 01 16	62	0.0112	74	14 01 23	12	-0.0424	119	16 01 52	7	0.4278
30	12 01 17	27	-0.1104	75	14 01 24	29	0.1389	120	16 01 53	51	0.4446
31	12 01 21	77	0.0860	76	14 01 25	5	0.0039	121	16 01 61	5	0.1943
32	12 01 22	24	0.0021	77	14 01 31	208	0.0904	122	16 01 71	7	0.0356
33	12 01 23	42	-0.0279	78	14 01 41	8	0.2814	123	16 01 72	24	0.0785
34	12 01 24	49	0.1364	79	14 01 42	11	0.4876	124	16 01 73	4	0.1403
35	12 01 25	5	0.4879	80	14 01 51	39	0.1823	125	16 01 74	4	0.2861
36	12 01 31	60	0.0437	81	14 01 52	5	0.4308	126	16 01 91	4215	0.0000
37	12 01 41	4	0.3056	82	14 01 53	8	0.4828	127	17 01 17	4	0.0000
38	12 01 42	17	0.4241	83	14 01 71	6	0.1112	128	17 01 21	160	0.1934
39	12 01 51	48	0.2405	84	14 01 72	8	0.0482	129	17 01 22	7	0.0803
40	12 01 52	11	0.4306	85	14 01 73	6	0.1103	130	17 01 23	35	0.0634
41	12 01 53	27	0.4505	86	14 01 74	4	0.2595	131	17 01 24	25	0.2197
42	12 01 61	5	0.2130	87	14 01 91	60	0.0924	132	17 01 25	5	0.2107
43	12 01 71	9	0.0620	88	15 01 15	10	0.0000	133	17 01 31	19	0.1703
44	12 01 72	38	0.0838	89	15 01 16	16	0.0034	134	17 01 41	5	0.4597
45	12 01 73	10	0.1252	90	15 01 17	6	-0.1083	135	17 01 42	4	0.4603

(Continues on next page)

Table 5.1
(continued)

	Type	Occur.	BCC		Type	Occur.	BCC		Type	Occur.	BCC
136	17 01 51	19	0.3267	181	23 01 53	5	0.3448	226	42 01 71	6	-0.2107
137	17 01 52	5	0.5416	182	23 01 61	5	0.1775	227	42 01 72	15	-0.3165
138	17 01 53	4	0.5166	183	23 01 71	4	0.0570	228	42 01 73	4	-0.2500
139	17 01 71	9	0.1403	184	23 01 72	7	0.0647	229	42 01 91	25	-0.3943
140	17 01 72	8	0.1659	185	23 01 73	4	0.1809	230	51 01 51	15	0.0000
141	17 01 73	16	0.2308	186	23 01 91	362	-0.0487	231	51 01 52	5	0.2577
142	17 01 74	4	0.3512	187	24 01 24	13	0.0000	232	51 01 53	4	0.2971
143	17 01 91	991	0.1369	188	24 01 25	4	0.1017	233	51 01 61	4	-0.0933
144	21 01 21	31	0.0000	189	24 01 31	34	-0.0983	234	51 01 71	4	-0.0543
145	21 01 22	16	-0.0305	190	24 01 41	6	0.1275	235	51 01 72	4	-0.0270
146	21 01 23	52	-0.1195	191	24 01 42	17	0.1675	236	51 01 73	4	0.1116
147	21 01 24	16	0.0865	192	24 01 51	19	0.0311	237	51 01 91	36	-0.1712
148	21 01 31	48	-0.0553	193	24 01 52	8	0.1723	238	52 01 52	1	0.0000
149	21 01 41	13	0.2522	194	24 01 53	14	0.2721	239	52 01 72	5	-0.2350
150	21 01 42	24	0.3084	195	24 01 61	4	0.1157	240	52 01 91	10	-0.4037
151	21 01 51	7	0.0755	196	24 01 71	4	-0.0554	241	53 01 71	7	-0.2040
152	21 01 52	11	0.3530	197	24 01 72	9	-0.0601	242	53 01 72	9	-0.2760
153	21 01 53	82	0.3354	198	24 01 73	5	0.0194	243	53 01 73	4	-0.1970
154	21 01 61	10	0.1033	199	24 01 74	4	0.1172	244	53 01 74	4	-0.1494
155	21 01 71	5	-0.0207	200	24 01 91	72	-0.2506	245	53 01 91	6	-0.4591
156	21 01 72	4	-0.0363	201	25 01 91	4	-0.2626	246	61 01 72	20	-0.0556
157	21 01 73	6	0.0870	202	31 01 31	6	0.0000	247	61 01 91	45	-0.0263
158	21 01 74	4	0.1889	203	31 01 41	18	0.2226	248	12 02 12	613	0.0000
159	21 01 91	878	-0.2048	204	31 01 42	196	0.2446	249	12 02 15	23	-0.0641
160	22 01 22	5	0.0000	205	31 01 51	5	0.0445	250	13 02 23	39	0.0540
161	22 01 23	4	-0.0332	206	31 01 52	9	0.2150	251	13 02 24	432	0.2884
162	22 01 24	11	0.1331	207	31 01 53	41	0.2094	252	13 02 25	13	0.1609
163	22 01 25	4	0.0214	208	31 01 61	32	0.0087	253	13 02 41	27	0.2490
164	22 01 31	10	-0.0060	209	31 01 71	4	0.0378	254	13 02 42	10	0.7081
165	22 01 41	9	0.1650	210	31 01 72	4	0.0019	255	14 02 31	301	0.2368
166	22 01 42	10	0.3188	211	31 01 73	4	0.0710	256	14 02 32	115	0.1891
167	22 01 51	4	0.1593	212	31 01 74	4	0.1723	257	14 02 33	217	0.2758
168	22 01 52	4	0.3657	213	31 01 91	386	-0.2017	258	14 02 51	88	0.2833
169	22 01 53	12	0.3322	214	41 01 41	4	0.0000	259	14 02 52	4	0.4445
170	22 01 71	4	-0.0157	215	41 01 42	6	0.2836	260	14 02 53	4	0.5622
171	22 01 72	4	-0.0022	216	41 01 51	5	-0.1100	261	15 02 15	6	0.0000
172	22 01 73	6	0.0836	217	41 01 61	6	0.0594	262	15 02 24	13	0.2201
173	22 01 74	4	0.2067	218	41 01 71	6	-0.2772	263	15 02 31	14	0.2338
174	22 01 91	213	-0.0866	219	41 01 72	8	-0.2065	264	15 02 42	4	0.7405
175	23 01 23	6	0.0000	220	41 01 73	4	-0.1311	265	15 02 51	6	0.3435
176	23 01 24	31	0.1801	221	41 01 74	4	0.0182	266	15 02 52	4	0.3054
177	23 01 31	11	0.0205	222	41 01 91	49	-0.1598	267	23 02 24	5	0.1740
178	23 01 41	6	0.2545	223	42 01 42	1	0.0000	268	24 02 24	47	0.0000
179	23 01 42	5	0.4322	224	42 01 51	10	-0.3747	269	24 02 25	12	0.0053
180	23 01 51	5	0.2031	225	42 01 52	4	0.0590	270	24 02 31	12	-0.0470

(Continues on next page)

Table 5.1
(continued)

	Type	Occur.	BCC		Type	Occur.	BCC		Type	Occur.	BCC
271	24 02 41	7	0.0534	299	16 07 51	178	0.2699	327	17 08 24	1003	0.2633
272	24 02 42	22	0.4572	300	17 07 17	97	0.0000	328	17 08 31	4	0.1294
273	24 02 51	6	0.1684	301	17 07 23	244	0.0878	329	17 08 41	6	0.2467
274	24 02 52	4	0.3751	302	17 07 24	813	0.2633	330	17 08 51	4	0.3139
275	24 02 53	5	0.4251	303	17 07 31	265	0.1294	331	23 08 24	4	0.1708
276	25 02 25	25	0.0000	304	17 07 41	4	0.2467	332	24 08 24	226	0.0000
277	31 02 41	4	0.1327	305	17 07 51	126	0.3139	333	24 08 41	8	0.0455
278	31 02 42	95	0.2890	306	23 07 23	4	0.0000	334	11 09 31	5	0.1632
279	31 02 51	5	0.1587	307	23 07 24	227	0.1708	335	11 09 51	4	0.0836
280	31 02 52	74	0.2719	308	23 07 31	4	0.0680	336	12 09 31	7	0.1839
281	31 02 53	385	0.2790	309	23 07 41	8	0.1809	337	12 09 51	5	0.1874
282	41 02 41	1	0.0000	310	23 07 51	4	0.2535	338	13 09 31	5	0.3052
283	42 02 51	14	-0.5805	311	24 07 24	87	0.0000	339	13 09 51	4	0.1961
284	51 02 52	7	0.3300	312	24 07 31	121	-0.0724	340	14 09 31	56	0.2655
285	15 03 15	88	0.0000	313	24 07 41	5	0.0455	341	14 09 51	23	0.1711
286	15 03 25	160	0.3240	314	24 07 51	131	0.0419	342	15 09 31	4	0.2553
287	25 03 25	17	0.0000	315	31 07 41	4	0.0711	343	15 09 51	4	0.3283
288	21 06 31	6	-0.0586	316	31 07 51	4	0.0984	344	16 09 31	4	0.2798
289	22 06 31	4	0.0461	317	41 07 51	4	0.0587	345	16 09 51	4	0.2542
290	23 06 31	7	0.1330	318	51 07 51	3	0.0000	346	17 09 31	4	0.3369
291	24 06 31	4	-0.0814	319	16 08 16	2551	0.0000	347	17 09 51	4	0.3475
292	25 06 31	4	-0.0527	320	16 08 17	504	-0.0652	348	23 09 31	245	-0.1500
293	25 06 51	4	0.1485	321	16 08 23	47	-0.0249	349	31 09 41	7	0.2364
294	16 07 16	2351	0.0000	322	16 08 31	4	-0.1518	350	31 09 42	108	0.3171
295	16 07 17	413	-0.0652	323	16 08 41	14	0.2237	351	31 09 51	4	0.1525
296	16 07 23	535	-0.0249	324	16 08 51	7	0.2699	352	31 09 52	8	0.3896
297	16 07 31	4	-0.1518	325	17 08 17	129	0.0000	353	31 09 53	30	0.3229
298	16 07 41	11	0.2237	326	17 08 23	11	0.0878	354	51 09 53	4	0.5217

6 Validation of AM1-BCC Charge Model

6.1 General Considerations

The goal of this chapter is to examine the ability of the AM1-BCC charge model to reproduce experimental or high-level *ab initio* data of non-bonded interactions between organic molecules. But first, the molecules used in the proof-of-concept section are revisited in order to investigate the differences between BCCs derived from the small training set (i.e. the training set used in Chapter 4; **TS1**, **TS2**, and **TS3**) and the BCCs derived from the training set used for global parameterization (i.e. the training set containing 2755 molecules). The AM1-BCC charge model is then used in the Cornell *et al.*[15] FF to test the ability to reproduce within 1 kcal/mol i) homo- and hetero-dimer energies of a set of diverse small organic molecules and ii) experimental relative free energies of solvation of small organic molecules. Conformational energies were not included in the validation set because the FF torsion parameters are dependent on the charges through the 1,4 electrostatic term; considering that the Cornell *et al.* FF torsion parameters were originally parameterized with the RESP charge model, the use of any other charge model would yield inaccurate conformations.

6.2 Differences Between the Proof-of-Concept Chapter and the Final AM1-BCC Protocol.

The final AM1-BCC protocol (i.e. the atom- and bond-types, geometry optimization before calculation of the AM1 atomic charges, and the BCCs) is different than the protocol used in the proof-of-concept chapter.

Firstly, for practical purposes, the geometry of molecules is optimized at the AM1 level before the AM1 atomic charges are calculated (MMFF was used in the proof-of-concept chapter to optimize the molecular geometries). This allows for a more robust AM1 charging process by providing a common geometry so that two different users will obtain the same AM1 atomic charges for the same molecule. Also, using the AM1 method for both geometry optimization and atomic charge calculation reduces the number of steps. As discussed in section 5.4, it is important to avoid over-polarizing atomic centers (i.e. avoiding internal hydrogen-bonds) when optimizing a molecular geometry due to AM1's sensitivity to polarization. The difference in AM1 atomic charges between MMFF and AM1 optimized geometries can be compared between Tables 4.3 and 6.1 for methanol, Tables 4.4 and 6.2 for imidazole, and Tables 4.5 and 6.3 for indole. In all three cases, the dipole moment decreases and the RMS with respect to the QM ESP increases with the AM1 optimized geometries, albeit by small amounts. This decrease in quality is more than compensated by the gain in robustness obtained by using the AM1 method to optimize the geometries.

Secondly, the atom- and bond-typing schemes have changed compared to the preliminary study of Chapter 4 in order to consider the larger variety of bond-types found in the global training set. Also, in some cases (e.g. aromatic carbon atom-types, sulfur atom-types, etc.) extra atom-types have been added to enable a better description of bond-types (see section 5.3).

Lastly, but most important, the BCCs themselves have changed between the training set containing 45 molecules and the training set containing 2755 molecules (cf. Tables 4.2 and 5.1).

With these new changes, the six molecules in the proof-of–concept chapter (i.e. methanol, imidazole, indole, D-glucose, aspirin, and eriodictyol) are revisited in Tables 6.1 to 6.6. Note that the RESP charges for methanol, imidazole, and indole have also changed as explained in the first paragraph of appendix D. In all six cases, the AM1-BCC dipole moments have decreased by a minimum of 0.0774 D for indole and a maximum of 0.3844 D for D-glucose. In all but two cases (methanol and eriodictyol), the RMS have increased by a maximum of 0.0228 a.u. for imidazole. This slight degradation of the AM1-BCC atomic charges is expected because the BCCs derived from the global parameterization reflect a consensus set of BCCs over the large training set and therefore must account for a larger variety of bond-types compared to the small training set.

TABLE 6.1.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Methanol for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	0.0416	-0.0733	0.1162
$H_{2,3,4}$	0.0564	0.0680	0.0287
O_5	-0.6308	-0.3260	-0.5988
H_6	0.4201	0.1954	0.3964
RMS (a.u.)	0.0414	0.0981	0.0390
Dipole moment (D)	2.1689	1.2875	2.0007

[a] Subscript refers to atom numbering for methanol (**14**) in Figure 4.1.

TABLE 6.2.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Imidazole for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	0.1248	-0.1065	0.3820
N_2	-0.4920	-0.1406	-0.6667
C_3	0.0945	-0.1743	0.2910
C_4	-0.3129	-0.1716	-0.2612
N_5	-0.2102	-0.2085	-0.3224
H_6	0.1288	0.1791	0.0422
H_7	0.2106	0.1761	0.1761
H_8	0.3120	0.2495	0.2992
H_9	0.1443	0.1967	0.0598
RMS (a.u.)	0.0332	0.1257	0.0534
Dipole moment (D)	3.8087	2.1289	3.8614

[a] Subscripts refer to atom numbering for imidazole (**36**) in Figure 4.1.

TABLE 6.3.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Indole for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	0.1919	-0.0839	-0.0957
C_2	0.1477	-0.0019	-0.0471
C_3	-0.2564	-0.1464	-0.1464
C_4	-0.1406	-0.1128	-0.1128
C_5	-0.1951	-0.1594	-0.1594
C_6	-0.2254	-0.0818	-0.0818
H_7	0.1588	0.1304	0.1304
H_8	0.1423	0.1280	0.1280
H_9	0.1451	0.1283	0.1283
H_{10}	0.1596	0.1330	0.1330
C_{11}	-0.3841	-0.1995	-0.1877
H_{12}	0.2024	0.1561	0.1561
C_{13}	-0.1272	-0.0817	-0.1088
H_{14}	0.1983	0.1632	0.1632
N_{15}	-0.3739	-0.2194	-0.1968
H_{16}	0.3567	0.2476	0.2973
RMS (a.u.)	0.0141	0.0485	0.0366
Dipole moment (D)	2.0336	1.2826	1.8223

[a] Subscript refers to atom numbering for indole (**41**) in Figure 4.1.

TABLE 6.4.

Charges, RMS Deviations from the QM ESP and Dipole Moments for D-Glucose for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	-0.0312	0.0140	0.1251
C_2	0.0580	-0.0295	0.0815
C_3	0.0367	-0.0266	0.0845
C_4	0.0605	-0.0120	0.0991
C_5	0.1743	0.1224	0.3052
O_6	-0.3237	-0.2948	-0.4384
H_7	0.1028	0.0852	0.0460
H_8	0.1345	0.0909	0.0516
H_9	0.1873	0.1459	0.1066
H_{10}	0.1793	0.1194	0.0801
H_{11}	0.1691	0.1458	0.1066
C_{12}	0.2253	-0.0399	0.1104
$H_{13, 14}$	0.0524	0.0948	0.0556
O_{15}	-0.6970	-0.3264	-0.5992
H_{16}	0.4431	0.2186	0.4196
O_{17}	-0.6157	-0.3243	-0.5970
H_{18}	0.4102	0.2104	0.4114
O_{19}	-0.6784	-0.3212	-0.5940
H_{20}	0.4345	0.2101	0.4111
O_{21}	-0.6568	-0.3317	-0.6045
H_{22}	0.4647	0.2327	0.4337
O_{23}	-0.6400	-0.3007	-0.5735
H_{24}	0.4589	0.2219	0.4229
RMS (a.u.)	0.0151	0.1109	0.0347
Dipole moment (D)	5.5880	3.2202	4.9043

[a] Subscript refers to atom numbering for D-glucose as shown in Figure 4.5*a*.

TABLE 6.5.
Charges, RMS Deviations from the QM ESP and Dipole Moments for Aspirin for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	-0.0036	-0.1163	-0.1369
C_2	0.2696	0.0988	0.1439
C_3	-0.2128	-0.1425	-0.1425
C_4	-0.1551	-0.0894	-0.0894
C_5	-0.1238	-0.1442	-0.1442
C_6	-0.2414	-0.0684	-0.0684
H_7	0.1832	0.1488	0.1488
H_8	0.1689	0.1404	0.1404
H_9	0.1504	0.1412	0.1412
H_{10}	0.1998	0.1553	0.1553
C_{11}	0.7064	0.3671	0.6668
O_{12}	-0.5755	-0.3440	-0.5330
O_{13}	-0.5993	-0.2947	-0.5858
H_{14}	0.4479	0.2402	0.4412
O_{15}	-0.4214	-0.2102	-0.3455
C_{16}	0.7796	0.3169	0.6460
O_{17}	-0.5420	-0.3101	-0.4991
C_{18}	-0.4466	-0.2262	-0.1584
$H_{19,20,21}$	0.1386	0.1124	0.0732
RMS (a.u.)	0.0108	0.0652	0.0300
Dipole moment (D)	1.9486	2.0204	1.9874

[a] Subscript refers to atom numbering for aspirin as shown in Figure 4.5*b*.

TABLE 6.6.

Charges, RMS Deviations from the QM ESP and Dipole Moments for Eriodictyol for Various Charge Models.

Atom[a]	RESP Fit Charges	AM1 Charges	AM1-BCC Charges
C_1	-0.3265	-0.3056	-0.3056
C_2	0.2688	0.1767	0.2219
C_3	-0.1891	-0.2398	-0.2398
C_4	0.0084	0.1843	0.2294
C_5	-0.0471	-0.3495	-0.3701
C_6	0.2116	0.2239	0.2690
H_7	0.1751	0.1679	0.1679
H_8	0.1900	0.1547	0.1547
O_9	-0.5587	-0.2556	-0.5018
H_{10}	0.4624	0.2639	0.4649
O_{11}	-0.5588	-0.2405	-0.4866
H_{12}	0.4386	0.2607	0.4317
O_{13}	-0.2666	-0.1867	-0.3037
C_{14}	0.5201	0.2907	0.6003
O_{15}	-0.5735	-0.3254	-0.5644
C_{16}	-0.2822	-0.2270	-0.1985
$H_{17, 18}$	0.1366	0.1323	0.0931
C_{19}	0.2264	0.0841	0.2025
H_{20}	0.0963	0.1019	0.0626
C_{21}	0.5928	-0.0886	-0.0959
C_{22}	-0.5958	-0.1020	-0.1020
C_{23}	0.4643	0.0665	0.1117
C_{24}	-0.0232	0.0094	0.0546
C_{25}	0.4643	-0.1814	-0.1814
C_{26}	-0.5958	-0.1132	-0.1132
H_{27}	-0.3494	0.1615	0.1615
H_{28}	-0.0754	0.1384	0.1384
H_{29}	0.2201	0.1357	0.1357
O_{30}	-0.6483	-0.2694	-0.5155
H_{31}	0.4710	0.2360	0.4370
O_{32}	-0.6274	-0.2452	-0.4913
H_{33}	0.4692	0.2389	0.4399
RMS (a.u.)	0.0252	0.0803	0.0331
Dipole moment (D)	4.0337	2.2686	3.8518

[a] Subscript refers to atom numbering for eriodictyol as shown in Figure 4.5*c*.

6.3 Dimer Energies

The binding energy between two static molecules can be calculated with the following equation:

$$E_{bind} = E_{AB} - (E_A + E_B) \tag{6.1}$$

where E_{AB} is the energy of the complex formed by molecules A and B, and E_A and E_B are the individual energies of molecules A and B, respectively, separated to infinity. A validation set of 47 organic dimers and 27 DNA dimers were chosen to test the ability of the charge models to reproduce *ab initio* non-covalent interactions between homo- and hetero-dimers representing molecules containing common organic functionalities. The validation criterion is to reproduce the *ab initio* dimer energies within 1 kcal/mol. A correct reproduction of these dimers is important in reproducing or predicting experimental results. All molecules were charged with the RESP, AM1, MMFF and AM1-BCC charge models with the exception of the TIP3P water. Initial geometries for the DNA dimers were reproduced from ref. [65], while the organic molecule dimers were obtained from the MMFF validation suite available from the Computational Chemistry List (CCL)[66]. See appendix D for details of the computational methods.

6.3.1 Results and Discussion

The hydrogen-bonded energies of diverse but simple organic homo- and hetero-dimers (see Figure 6.1 for representative structures) are compared between *ab initio* (HF/6-31G* and MP2/6-31+G** referred to as HF and MP2, respectively hereafter) and RESP, AM1, MMFF, and AM1-BCC charge models. The MMFF charge model performed best overall with a mean unsigned error of 0.74 kcal/mol with respect to MP2 even though the MMFF charge model was initially parameterized for a "buffered" vdW and electrostatic functional form[44], not the AMBER electrostatic functional form[16] in which it was used for this work. However, the good performance of MMFF must be interpreted in light of the fact that dimers **1** through **40** were part of the dimer parameterization set[45] of MMFF.

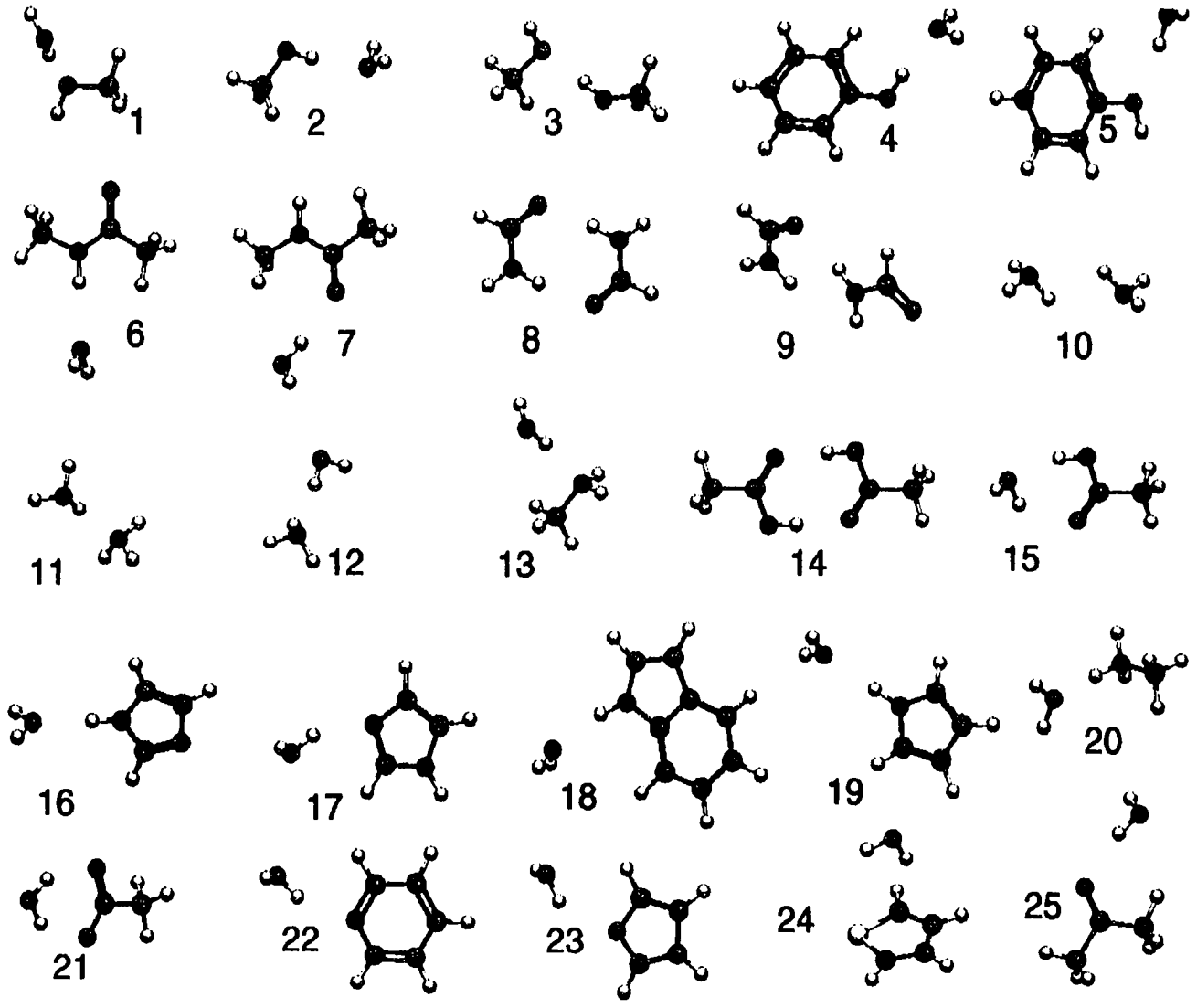


Figure 6.1 Representative structures for the hydrogen-bonded organic homo- and hetero-dimers. A description of the dimers can be found in the first column of Table 6.1 (continues on next page). Colour representations: green (C), white (H), red (O), blue (N), yellow (S).

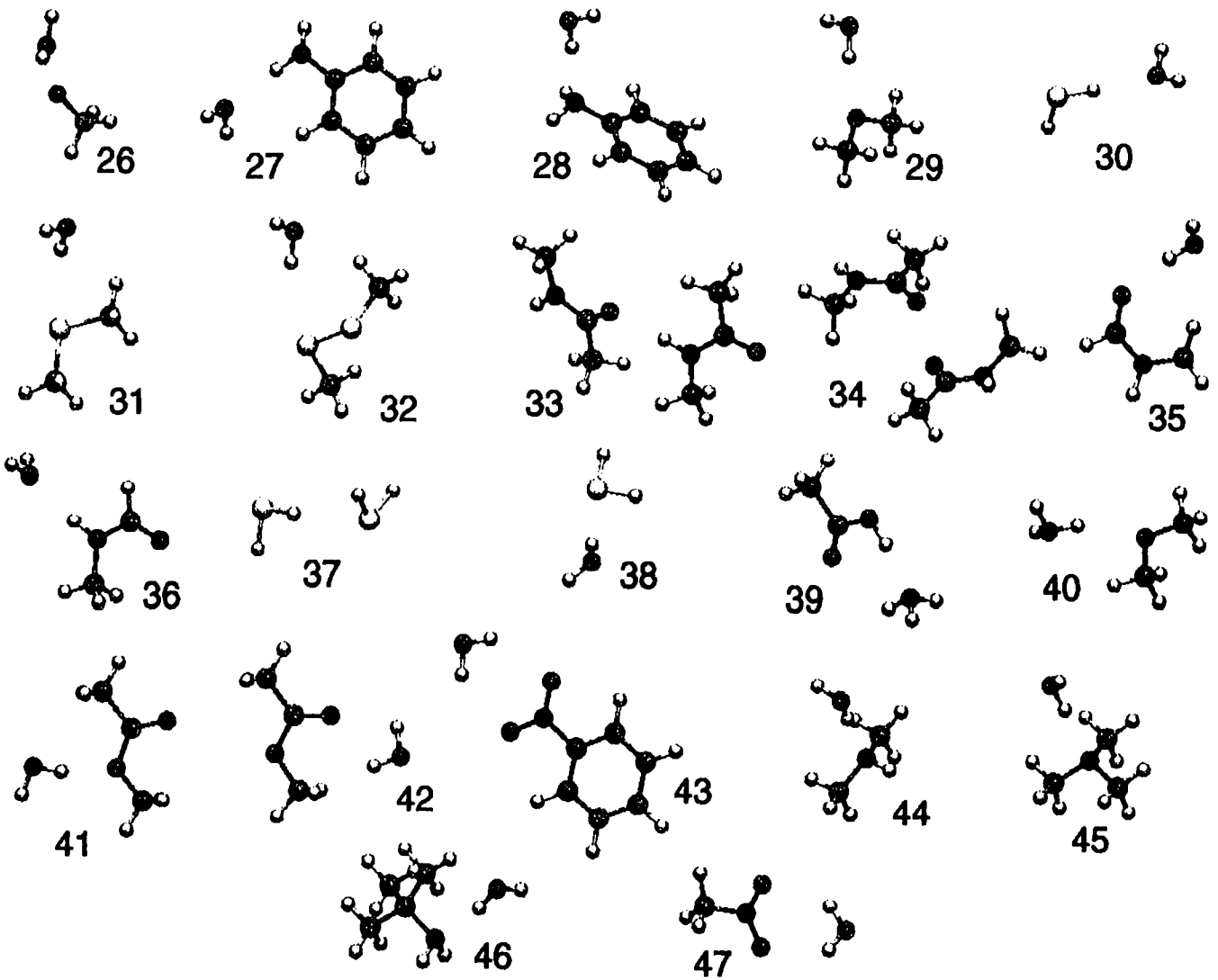


Figure 6.1. (continued)

TABLE 6.7

Comparison of Hydrogen-Bonded Organic Homo- and Hetero-Dimer Energies[a] Between Various Charge Models and HF/6-31G* and MP2/6-31+G.**

Dimer	HF/6-31G*	MP2/6-31+G**	RESP	ΔHF	ΔMP2	AM1	ΔHF	ΔMP2	MMFF	ΔHF	ΔMP2	AM1-BCC	ΔHF	ΔMP2
HOH...$OHCH_3$ **1**	-5.56	-6.80	-6.33	0.77	-0.47	-3.39	-2.17	-3.41	-6.44	0.88	-0.36	-5.70	0.14	-1.10
CH_3OH...OH_2 **2**	-5.59	-6.21	-6.90	1.31	0.69	-2.41	-3.18	-3.80	-6.36	0.77	0.15	-6.57	0.98	0.36
CH_3OH...$OHCH_3$ **3**	-5.54	-6.89	-6.67	1.13	-0.22	-1.90	-3.64	-4.99	-6.32	0.78	-0.57	-5.82	0.28	-1.07
C_6H_5OH...OH_2 **4**	-7.37	-8.61	-8.41	1.04	-0.20	-3.95	-3.42	-4.66	-8.99	1.62	0.38	-8.19	0.82	-0.42
HOH...OHC_6H_5 **5**	-4.71	-5.69	-5.54	0.83	-0.15	-4.59	-0.12	-1.10	-5.49	0.78	-0.20	-4.98	0.27	-0.71
T-NMA...OH_2 **6**	-5.43	-6.65	-6.60	1.17	-0.05	-3.90	-1.53	-2.75	-7.09	1.66	0.44	-6.25	0.82	-0.40
HOH...T-NMA **7**	-7.30	-7.97	-8.59	1.29	0.62	-6.43	-0.87	-1.54	-8.99	1.69	1.02	-9.29	1.99	1.32
formamide, cyclic **8**	-13.44	-14.36	-14.10	0.66	-0.26	-5.80	-7.64	-8.56	-12.73	-0.71	-1.63	-11.27	-2.17	-3.09
formamide, 1-HB **9**	-7.38	-7.86	-7.80	0.42	-0.06	-4.12	-3.26	-3.74	-8.1	0.72	0.24	-7.21	-0.17	-0.65
H_2NH...NH_3, bifurcated **10**	-3.19	-3.65	-4.43	1.24	0.78	-0.57	-2.62	-3.08	-3.9	0.71	0.25	-3.30	0.11	-0.35
H_2NH...NH_3, linear **11**	-3.07	-3.68	-4.47	1.40	0.79	-0.58	-2.49	-3.10	-3.95	0.88	0.27	-3.35	0.28	-0.33
HOH...NH_3 **12**	-6.56	-7.65	-8.01	1.45	0.36	-2.08	-4.48	-5.57	-7.48	0.92	-0.17	-6.81	0.25	-0.84
HOH...NH_2CH_3 **13**	-6.53	-8.00	-7.89	1.36	-0.11	-2.92	-3.61	-5.08	-7.76	1.23	-0.24	-7.25	0.72	-0.75
H_3CCOOH...$HOOCCH_3$ **14**	-15.55	-16.10	-15.96	0.41	-0.14	-5.65	-9.90	-10.45	-17.7	2.15	1.60	-14.70	-0.85	-1.40
H_2CCOOH...OHH **15**	-10.66	-10.75	-10.55	-0.11	-0.20	-5.93	-4.73	-4.82	-11.73	1.07	0.98	-10.23	-0.43	-0.52
imidazole...OH_2 **16**	-6.37	-7.88	-7.13	0.76	-0.75	-4.65	-1.72	-3.23	-8.05	1.68	0.17	-6.47	0.10	-1.41
HOH...imidazole **17**	-7.06	-7.82	-6.52	-0.54	-1.30	-5.38	-1.68	-2.44	-8.13	1.07	0.31	-7.54	0.48	-0.28
indole...OH_2 **18**	-5.75	-7.42	-6.90	1.15	-0.52	-4.53	-1.22	-2.89	-6.68	0.93	-0.74	-6.09	0.34	-1.33
pyrrole...OH_2 **19**	-5.36	-7.11	-6.41	1.05	-0.70	-4.29	-1.07	-2.82	-6.42	1.06	-0.69	-5.67	0.31	-1.44
$CH_3NH_3^+$...OH_2 **20**	-19.30	-19.80	-18.31	-0.99	-1.49	-17.48	-1.82	-2.32	-19.49	0.19	-0.31	-19.53	0.23	-0.27
OHH...(-)O_2CCH_3 **21**	-21.85	-20.74	-24.34	2.49	3.60	-21.32	-0.53	0.58	-26.15	4.30	5.41	-24.74	2.89	4.00
HOH...pyridine **22**	-6.03	-7.28	-6.13	0.10	-1.15	-4.72	-1.31	-2.56	-7.32	1.29	0.04	-6.33	0.30	-0.95
HOH...furan **23**	-3.64	-4.10	-4.62	0.98	0.52	-4.50	0.86	0.40	-4.45	0.81	0.35	-4.04	0.40	-0.06
HOH...thiophene **24**	-2.42	-3.54	-4.48	2.06	0.94	-6.48	4.06	2.94	-3.91	1.49	0.37	-4.13	1.71	0.59
HOH...acetone **25**	-6.25	-6.57	-6.76	0.51	0.19	-4.44	-1.81	-2.13	-7.56	1.31	0.99	-6.67	0.42	0.10
HOH...FCH_3 **26**	-4.42	-4.30	-3.32	-1.10	-0.98	-2.55	-1.87	-1.75	-3.94	-0.48	-0.36	-3.03	-1.39	-1.27
aniline...OHH **27**	-4.20	-4.04	-4.76	0.56	0.72	-2.81	-1.39	-1.23	-4.78	0.58	0.74	-5.20	1.00	1.16
HOH...aniline **28**	-5.17	-5.02	-5.62	0.45	0.60	-4.82	-0.35	-0.20	-6.97	1.80	1.95	-5.97	0.80	0.95
HOH...$O(CH3)_2$ **29**	-5.31	-6.72	-4.68	-0.63	-2.04	-4.17	-1.14	-2.55	-6.42	1.11	-0.30	-5.14	-0.17	-1.58
SHH...OHH **30**	-2.66	-3.66	-3.07	0.41	-0.59	-0.98	-1.68	-2.68	-3.57	0.91	-0.09	-4.47	1.81	0.81

(Continues on next page)

Table 6.7
(continued)

Dimer	RHF/6-31G*	MP2/6-31+G**	RESP	ΔHF	ΔMP2	AM1	ΔHF	ΔMP2	MMFF	ΔHF	ΔMP2	AM1-BCC	ΔHF	ΔMP2
HOH...$S(CH_3)_2$ **31**	-3.24	-3.05	-3.51	0.27	0.46	-1.39	-1.85	-1.66	-4.36	1.12	1.31	-3.50	0.26	0.45
OHH...CH_3SSCH_3 cyclic **32**	-3.44	-4.11	-3.31	-0.13	-0.80	-1.87	-1.57	-2.24	-3.97	0.53	-0.14	-3.57	0.13	-0.54
T-NMA dimer, parallel **33**	-7.00	-8.95	-9.65	2.65	0.70	-5.57	-1.43	-3.38	-10.43	3.43	1.48	-8.74	1.74	-0.21
T-NMA antiparallel stacked **34**	-4.92	-7.46	-7.75	2.83	0.29	-5.57	0.65	-1.89	-9.54	4.62	2.08	-7.69	2.77	0.23
HOH...N-Me formamide **35**	-7.18	-7.65	-8.30	1.12	0.65	-6.13	-1.05	-1.52	-8.89	1.71	1.24	-8.97	1.79	1.32
N-Me formamide...OH_2 **36**	-5.50	-6.80	-6.25	0.75	-0.55	-4.00	-1.50	-2.80	-6.93	1.43	0.13	-6.31	0.81	-0.49
HSH...SH_2 **37**	-0.88	-1.57	-1.32	0.44	-0.25	-0.41	-0.47	-1.16	-1.66	0.78	0.09	-2.43	1.55	0.86
HOH...SH_2 **38**	-2.06	-2.88	-2.16	0.10	-0.72	-0.71	-1.35	-2.17	-2.51	0.45	-0.37	-3.15	1.09	0.27
$CH_3COOH...NH_3$, bidentate **39**	-10.85	-12.11	-11.73	0.88	-0.38	-2.02	-8.83	-10.09	-12.55	1.70	0.44	-9.65	-1.20	-2.46
$H_2NH...O(CH_3)_2$ **40**	-2.84	-4.10	-3.07	0.23	-1.03	-1.25	-1.59	-2.85	-3.79	0.95	-0.31	-2.97	0.13	-1.13
HOH...Me acetate (-O-) **41**	-3.29	-4.82	-3.60	0.31	-1.22	-3.72	0.43	-1.10	-3.03	-0.26	-1.79	-4.13	0.84	-0.69
HOH...Me acetate (O=) **42**	-6.02	-6.59	-7.89	1.87	1.30	-5.67	-0.35	-0.92	-7.6	1.58	1.01	-7.62	1.60	1.03
OHH...nitrobenzene **43**	-5.04	-5.28	-5.83	0.79	0.55	-5.21	0.17	-0.07	-5.95	0.91	0.67	-3.57	-1.47	-1.71
OHH...nitromethane **44**	-4.86	-4.95	-5.39	0.53	0.44	-4.96	0.10	0.01	-6.29	1.43	1.34	-3.43	-1.43	-1.52
HOH...dimethyl amine **45**	-6.19	-8.05	-6.63	0.44	-1.42	-3.72	-2.47	-4.33	-8.01	1.82	-0.04	-7.80	1.61	-0.25
HOH...trimethylamine **46**	-5.90	-8.46	-3.74	-2.16	-4.72	-4.16	-1.74	-4.30	-7.99	2.09	-0.47	-8.07	2.17	-0.39
HOH...t-butylamine **47**	-7.08	-8.32	-3.42	-3.66	-4.90	-3.42	-3.66	-4.90	-7.96	0.88	-0.36	-7.54	0.46	-0.78
Mean error				0.61	-0.28		-1.98	-2.87		1.24	0.35		0.53	-0.36
Mean unsigned error				1.01	0.88		2.24	3.04		1.30	0.74		0.93	0.93
Maximum error				3.66	4.90		9.9	10.45		4.62	5.41		2.89	4.00
Standard deviation				1.12	1.33		2.36	2.47		0.97	1.10		1.07	1.15

[a] All energies are in units of kcal/mol.

The AM1-BCC model has the same mean unsigned error with respect to both HF and MP2. However, examination of the mean error reveals that AM1-BCC is overestimated with respect to HF (0.53 kcal/mol) and underestimated with respect to MP2 (–0.36 kcal/mol), also observed with RESP (0.61 and –0.28 kcal/mol, respectively). Conversely, the AM1 charge model underestimates (-1.98 and –2.87 kcal/mol) and the MMFF charge model overestimates (1.24 and 0.35 kcal/mol) with respect to both HF and MP2, respectively. When compared to the HF dimer energies, the AM1-BCC charge model performs best overall with a mean unsigned error of 0.93 kcal/mol while RESP has a mean unsigned error of 1.01 kcal/mol. However, compared to MP2, RESP slightly outperforms AM1-BCC (0.88 and 0.93 kcal/mol). The AM1 charge model has the highest mean unsigned error with respect to both HF and MP2 (2.24 and 3.04 kcal/mol) with maximum error greater than 9 kcal/mol. Overall, the AM1-BCC charge model produces dimer energies that tend to fall in between HF and MP2 values. Compared to MP2, the RESP charge model has a lower mean unsigned error than AM1-BCC (0.88 *vs.* 0.93 kcal/mol, respectively) while AM1 has the highest (3.04 kcal/mol). This behaviour is consistent with the results found in Chapter 4 (see Figures 4.3, 4.4, 4.5, and Table 4.1) where the ESP generated with the RESP model had an overall lower RMS than AM1-BCC while AM1 had the highest RMS with respect to the QM ESP. However, compared to the HF, the AM1-BCC charge model has the lowest mean unsigned error, albeit by only 0.05 kcal/mol.

Representative structures for the DNA dimers are shown in Figure 6.2 while Table 6.8 compares the dimer energies of the RESP, AM1, MMFF, and AM1-BCC charge models to those calculated by *ab initio* HF/6-31G** and MP2/6-31G*(.25) levels

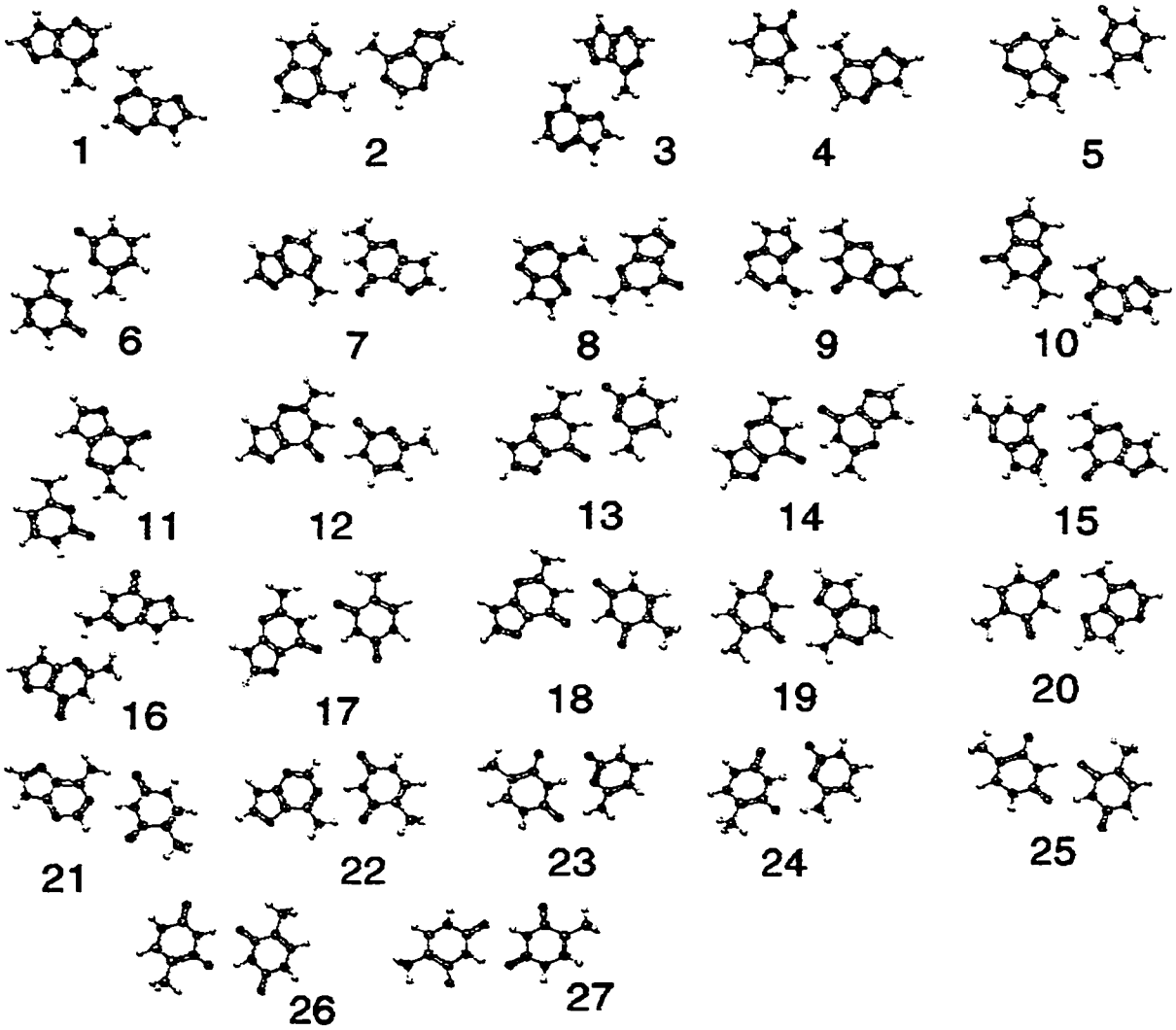


Figure 6.2. Representative structures for the DNA dimers. 1) AA1 2) AA2 3) AA3 4) AC1 5) AC2 6) CC 7) GA1 8) GA2 9) GA3 10) GA4 11) GC1 12) GCNEW 13) GCWC 14) GG1 15) GG3 16) GG4 17) GT1 18) GT2 19) TAH 20) TARH 21) TARWC 22) TAWC 23) TC1 24) TC2 25) TT1 26) TT2 27) TT3. Definitions: A (adenine), C (cytosine), G (guanine), and T (thymine). Colour representations: green (C), white (H), red (O), blue (N).

TABLE 6.8

Comparison of Hydrogen-Bonded DNA Dimer Energies[a] with HF/6-31G[b] and MP2/6-31G*(.25)[b, c] Energies.**

Dimers[d]	HF/6-31G**	MP2/6-31G*(.25)	RESP	ΔHF	ΔMP2	AM1	ΔHF	ΔMP2	MMFF	ΔHF	ΔMP2	AM1-BCC	ΔHF	ΔMP2
AA1	-8.7	-11.5	-10.19	1.5	-1.3	-5.65	-3.1	-5.9	-7.87	-0.8	-3.6	-11.60	2.9	0.1
AA2	-8.0	-11.0	-10.50	2.5	-0.5	-7.06	-0.9	-3.9	-8.41	0.4	-2.6	-11.55	3.6	0.6
AA3	-6.9	-9.8	-10.79	3.9	1.0	-8.16	1.3	-1.6	-8.49	1.6	-1.3	-11.91	5.0	2.1
AC1	-11.9	-14.3	-12.26	0.4	-2.0	-6.28	-5.6	-8.0	-9.05	-2.9	-5.3	-13.58	1.7	-0.7
AC2	-11.4	-14.1	-12.98	1.6	-1.1	-6.16	-5.2	-7.9	-10.08	-1.3	-4.0	-13.67	2.3	-0.4
CC	-17.3	-18.8	-16.57	-0.7	-2.2	-9.31	-8.0	-9.5	-14.74	-2.6	-4.1	-19.07	1.8	0.3
GA1	-12.6	-15.2	-14.16	1.6	-1.0	-10.23	-2.4	-5.0	-12.83	0.2	-2.4	-13.96	1.4	-1.2
GA2	-7.5	-10.3	-10.35	2.9	0.0	-8.41	0.9	-1.9	-10.58	3.1	0.3	-12.51	5.0	2.2
GA3	-11.0	-13.8	-15.13	4.1	1.3	-11.34	0.3	-2.5	-13.84	2.8	0.0	-14.83	3.8	1.0
GA4	-8.8	-11.4	-9.77	1.0	-1.6	-11.34	2.5	-0.1	-9.57	0.8	-1.8	-12.21	3.4	0.8
GC1	-12.7	-14.3	-12.28	-0.4	-2.0	-8.95	-3.8	-5.4	-12.13	-0.6	-2.2	-15.00	2.3	0.7
GCNEW	-22.8	-22.2	-21.73	-1.1	-0.5	-9.64	-13.2	-12.6	-18.88	-3.9	-3.3	-19.92	-2.9	-2.3
GCWC	-25.5	-25.8	-25.35	-0.1	-0.4	-11.48	-14.0	-14.3	-23.11	-2.4	-2.7	-26.00	0.5	0.2
GG1	-25.0	-24.7	-24.41	-0.6	-0.3	-10.64	-14.4	-14.1	-21.65	-3.4	-3.1	-22.98	-2.0	-1.7
GG3	-16.8	-17.8	-17.41	0.6	-0.4	-11.28	-5.5	-6.5	-17.61	0.8	-0.2	-17.07	0.3	-0.7
GG4	-7.3	-10.0	-8.61	1.3	-1.4	-10.84	3.5	0.8	-10.26	3.0	0.3	-12.37	5.1	2.4
GT1	-14.0	-15.1	-15.62	1.6	0.5	-6.97	-7.0	-8.1	-15.65	1.7	0.6	-15.33	1.3	0.2
GT2	-13.7	-14.7	-15.36	1.7	0.7	-6.96	-6.7	-7.7	-15.36	1.7	0.7	-14.50	0.8	-0.2
TAH	-10.9	-13.3	-13.66	2.8	0.4	-7.98	-2.9	-5.3	-13.13	2.2	-0.2	-13.77	2.9	0.5
TARH	-10.9	-13.2	-13.70	2.8	0.5	-6.28	-4.6	-6.9	-12.75	1.9	-0.4	-13.66	2.8	0.5
TARWC	-10.2	-12.4	-12.51	2.3	0.1	-6.40	-3.8	-6.0	-12.15	2.0	-0.3	-13.27	3.1	0.9
TAWC	-10.3	-12.4	-12.50	2.2	0.1	-6.47	-3.8	-5.9	-12.52	2.2	0.1	-13.47	3.2	1.1
TC1	-9.1	-11.4	-11.31	2.2	-0.1	-5.76	-3.3	-5.6	-11.52	2.4	0.1	-11.78	2.7	0.4
TC2	-9.2	-11.6	-11.43	2.2	-0.2	-5.83	-3.4	-5.8	-11.85	2.7	0.3	-12.35	3.2	0.8
TT1	-9.1	-10.6	-12.06	3.0	1.5	-5.66	-3.4	-4.9	-12.56	3.5	2.0	-11.79	2.7	1.2
TT2	-9.0	-10.6	-11.93	2.9	1.3	-5.61	-3.4	-5.0	-12.84	3.8	2.2	-11.91	2.9	1.3
TT3	-9.1	-10.6	-12.23	3.1	1.6	-5.72	-3.4	-4.9	-12.85	3.8	2.3	-11.72	2.6	1.1
Mean error				1.7	-0.2		-4.2	-6.1		0.8	-1.1		2.3	0.4
Mean unsigned error				1.9	0.9		4.8	6.2		2.2	1.7		2.7	0.9
Maximum error				4.1	2.2		14.4	14.3		3.9	5.3		5.1	2.4
Standard deviation				1.4	1.1		4.5	3.6		2.3	2.0		1.8	1.1

[a] All energies are unit of kcal/mol.
[b] HF and MP2 energies are for HF/6-31G** optimized geometries and are taken from Ref. 65.
[c] The 6-31G*(.25) basis set represents a standard split valence 6-31G* basis set augmented with d-polarization functions on the non-hydrogen atoms with an exponent of 0.25 added to the second row elements (see Ref. 65).

On average, RESP, AM1, and MMFF underestimate the hydrogen-bond energies (i.e. not negative enough) with respect to the MP2 energies (mean errors of -0.2, -6.1, and -1.1 kcal/mol, respectively) while AM1-BCC, on average, tends to overestimate the hydrogen-bond energies (i.e. too negative; mean error of 0.4 kcal/mol). Compared to HF, the RESP, MMFF and AM1-BCC charge models overestimate (1.7, 0.8 and 2.30 kcal/mol, respectively) while the AM1 charge model underestimates (-4.2 kcal/mol). Compared to MP2, the mean unsigned errors show that both RESP and AM1-BCC have the smallest and equal errors (0.9 kcal/mol). Although MMFF had the smallest error in the organic homo- and hetero-dimers (0.74 kcal/mol, see Table 6.7), it performed poorly in reproducing the DNA dimer energies (mean unsigned error of 1.7 kcal/mol) which were not included in the MMFF parameterization set. In contrast, the AM1-BCC charge model maintains a mean unsigned error similar to that of the organic homo- and hetero-dimers (0.93 and 0.9 kcal/mol for the organic and DNA dimers, respectively) and demonstrates it's ability to account for delocalization within the densely functionalized DNA bases. The inability of MMFF to properly account for the electrostatics of densely functionalized molecules is a major drawback of the MMFF charge model and is examined further in the next section. Compared to HF, RESP, MMFF, and AM1-BCC show an increase in their mean unsigned errors while AM1 shows a decrease. All four charge models tested in this work reproduce the correct energetic ordering of the AT base pairs i.e. the AT Hoogsteen base pair is stronger than the Watson-Crick AT base pair[15]. The large mean unsigned error for the AM1 charge model (6.2 kcal/mol) and the large maximum errors (> 14 kcal/mol) clearly demonstrate that AM1 atomic charges are unsuitable for calculations of non-bonded interactions. In general, AM1 atomic charges

perform very poorly in reproducing *ab initio* interaction energies but are dramatically improved upon correction with BCCs while RESP tends to produce the best results.

6.4 Densely Functionalized Molecules

A crucial requirement for a charge model in the pharmaceutical industry is to properly account for the electrostatics of densely functionalized molecules. Nearly all molecules in The Merck Index, a wide representation of drug and drug-like molecules, contain multiple functional groups crowded in a local volume. In this section, RESP, AM1, MMFF, and AM1-BCC are investigated for their ability to correctly capture the electrostatics (compare to the QM ESP) of ethanol, a simple mono-functionalized molecule; homarine a zwitterion; *p*-methoxy benzenesulfonate, a multi-functionalized molecule; and a variant of the carcinogen MNNG[67], N-methyl-N'-cyano-N-nitrosoguanine cation (MCNG), a multi- and densely-functionalized cation (these molecules are depicted in Figure 6.3).

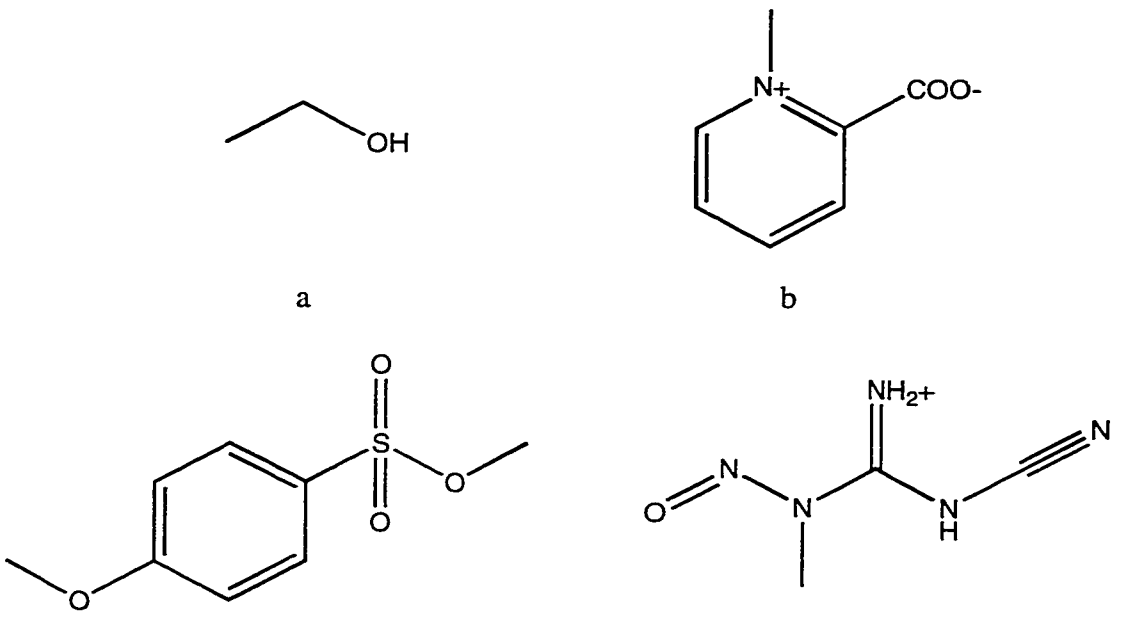


Figure 6.3. Representation of a) ethanol, a simple mono-functionalized molecule; b) homarine a zwitterion; c) *p*-methoxy benzenesulfonate, a multi-functionalized molecule; and d) N-methyl-N'-cyano-N-nitrosoguanine cation (MCNG), a multi- and densely-functionalized cation.

The RESP, AM1, MMFF, and AM1-BCC atomic charges for each molecule in Figure 6.3 were calculated, the ESPs generated, and the RMS with respect to the QM ESP calculated. Table 6.9 shows the RMS deviation of the four charge models with respect to the QM ESP for the molecules in Figure 6.3. The RESP, AM1, and AM1-BCC charge models perform as expected: AM1 atomic charges are much improved by the addition of the BCCs while RESP performs best overall. For the ethanol molecule, with a single functional group, MMFF performs moderately with an RMS of 0.0478 a.u. However, for the molecules with more than one functional group (i.e. homarine, *p*-methoxy benzenesulfonate, and N-methyl-N'-cyano-N-nitrosoguanine cation), the quality of the MMFF ESPs deteriorate drastically and are higher than the RMS of AM1 (see

Table 6.9). For the homarine molecule, a zwitterion with two functional groups, the MMFF RMS is approximately 24 times greater than AM1-BCC (1.0280 vs. 0.0424 a.u., respectively). For *p*-methoxy benzenesulfonate which contains π-delocalized electron withdrawing and donating functional groups, the MMFF RMS is approximately 33 times greater than AM1-BCC (0.7537 vs. 0.0230 a.u., respectively) and for the N-methyl-N'-cyano-N-nitrosoguanine cation which containing multiple functional groups in a local volume, the MMFF RMS is approximately 9 times greater than that of the AM1-BCC (0.5029 vs. 0.0547 a.u., respectively). MMFF was parameterized to reproduce dipole moments and scaled dimer energies, not ESPs. Although it is able to reproduce the ESP of the mono-functionalized molecule, multi-functionalized molecules demonstrate MMFF's deficiency in describing highly-delocalized systems.

TABLE 6.9.
RMS Deviations from the QM ESP[a] for RESP, MMFF, and AM1-BCC Charge models for the Molecules in Figure 6.3.

	RMS Deviation from QM ESP (a.u.)			
Molecule	RESP	AM1	MMFF	AM1-BCC
Ethanol	0.0254	0.0816	0.0408	0.0259
Homarine	0.0214	0.1079	1.0277	0.0424
p-methoxy benzenesulfonate	0.0190	0.1159	0.7537	0.0230
MCNG	0.0294	0.1593	0.5029	0.0547

[a] The ESP was calculated at the HF/6-31G* level of theory.

6.5 Relative Free Energies of Solvation

The calculation of the relative free energies of solvation (RFES) is the critical test of the ability of a FF to correctly capture the dynamic non-covalent interactions between solute-solvent and solvent-solvent molecules. Unlike the dimer energies, where

comparisons were made to quantum mechanical data, the relative free energies of solvation (RFES) allow a direct comparison to experimental data and therefore provide the most important comparisons. Also, the RFES reflect an ensemble average, not an energy minimization as with the dimers. The validation criterion, as with the dimer calculations, is to reproduce the RFES within 1 kcal/mol. The RFES in this work were chosen to be between systems as isosteric as possible, but with one system as non-polar as possible. This places the burden on the electrostatic term for the RFES and thus focuses on the charge model, minimizing the vdW contributions. See appendix D for details of the computational methods.

6.5.1 Results and Discussion

Table 6.10 shows a comparison of relative free energies of solvation for organic molecules using various charge models.

TABLE 6.10.
Relative Free Energies[a] of Solvation of Organic Molecules Using Various Charge Models.

		RESP		AM1		MMFF		AM1-BCC	
Perturbation	Exp.[b]	ΔΔG	error	ΔΔG	error	ΔΔG	error	ΔΔG	error
Ammonia → Methane	6.31	8.27	-1.96	1.21	5.11	7.36	-1.05	6.19	0.12
CH_3NH_2 → Ethane	6.39	7.63	-1.24	1.61	4.79	7.08	-0.69	6.32	0.07
$(CH_3)_2NH$ → Propane	6.25	5.15	1.10	1.86	4.39	6.40	-0.15	5.98	0.27
$(CH_3)_3N$ → Isobutane	5.56	1.68	3.89	2.02	3.55	5.63	-0.07	5.64	-0.08
Phenol → Aniline	-0.10	2.40	-2.50	-1.49	1.39	1.58	-1.68	0.15	-0.25
Pyridine → Benzene	3.83	2.71	1.12	0.65	3.18	4.79	-0.96	2.64	1.19
THF → Cyclopentane	4.67	4.61	0.06	2.66	2.02	5.06	-0.39	3.85	0.83
Cyclopentene → Cyclopentane	0.64	0.13	0.51	0.73	-0.09	0.60	0.04	0.33	0.31
Pyrrolidine → Cyclopentane	6.68	5.93	0.76	1.21	5.48	6.42	0.26	5.73	0.96
Benzene → Cyclohexane	2.09	1.62	0.47	1.48	0.61	2.65	-0.56	1.50	0.59
Piperidine → Cyclohexane	6.34	4.74	1.61	1.84	4.51	6.03	0.32	5.77	0.58
Acetaldehyde → Ethane	5.33	5.80	-0.47	2.62	2.71	8.15	-2.82	5.84	-0.51
Bromomethane → Ethane	2.63	2.09	0.54	2.99	-0.36	2.09	0.54	1.82	0.81
Chloromethane → Ethane	2.43	0.08	2.35	-0.67	3.10	0.78	1.65	-0.13	2.56
Ethene → Ethane	0.56	0.42	0.15	-0.13	0.69	0.22	0.35	-0.11	0.67
Fluoromethane → Ethane	2.03	1.95	0.09	1.11	0.93	2.68	-0.65	1.65	0.39
Methanol → Ethane	6.95	8.47	-1.52	3.01	3.94	8.11	-1.16	7.28	-0.33
Methylthiol → Ethane	3.07	2.98	0.09	0.52	2.55	2.94	0.13	2.84	0.23
Anisole → Ethylbenzene	0.24	0.87	-0.63	0.71	-0.47	0.75	-0.51	1.47	-1.23
Acetamide → Isobutane	12.03	12.54	-0.51	6.89	5.15	12.16	-0.13	12.26	-0.23
Acetic acid → Isobutane	9.02	10.97	-1.95	5.30	3.73	11.11	-2.09	10.00	-0.98
Nitrobenzene → *i*-Pr benzene	3.80	6.91	-3.11	5.67	-1.87	5.77	-1.97	4.48	-0.68
Me 2Pr ether → Isopentane	4.39	2.76	1.63	2.53	1.86	4.64	-0.25	3.48	0.91
Me acetate → Isopentane	5.70	6.59	-0.89	4.57	1.13	6.17	-0.47	6.66	-0.96
Me butene → Isopentane	1.07	-0.27	1.34	0.76	0.31	0.80	0.27	0.22	0.85
Me Et ketone → Isopentane	6.02	6.35	-0.33	3.21	2.81	8.17	-2.15	6.45	-0.43
Nitroethane → Isopentane	6.08	9.39	-3.31	8.60	-2.52	11.13	-5.05	5.57	0.51
NMA → Isopentane	12.50	10.98	1.53	6.67	5.84	11.77	0.73	12.19	0.32
Mepiperidine→Mecyclohexane	5.60	1.25	4.36	1.80	3.81	5.25	0.36	5.48	0.12
Toluene → Me cyclohexane	2.59	1.77	0.83	2.29	0.31	3.32	-0.73	2.05	0.55
Me imidazole → Me pyrrole	5.52	3.64	1.89	1.02	4.51	7.05	-1.53	4.19	1.34
Me acetate → NMA	-6.76	-3.59	-3.17	-1.30	-5.46	-4.80	-1.96	-4.73	-2.03
Phenol → Toluene	5.73	6.69	-0.96	0.51	5.23	6.41	-0.68	5.81	-0.07
$(CH_3)_2O$ → Propane	3.85	3.18	0.67	2.39	1.46	4.93	-1.08	3.43	0.42
$(CH_3)_2S$ → Propane	3.49	2.52	0.97	1.15	2.34	4.18	-0.69	2.82	0.67
Imidazole → Pyrrole	4.85	3.57	1.28	1.12	3.73	8.63	-3.78	4.50	0.35
Benzaldehyde → Toluene	3.14	4.47	-1.33	1.77	1.38	6.42	-3.28	4.51	-1.37
Bromobenzene → Toluene	0.61	0.08	0.53	0.57	0.04	-0.19	0.80	0.06	0.55
Chlorobenzene → Toluene	0.21	-1.88	2.09	-1.95	2.16	-2.09	2.30	-1.98	2.19
Thiobenzene → Toluene	1.66	2.28	-0.62	1.60	0.06	1.20	0.47	1.33	0.34
Mean error			0.13		2.10		-0.71		0.24
Mean unsigned error			1.36		2.64		1.12		0.69
Maximum error			4.36		5.84		5.05		2.56
Standard deviation			1.73		2.42		1.42		0.88

[a] All energies are in kcal/mol

[b] Experimental values obtained from refs. 68 and 69.

In general, RESP, AM1, and AM1-BCC underestimate the relative free energies of solvation while MMFF overestimates as shown by the mean errors. The mean unsigned errors show that AM1-BCC reproduced the experimental data better than RESP (0.69 vs. 1.36, respectively). This is at least in part due to five BCCs that, *a posteriori*, were adjusted in order to improve agreement with the experimental data. Adjustments were made to the $N_{2-,3,4+} - H$ and $C_4 - N_{2-,3,4+}$ bond-types in order to reproduce the experimentally observed free energy of solvation trends of the amine series[70]: ammonia (-4.3 kcal/mol), N-Me amine (-4.6 kcal/mol), NN-diMe amine (-4.3 kcal/mol), and NNN-triMe amine (-3.2 kcal/mol). Testing the validity of these changes, these adjustments also improved the RFES of pyrrolidine to cyclopentane from 3.96 to 5.78 kcal/mol (experimental value of 6.68 kcal/mol); piperidine to cyclohexane from 4.13 to 5.74 kcal/mol (experimental value of 6.34 kca/mol); and Me piperidine to Me cyclohexane from 2.96 to 5.45 kcal/mol (experimental value of 5.60 kcal/mol). Adjustments were also made to the N_3^{hdeloc} $\underline{deloc}$ $O_{1,2}$ BCC, used for the nitro containing functionalities which appear to be problematic.[71] The BCC value for bond-type N_3^{hdeloc} $\underline{deloc}$ $O_{1,2}$ required a large adjustment from 0.1203 to -0.1500 changing the RFES of nitroethane to isopropane from 10.54 to 5.75 kcal/mol, in close agreement with the experimental value of 6.08 kcal/mol. Similarly, the relative free energy of solvation of nitrobenzene to isopropylbenzene changed from 7.44 to 4.29 kcal/mol, approaching the experimental value of 3.80 kcal/mol. For perturbations between hydrocarbons (e.g. cyclopentene to cyclopentane), the AM1 charge model better reproduced the RFES than AM1-BCC, suggesting that BCCs were not needed for unsaturated bonds containing hydrocarbons. Consequently, unsaturated hydrocarbon BCCs (i.e. $C_3^{=C} - H_1$ and $C_{ar} - H_1$) were set to

zero and the RFES were recalculated. Setting $C_4 - H_1$ to zero had an overall unfavourable effect on the RFES of molecules containing methyl groups; therefore the BCC for $C_4 - H_1$ was not adjusted. On the other hand a positive and predictable effect was observed for $C_3^{=C} - H_1$ and $C_{ar} - H_1$ where the experimental RFES were better reproduced. The change from 0.0116 to 0.0000 for $C_3^{=C} - H_1$ made only modest improvements while the adjustment from 0.0128 to 0.0000 for $C_{ar} - H_1$ improved the relative free energy of solvation of benzene to cyclohexane from 0.87 to 1.50 kcal/mol approaching the experimental value of 2.09 kcal/mol.

Because a two-body additive force field is incapable of reproducing the free energy of solvation of charged solutes,[72] ions were not included in these calculations. Although HF/6-31G* derived charge models embed averaged implicit polarization, an explicit polarization term is needed to accurately reproduce the free energy of solvation of ions.[73]

6.6 Correlation Coefficients

The correlation coefficients (see Appendix E for definition of correlation coefficient) between RESP, AM1, MMFF, and AM1-BCC for the *ab initio* dimer energies and the experimental relative free energies of solvation are shown in Table 6.11.

TABLE 6.11
Correlation Coefficients[a] between the Various Charge Models for the *Ab Initio* Dimer Energies and the Experimental Relative Free Energies of Solvation.

	Organic Dimers		DNA Dimers		RFES
	HF	MP2	HF	MP2	Exp.
RESP	0.96	0.95	0.97	0.97	0.88
AM1	0.82	0.79	0.50	0.54	0.70
MMFF	0.98	0.97	0.91	0.88	0.93
AM1-BCC	0.97	0.96	0.97	0.98	0.97

[a] Definition of correlation coefficient is given in Appendix E.

For the organic dimer energies, except for the AM1 charge model, all models correlate well with both *ab initio* methods with correlation coefficients above 0.95. The greatest correlation is for MMFF with the HF method (correlation coefficient of 0.98) but as explained above, MMFF was *parameterized* to reproduce *these* HF dimers. The AM1-BCC model correlates well with both *ab initio* methods, having been parameterized to neither.

The best correlation coefficient for the DNA dimer energies, between *ab initio* and charge model is for the AM1-BCC method with a correlation coefficient of 0.98 with respect to MP2. In both the organic and DNA dimer sets, AM1-BCC maintains good correlation with the *ab intio* methods (0.97 and 0.96 with respect to HF and MP2, respectively for the organic dimers and 0.97 and 0.98 with respect to HF and MP2, respectively for the DNA dimers), while MMFF loses correlation for the DNA dimers (0.98 and 0.97, respectively for the organic dimers vs. 0.91 and 0.88 with respect to HF

and MP2, respectively for the DNA dimers). RESP also maintains a good correlation for both organic and DNA dimers.

The AM1-BCC model achieves the highest correlation with experiment (correlation coefficient of 0.97) for the relative free energies of solvation. AM1-BCC is the only charge model to consistently maintain a high correlation coefficient for all the tests while RESP, AM1, and MMFF produce inconsistent correlation coefficients.

6.7 Validation Discussion and Conclusions

The concept of modifying AM1 population atomic charges with simple additive constants in order to generate atomic charges that emulate the HF/6-31G* ESP was demonstrated in Chapter 4 to be valid. As reviewed in section 1.5, many attempts have been made to create a charge model that quickly and efficiently generates atomic charges of high-quality. However, important electronic effects (i.e. formal charge and electron density distribution) have been left out of these charge models. Capturing both local and non-local electron distributions in the charges has revealed to be the *key* to generating high-quality atomic charges. Although the AM1 semi-empirical methodology is a crude model with many inherent approximations and deficiencies, it nonetheless systematically provides a basic electronic structural foundation in a fraction of the time required by an *ab initio* calculation. With the addition of simple empirical bond charge corrections (BCCs), this semi-empirical method is able to reproduce an *ab initio* level ESP (i.e. the HF/6-31G* ESP) and does so in a simple and efficient manner. The atom-types, which

compose the BCCs, are themselves simple, and are based on firm chemical ideas (i.e. hybridization states and bonded environments).

Although some charge models do include both local and non-local electronic effects, they underestimate the importance of reproducing the ESP of a molecule. The combination of capturing the electronic effects *and* reproducing the ESP has been shown in this chapter to be crucial in generating high-quality atomic charges.

It is evident from the organic dimer energies in Table 6.7 and the DNA dimer energies in Table 6.8 that the AM1-BCC method captures the required electron delocalization *and* reproduces the QM ESP because it is able to reproduce these *ab initio* energies within 1 kcal/mol. If the AM1-BCC model was unable to capture electron delocalization in combination with the QM ESP, then the DNA dimer energies would not agree with the *ab initio* data as demonstrated with the MMFF model. The MMFF model performs well with the organic dimer energies, which are simple monofunctional species, but loses agreement with the *ab initio* DNA dimer energies, which are electron delocalized ring systems. This is further substantiated with the densely functionalized molecules where it is evident that capturing the electron delocalization and formal charge distribution is important (i.e. for homarine, *p*-methoxy benzenesulfonate, and MCNG) because the AM1 model out-performs the MMFF model, which is unable to capture electron delocalization or formal charge distribution.

The relative free energies of solvation (RFES) in Table 6.10 show that the AM1-BCC model performs very well in reproducing these experimental results (average error of 0.69 kcal/mol). However, some AM1-BCC values have errors near or above 1 kcal/mol and it is not surprising that these systems contain molecules that bear lone pairs of electrons, e.g. the perturbations from pyridine to benzene (1.19 kcal/mol), pyrrolidine to cyclopentane (0.96 kcal/mol), anisole to ethylbenzene (-1.23 kcal/mol), acetic acid to isobutane (-0.98 kcal/mol), methyl acetate to isopentane (-0.96 kcal/mol), and methyl imidazole to methyl pyrrole (1.34 kcal/mol). Although atomic charge compensation (ACC) is allowed in the AM1-BCC model for aromatic carbons adjacent to either nitrogens or oxygens with lone pairs (which benefits pyridine and me imidazole in this case) the explicit inclusion of point-centers located in the lone pair volume on *all* molecules bearing lone pairs should improve the RFES. For example, lone pair centers would be expected to benefit amine nitrogens (e.g. pyrrolidine), carbonyl oxygens (e.g. acetic acid), conjugated oxygens (e.g. anisole), etc. Explicit inclusion of lone pair centers on aromatic nitrogen atoms would remove the need for ACC provided by the C^{lp}_{arom} atom-type and consequently reduce the number of bond-types, also reducing the degrees of freedom in the fit. Furthermore, it has been demonstrated that explicit inclusion of lone pair centers not only improves the quality of the ESP but also provides directionality to hydrogen-bonds between donors and acceptors.[63] The disadvantages of explicitly including lone pair centers are the increase in the number of non-bonded interactions to calculate and the creation of an artificial off-center point.

An alternative to explicitly including lone pair centers is to include point-dipoles on those centers that bear lone pairs. This method would improve the quality of the ESP, provide directionality to hydrogen-bonds, and would not be an off-center point bearing a mass. However, a point-dipole model would require special treatment for non-bonded interactions in the FF code because the angle between a point charge and a point-dipole is required for the energy calculation.

It has been demonstrated in this chapter that the AM1-BCC charge model is able to reproduce *ab initio* determined dimer energies of homo- and hetero-dimers of a diverse set of organic molecules within 0.95 kcal/mol. Also, AM1-BCC is able to capture electron delocalization in highly functionalized molecules such as the DNA dimers within 0.9 kcal/mol of *ab initio* and the densely functionalized molecules with RMS values less than 0.055 a.u. Furthermore, the AM1-BCC charge model is able to reproduce experimentally determined relative free energies of solvation within 0.69 kcal/mol of experiment and is the only charge model, amongst the ones tested in this work, to consistently maintain a correlation coefficient above 0.96 for *all* the validation tests with respect to *ab initio* and experiment.

7 Conclusions and Future Research

The AM1-BCC atomic charge method corrects an initial set of AM1 charges that do not emulate the HF/6-31G* electrostatic potential (ESP), into a new set of charges that do i.e. the AM1-BCC charges. The hypothesis was that this could be done using simple additive correction terms applied to the AM1 charges.

The AM1-BCC charge model is a fast and efficient method to generate high-quality atomic charges suitable for simulations of solution phase systems. A semi-empirical AM1 calculation is initially performed on a molecule to obtain AM1 atomic charges. Although these charges alone are unsuitable for solution phase simulations, they nevertheless capture basic electronic features of the molecule (e.g. delocalization and formal charge). The AM1 atomic charges are then corrected by applying a set of bond charge corrections (BCCs) that have been parameterized so that the resulting AM1-BCC charges reproduce the HF/6-31G* ESP; thus the user of the AM1-BCC model simply needs to calculate AM1 atomic charges and apply the BCCs, there are no *ab initio* calculations to perform. A training set of 2755 organic molecules was created to parameterize the 309 unique BCCs of the 354 that sample bond-types composed of H, C, N, O, F, P, S, Si, Cl, Br, and I atoms. The AM1-BCC charge model is capable of generating atomic charges for all the molecules in The Merck Index and the NCI database of small molecules, except molecules that contain boron and molecules containing metal atoms. Although silicon BCCs are present, a complete set of silicon BCCs has not been parameterized in this work. The AM1-BCC charge model is able to reproduce correlated

ab initio hydrogen-bonded organic homo- and hetero-dimer energies within, on average, 0.95 kcal/mol, hydrogen-bonded DNA dimer energies within, on average, 0.9 kcal/mol and tends to produce energies overestimated with respect to HF and underestimated with respect to MP2. This charge model was also validated against relative free energies of solvation of isosteric small organic molecules by performing free energy perturbations. Five BCCs were adjusted to better reproduce the relative free energies of solvation of the amine series, aromatic and olefinic hydrocarbons and nitro containing functionalities. This charge model is able to reproduce experimental relative free energies of solvation of a spectrum of isosteric organic molecules within, on average, 0.69 kcal/mol. Furthermore, it is the only charge model, amongst the ones tested in this work, to consistently maintain a correlation coefficient above 0.96 for *all* the validation tests with respect to *ab initio* and experiment. The AM1-BCC charge model generates very stable atomic charges compared to ESP-fit methodologies as demonstrated with the highly transferable atomic charges of the ether functional group between various molecules. These transferable charges should improve the quality of organic FFs by allowing the torsion parameters to be parameterized in a better behaved and general fashion. The AM1-BCC charge model is proposed as a general method to generate charges for bioorganic molecules for use in solution phase FFs.

Future effort on this work can be based on two aspects: i) improving the validation scheme and ii) improving the model itself. The validation of the AM1-BCC model, specially for the relative free energies of solvation should contain a more diverse and challenging set of molecules, i.e. molecules with multiple functional groups and

molecules that contain functional groups with electron-withdrawing and –donating groups instead of the monofunctional molecules which were used in this work. The charge model can be improved by: a) adding explicit lone pair centers to atoms bearing lone pairs or b) adding point-dipoles to atoms bearing lone pairs. The addition of either lone pair centers or point-dipoles would improve the quality of the ESP and provide directionality to hydrogen-bonds.

Appendix A

A.1 AM1 Atomic Charges

The electron density, $\rho(\mathbf{r})$, is a three-dimensional function that is normalized to the total number of electrons, n, in the system, i.e.

$$\int \rho(\mathbf{r}) d\mathbf{r} = n . \tag{A.1}$$

Within the linear combination of atomic orbitals (LCAO) approximation, AM1 atomic charges are derived from the on-diagonal elements of the density matrix of the AM1 semi-empirical wavefunction. The above integral can then be expanded in terms of K basis functions, ϕ_μ,

$$\rho(\mathbf{r}) = \sum_{\mu}^{K} \sum_{\nu}^{K} P_{\mu\nu}^{\alpha} \phi_\mu \phi_\nu \tag{A.2}$$

where $P_{\mu\nu}^{\alpha}$ are elements of the Coulson density matrix and can be obtained from the wavefunction when the self-consistent field (SCF) has converged.

$$P_{\mu\nu}^{\alpha} = \sum_{i=1}^{occ} c_{\mu i}^{\alpha} c_{\nu i}^{\alpha} \tag{A.3}$$

The summation runs over all occupied spin molecular orbitals, $c_{\mu i}^{\alpha}$ and $c_{\nu i}^{\alpha}$ are the molecular coefficients. Integration of equation A.2 gives,

$$\int \rho(\mathbf{r})d\mathbf{r} = \int \sum_{\mu}^{K}\sum_{\nu}^{K} P_{\mu\nu}^{\alpha}\phi_{\mu}\phi_{\nu}d\mathbf{r} = \sum_{\mu}^{K}\sum_{\nu}^{K} P_{\mu\nu}^{\alpha}S_{\mu\nu} = n \qquad \text{(A.4)}$$

where $S_{\mu\nu}$ are elements of the overlap matrix. In the neglect of diatomic differential overlap (NDDO) approximation, the overlap of two atomic orbitals located on different centers is set to zero, i.e. $S_{\mu\nu} = 0$, and normalization gives $S_{\mu\mu} = 1$ (i.e. the overlap matrix is set to the identity matrix **I**). Consequently, the on-diagonal elements of the density matrix, i.e. $P_{\mu\mu}^{\alpha}$, represent the number of electrons associated with a particular basis function ϕ_{μ}:

$$q_{\mu} = P_{\mu\mu}^{\alpha}. \qquad \text{(A.5)}$$

Therefore, the sum over the on-diagonal elements is equal to the total number of electrons in the system:

$$\sum_{\mu}^{K} P_{\mu\mu}^{\alpha} = n. \qquad \text{(A.6)}$$

To obtain the number of electrons on atom *i*, the summation is run over the basis functions, ϕ_{μ}, *on atom i* and subtract the nuclear charge of atom *i*, Z_i, from it:

$$q_i = Z_i - \sum_{\mu}^{i} q_{\mu\mu} \,. \tag{A.7}$$

The AM1 atomic charges produced from MOPAC-6 are these q_i's.

Appendix B

B.1 van der Waals Radii

Table B.1 lists the van der Waals radii used during the course of this work.

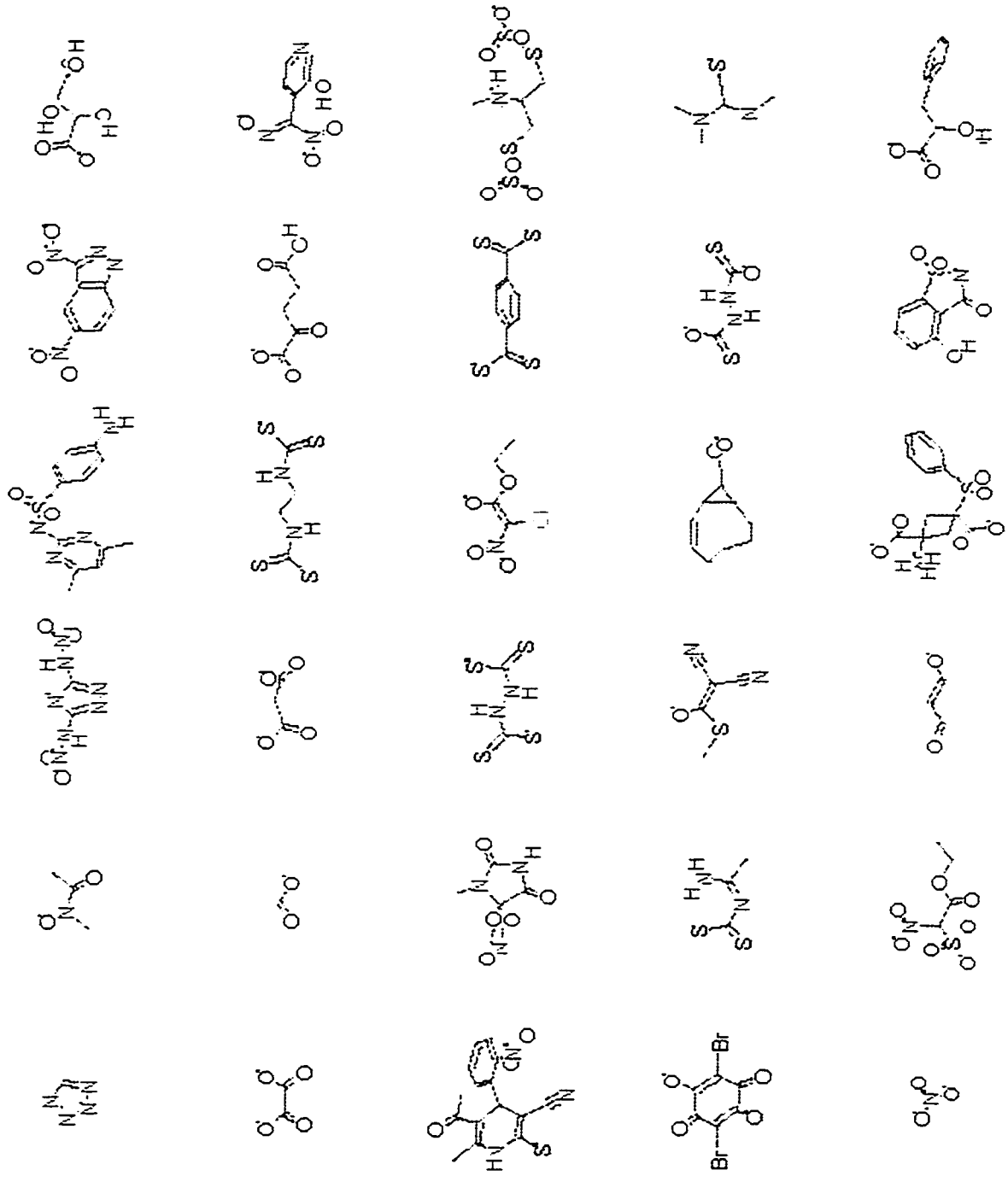
TABLE B.1
Table of van der Waals Radii Used.

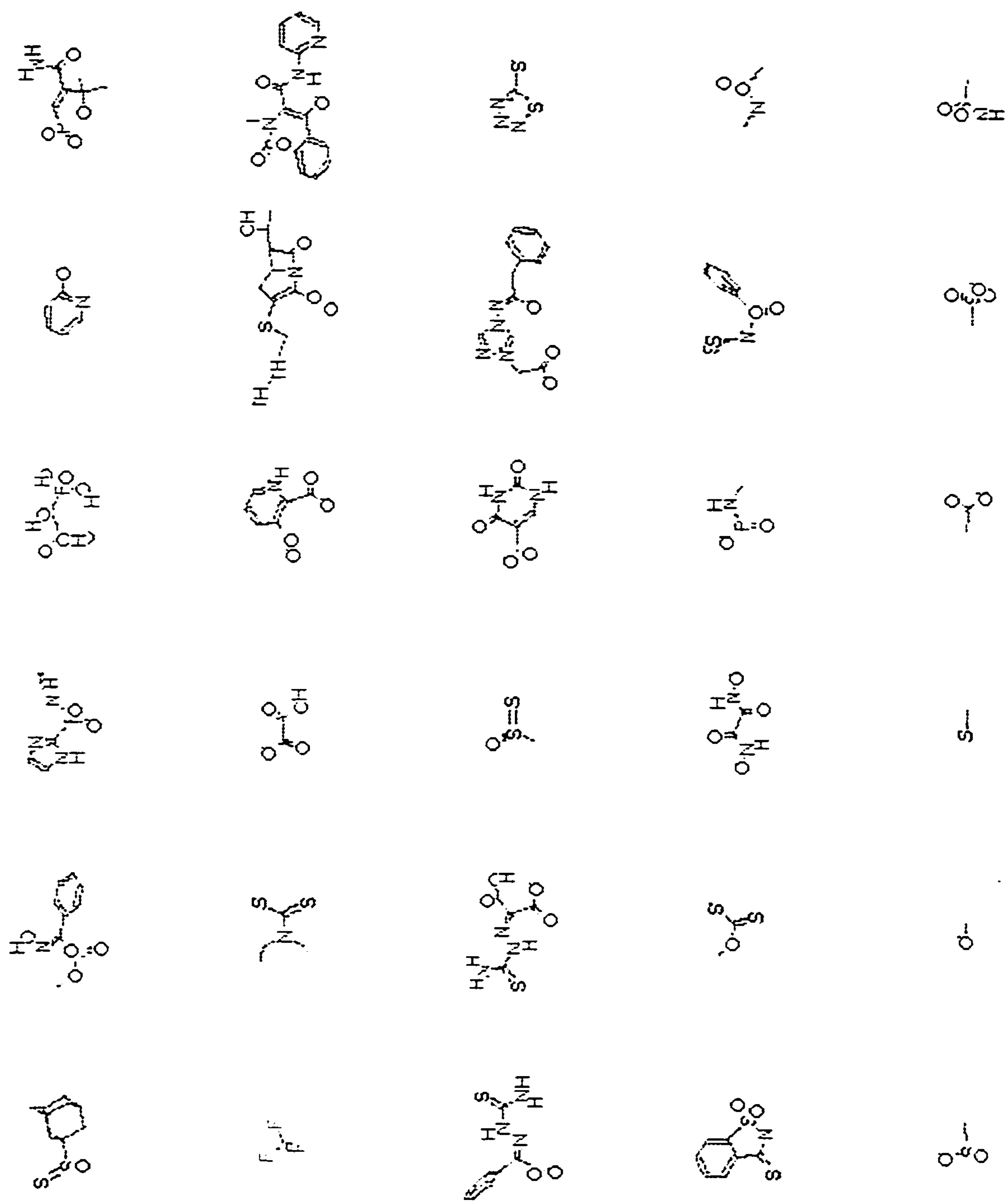
Atom	Radius[a] (Å)
H	1.2
C	1.7
N	1.63[b]
O	1.52
F	1.47
Si	2.1
P	1.8
S	1.85[c]
Cl	1.75
Br	1.85
I	1.98

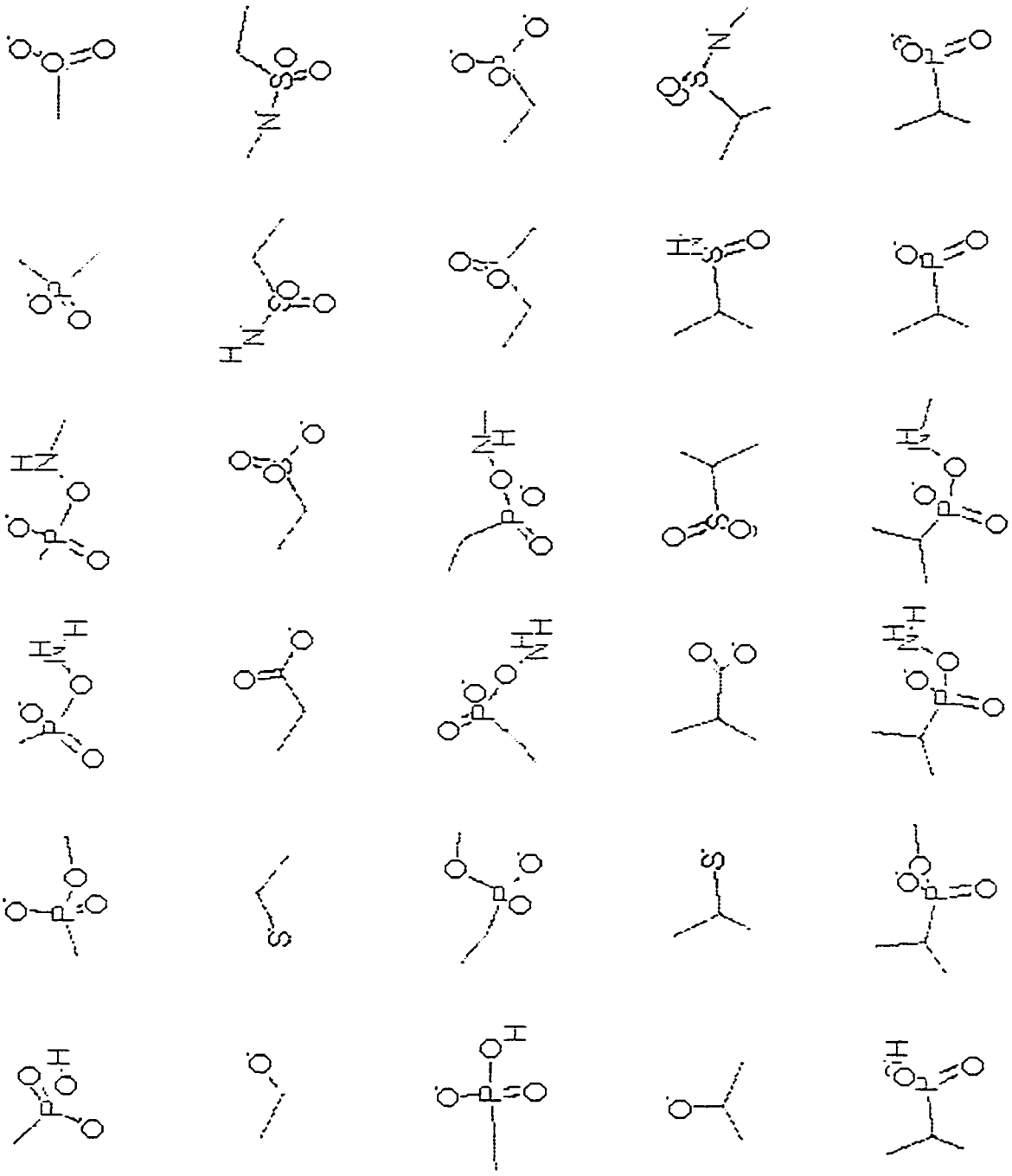
[a] From Table 1 in A. Bondi, *J. Phys. Chem.*, **68**, 441 (1964).
[b] Increased from 1.55 in reference a.
[c] Increased from 1.8 in reference a.

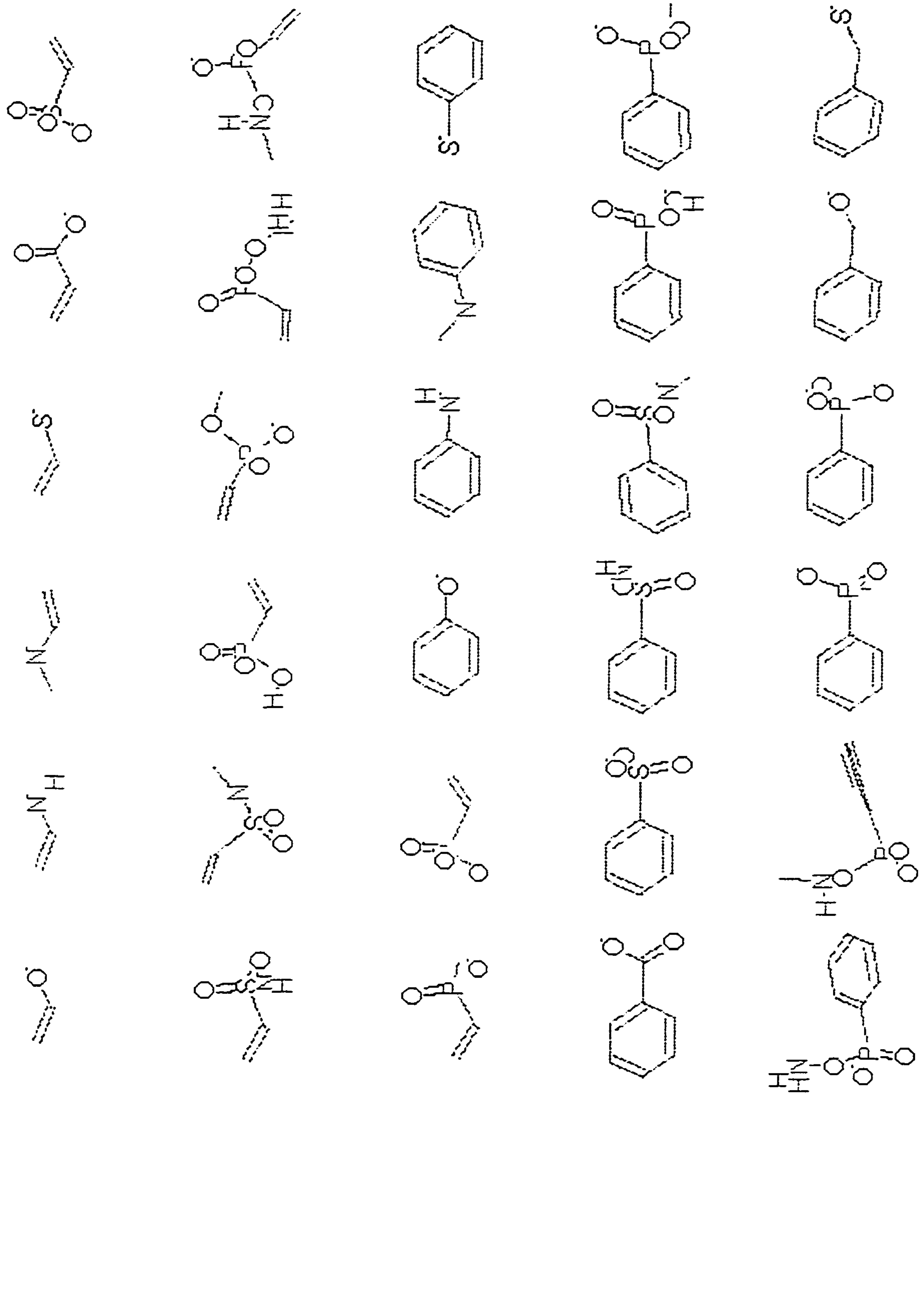
Appendix C

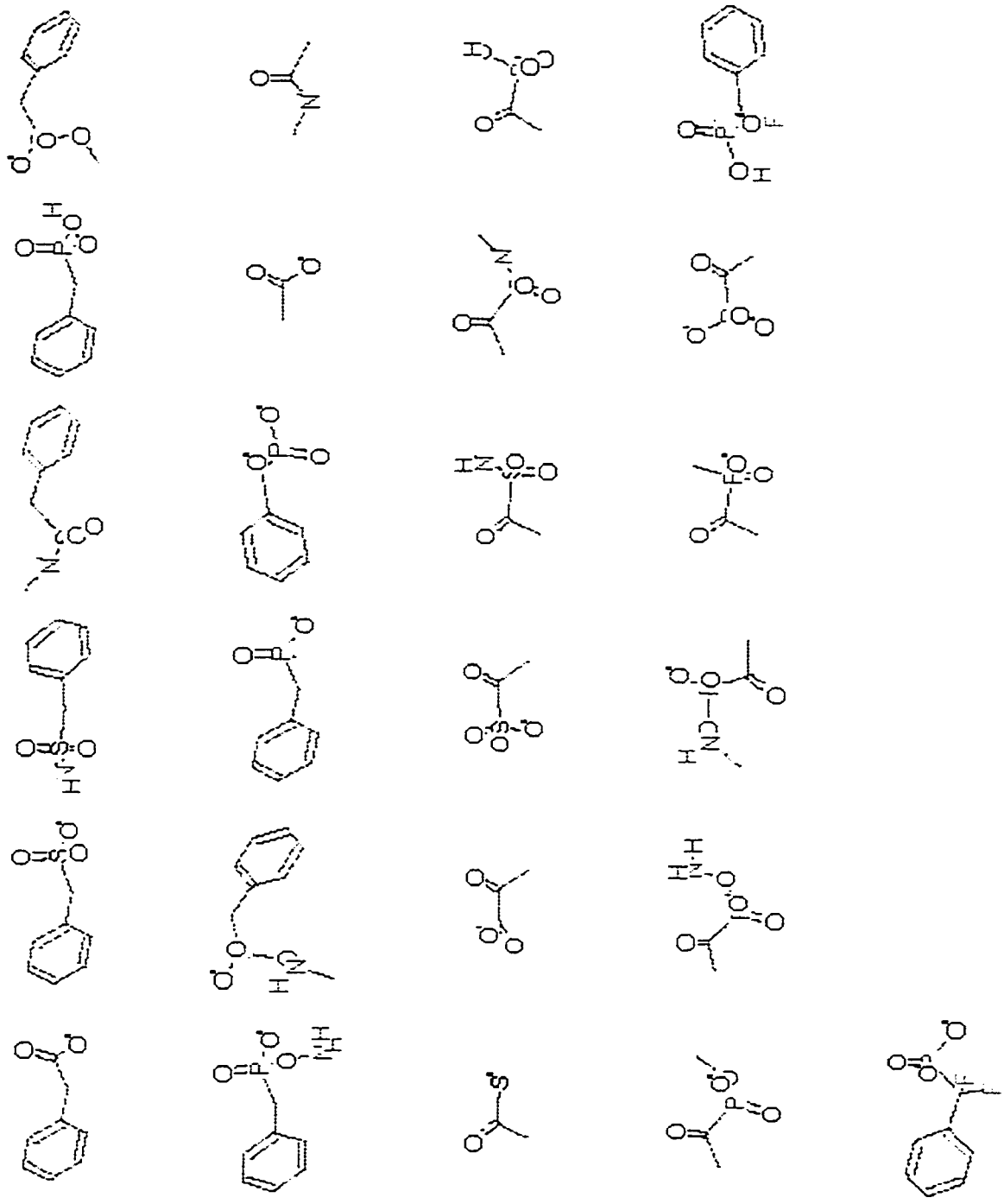
C.1 Anions



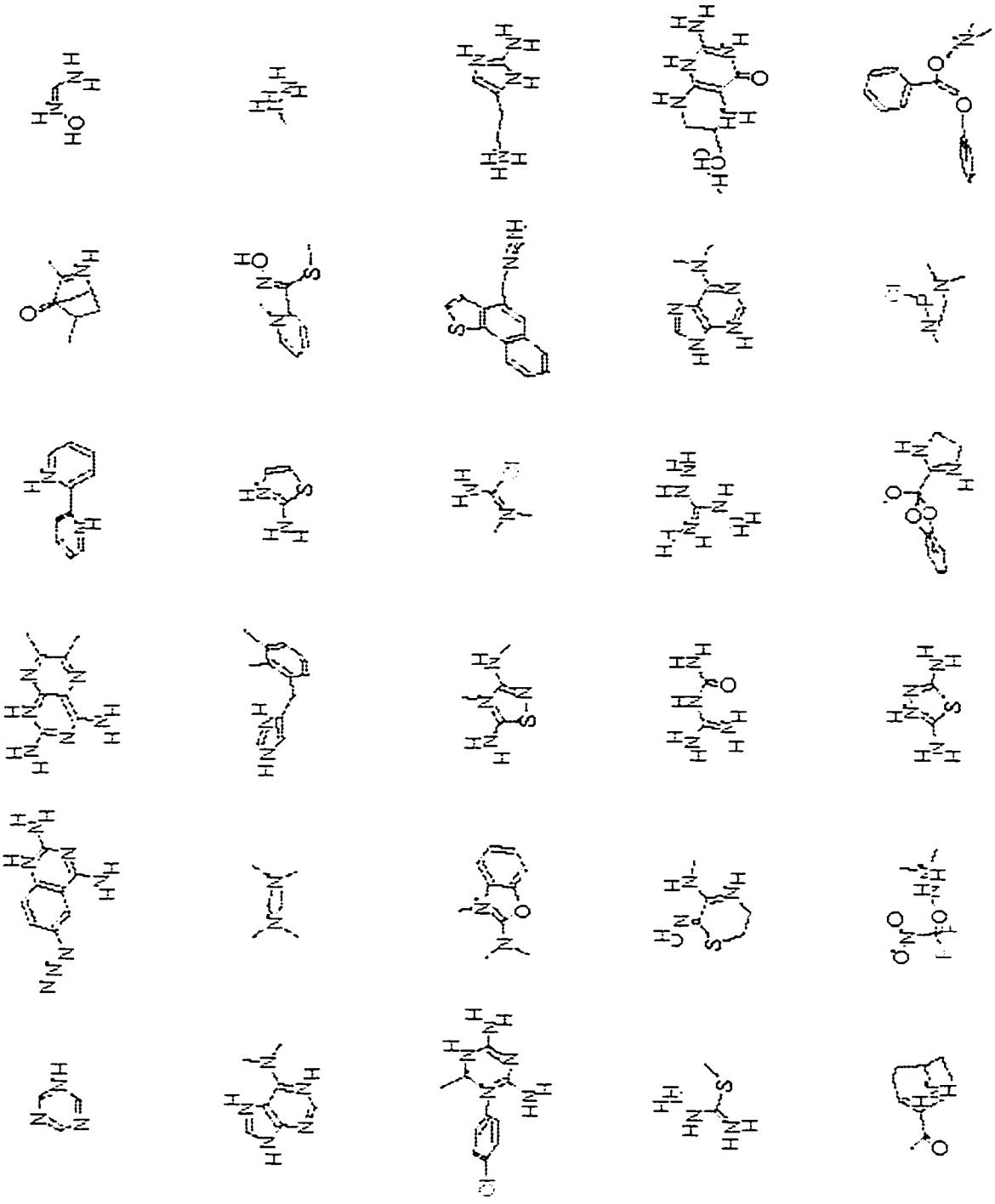


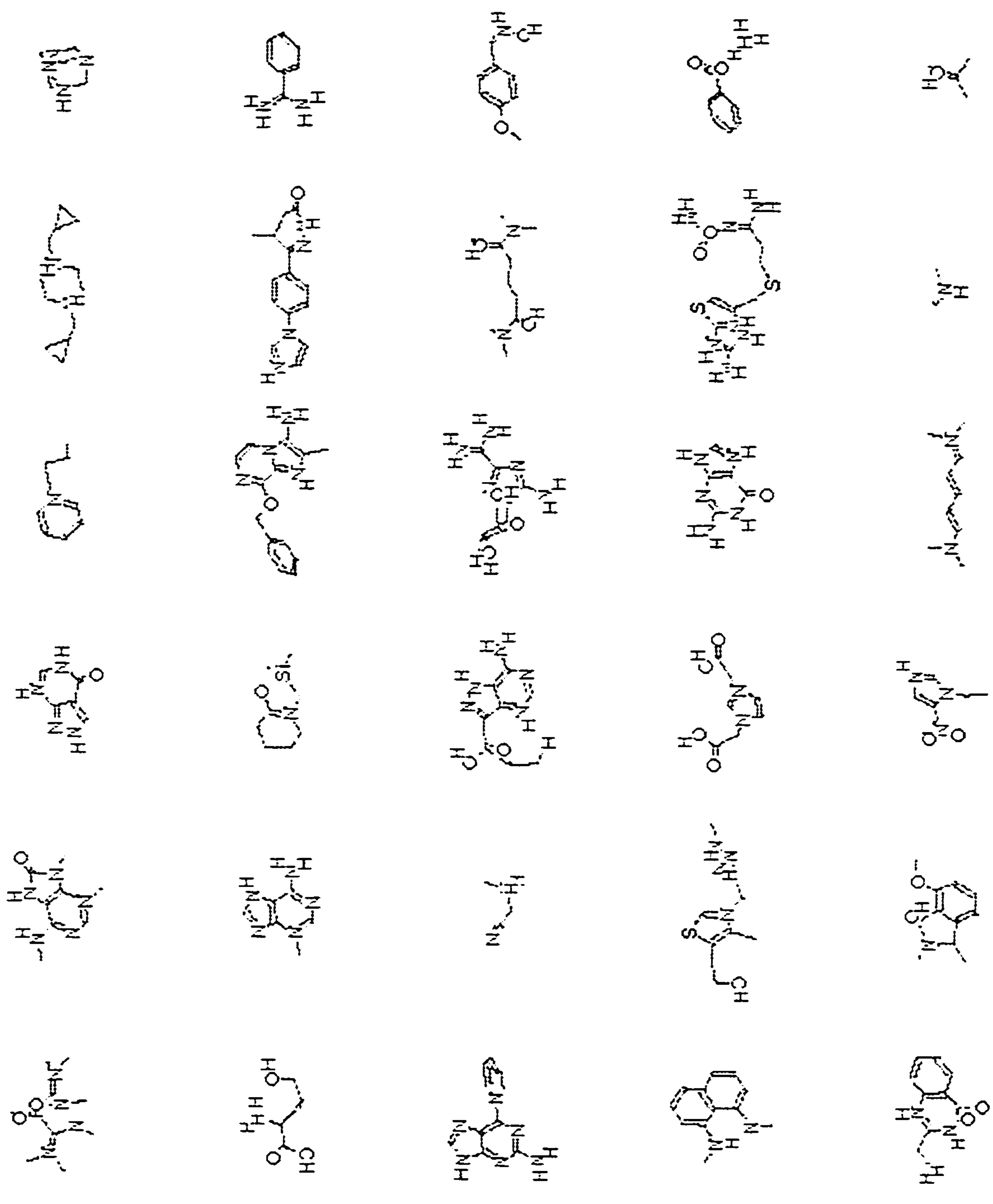


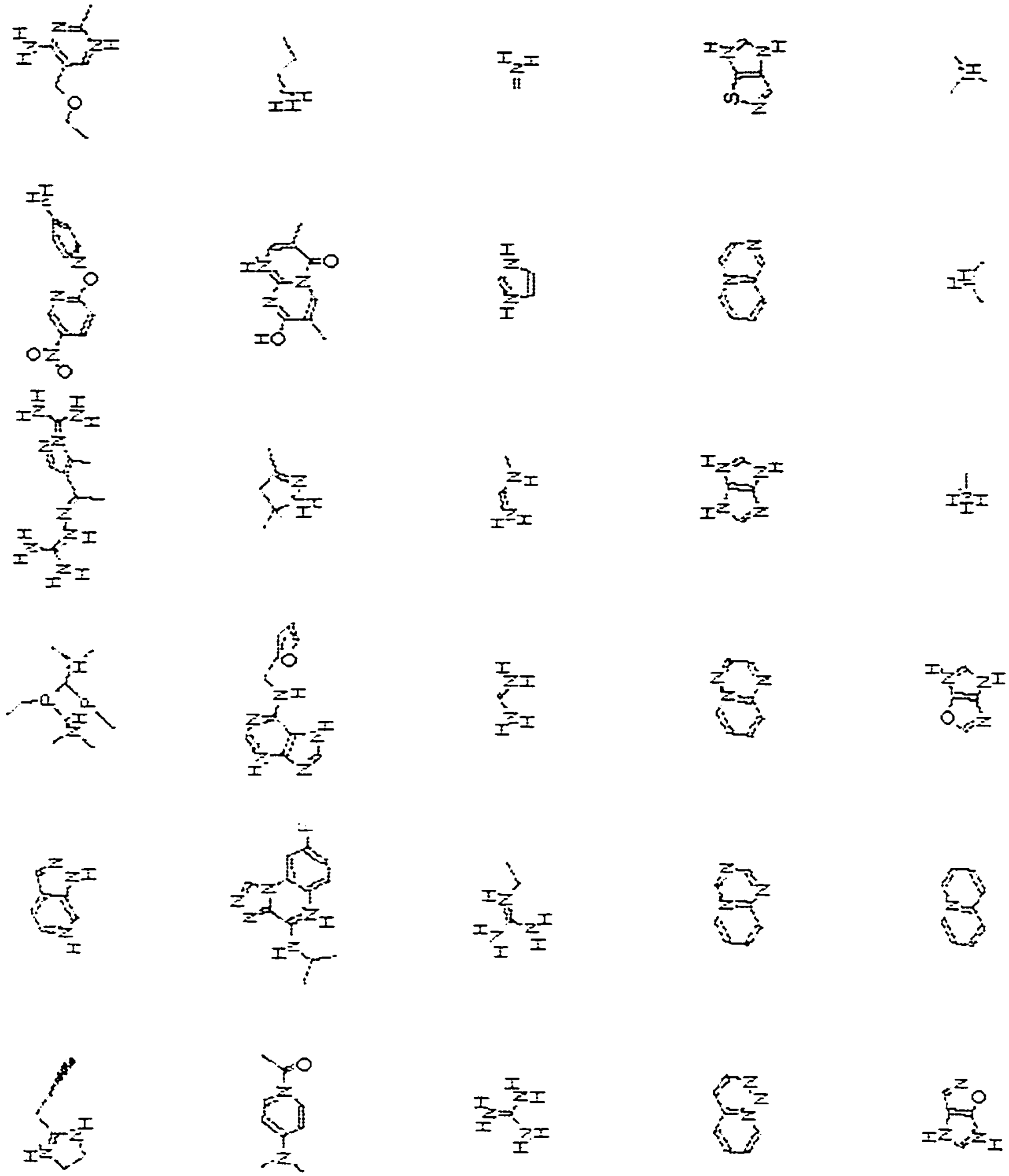


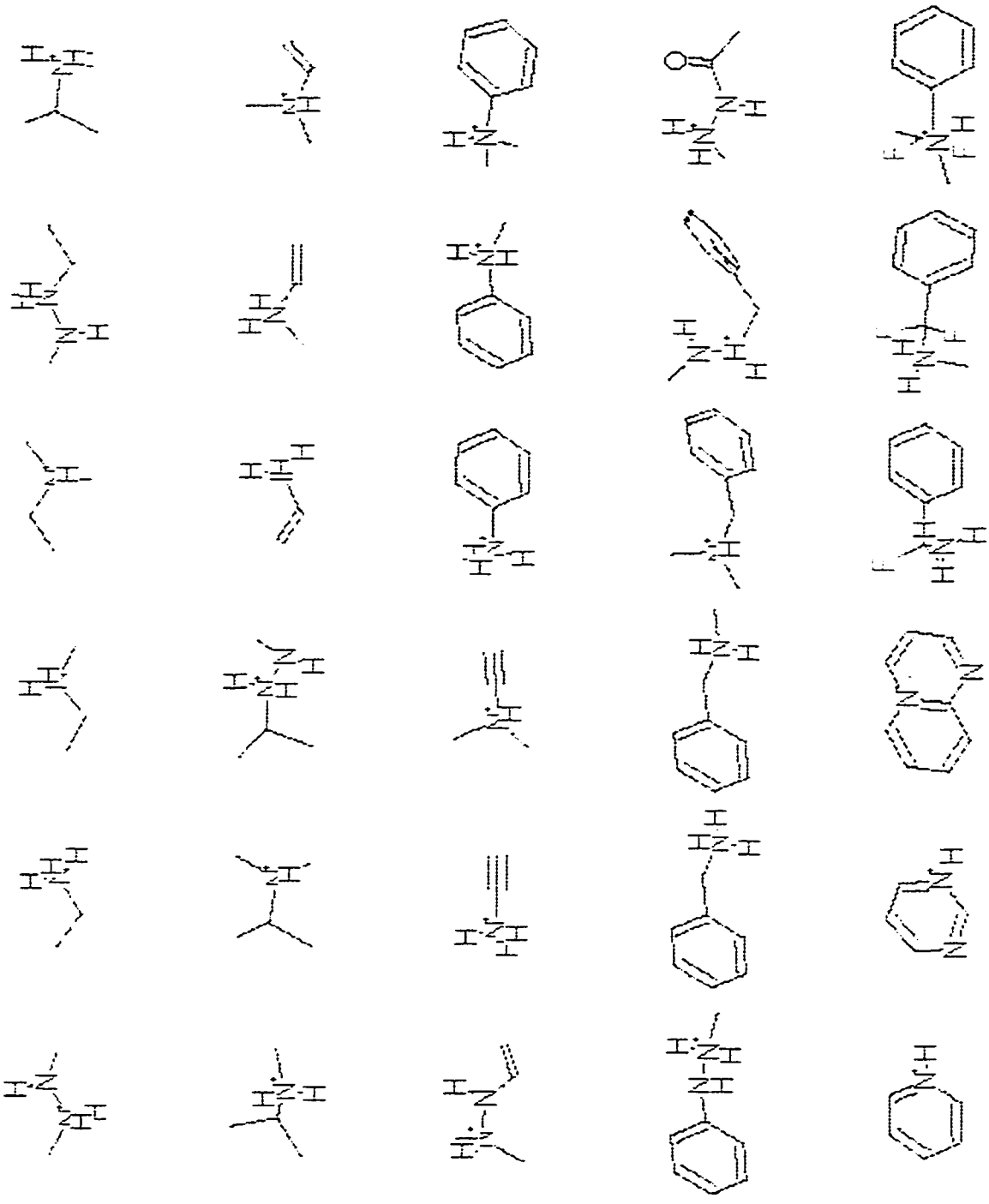


C.2 Cations

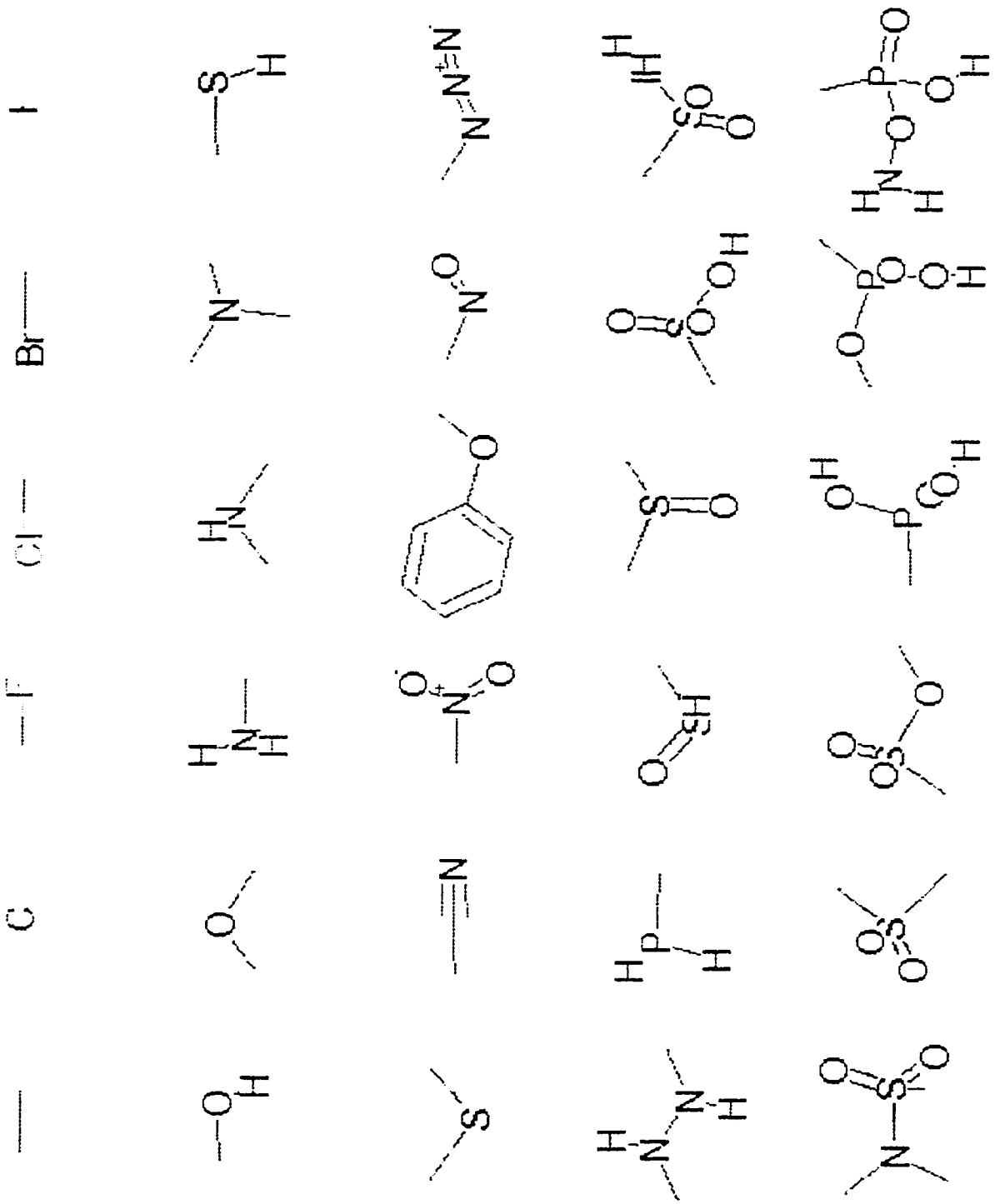


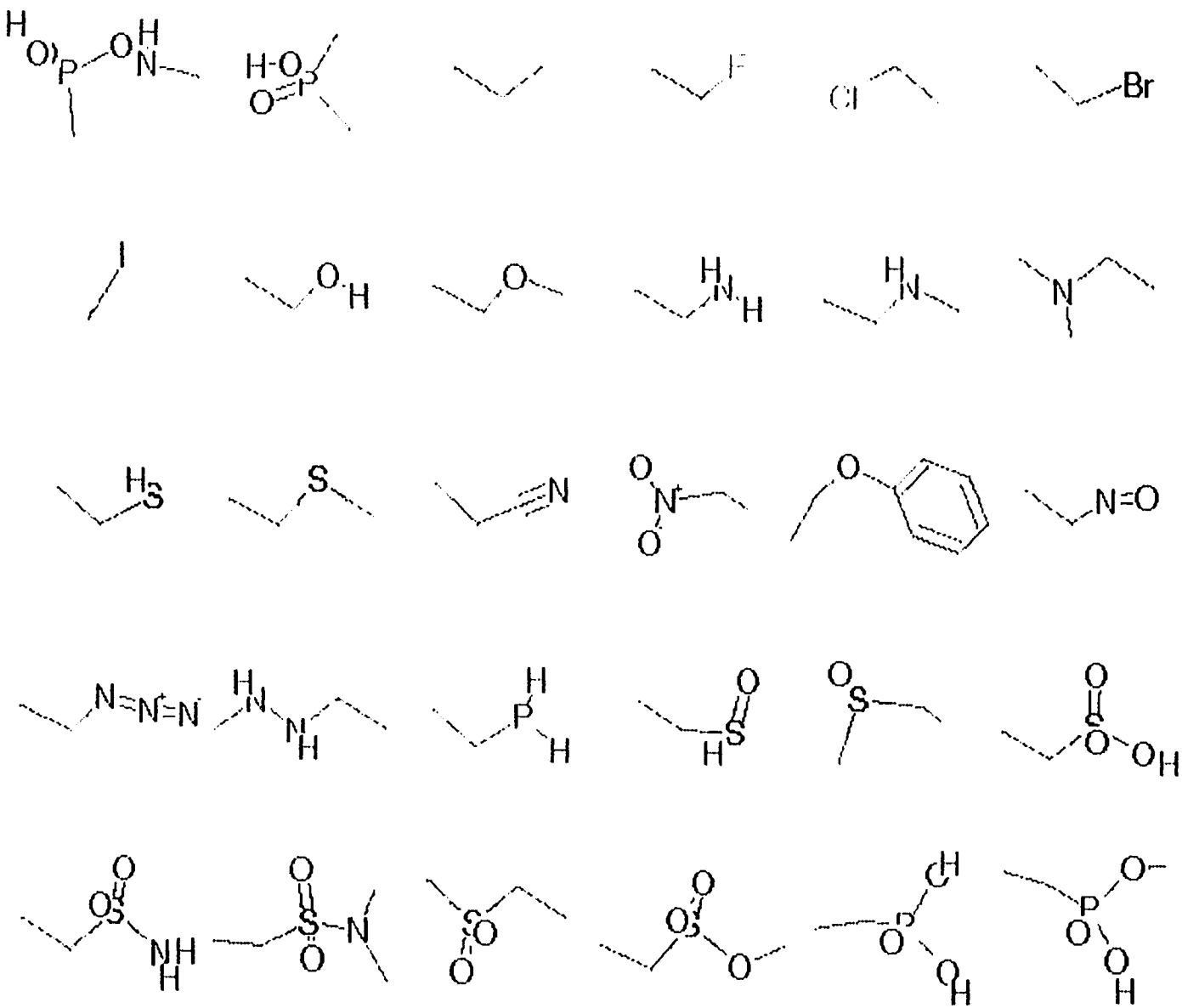


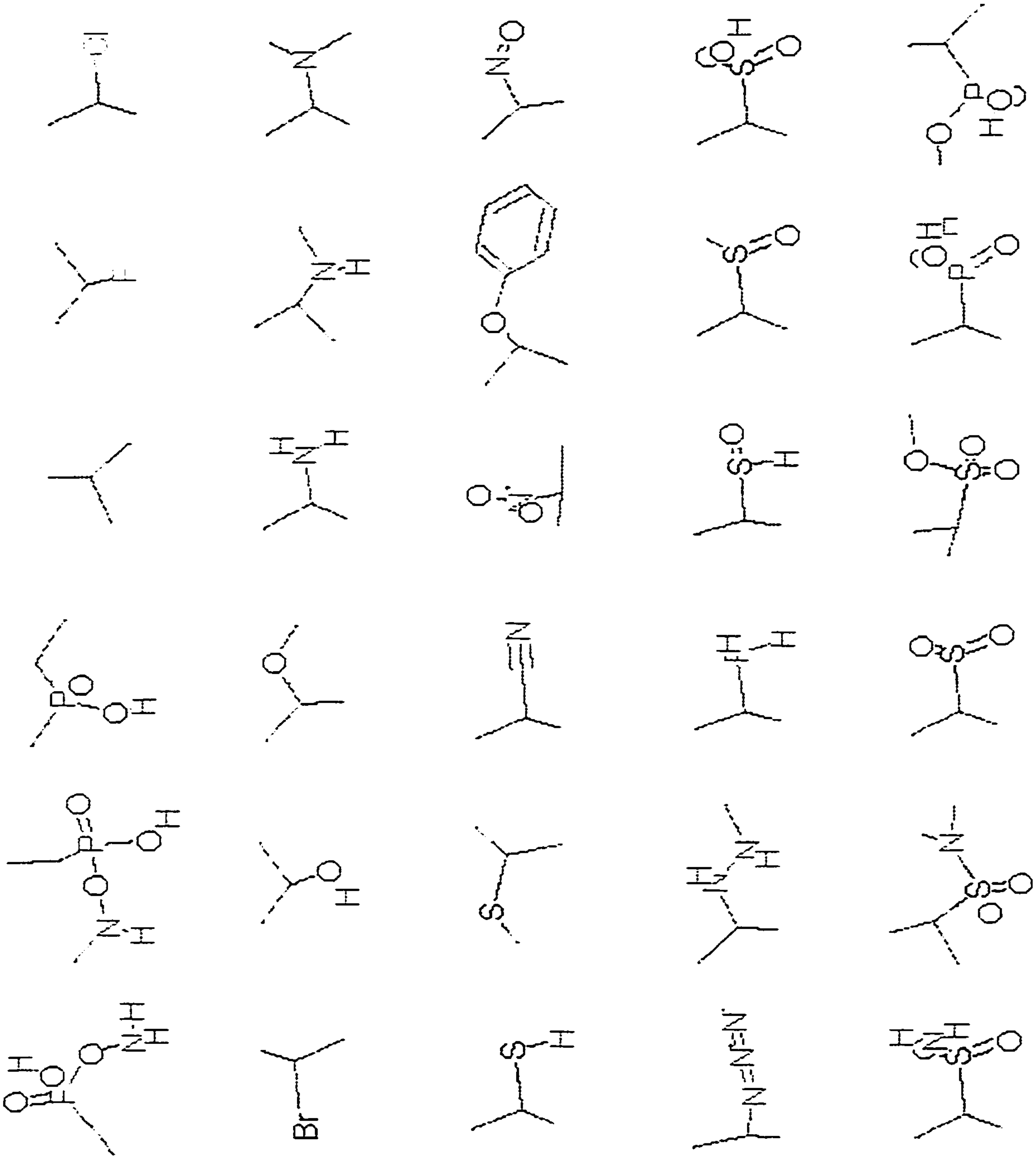


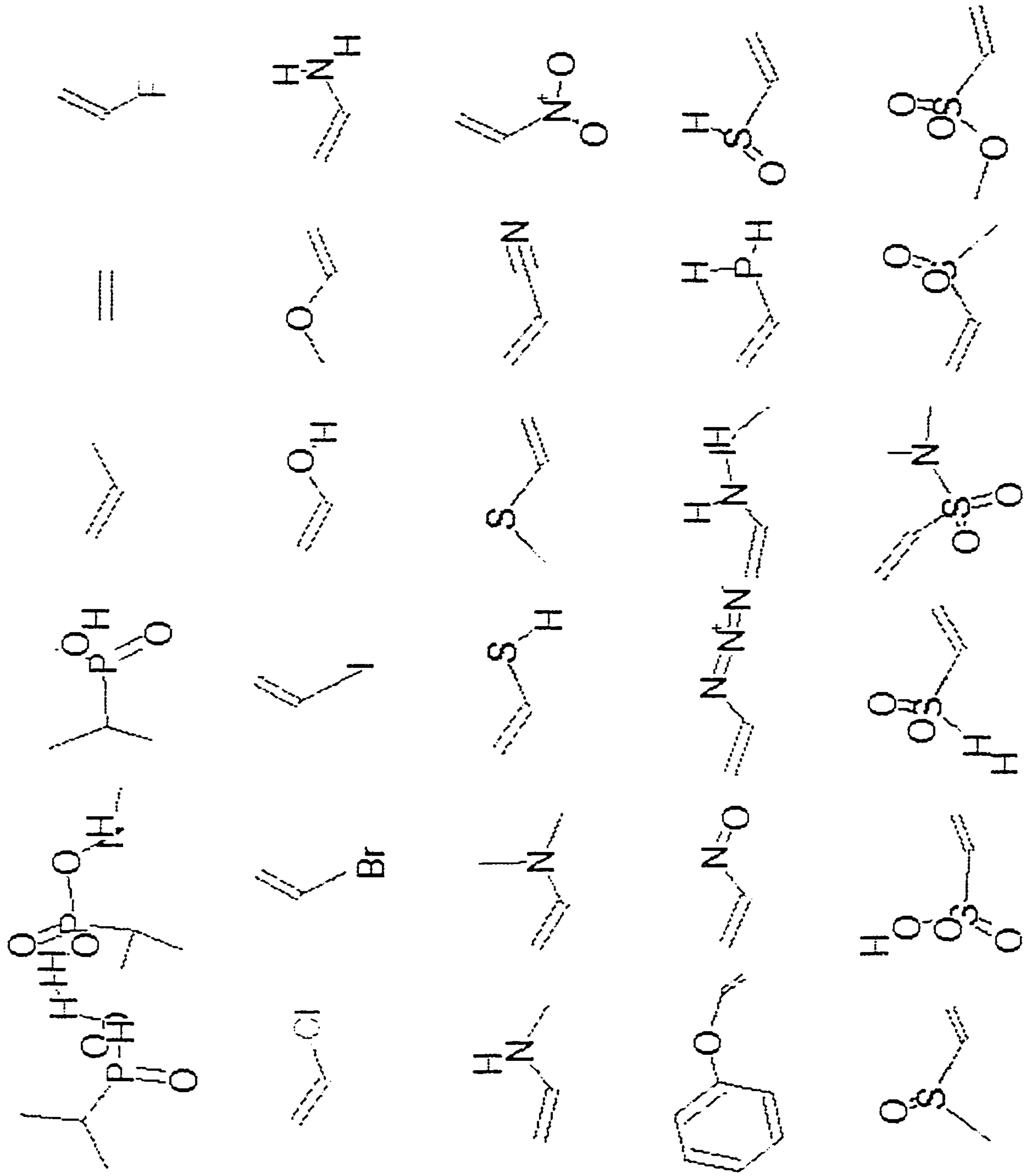


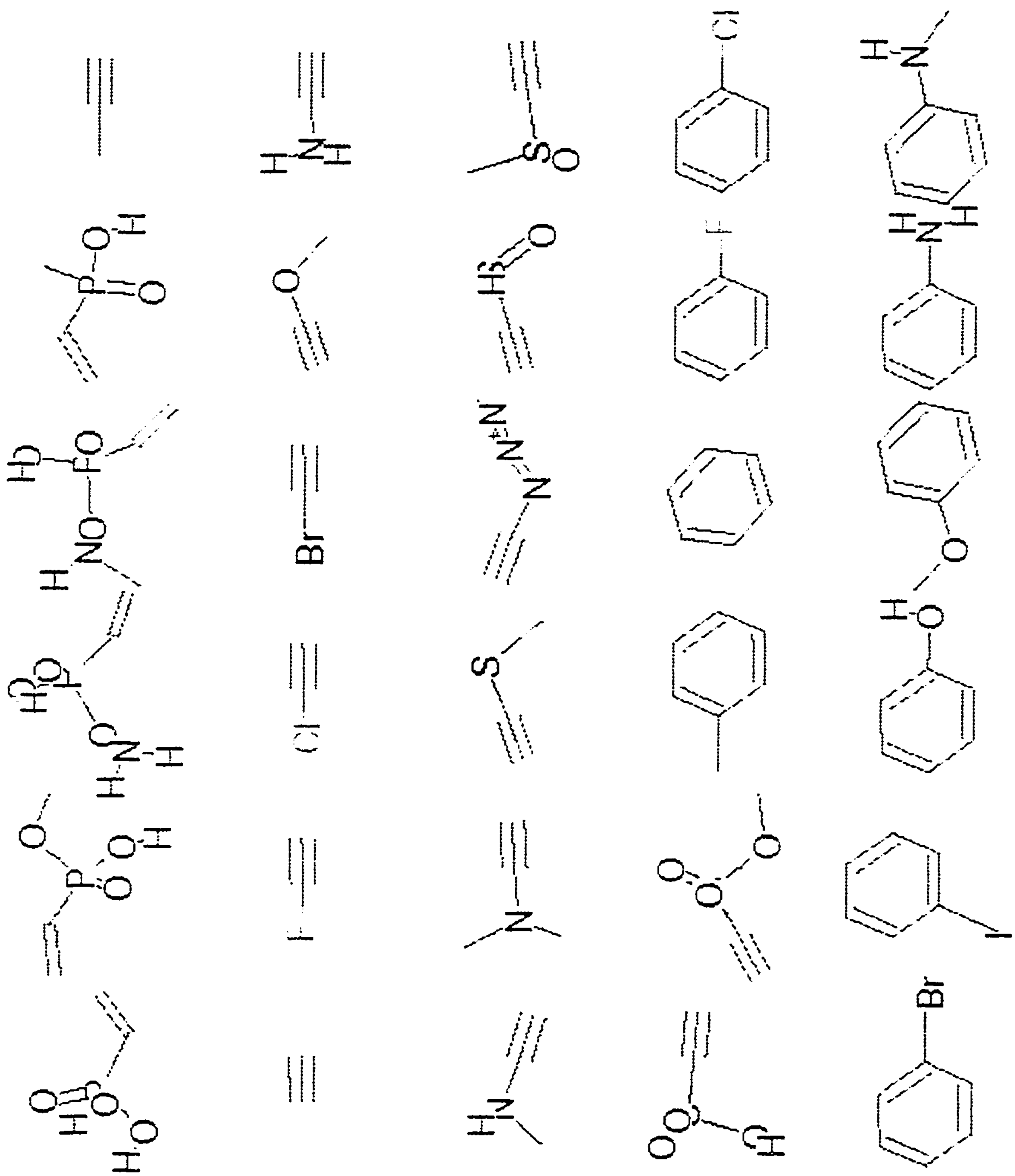
C.3 Scaffold/Functional Group

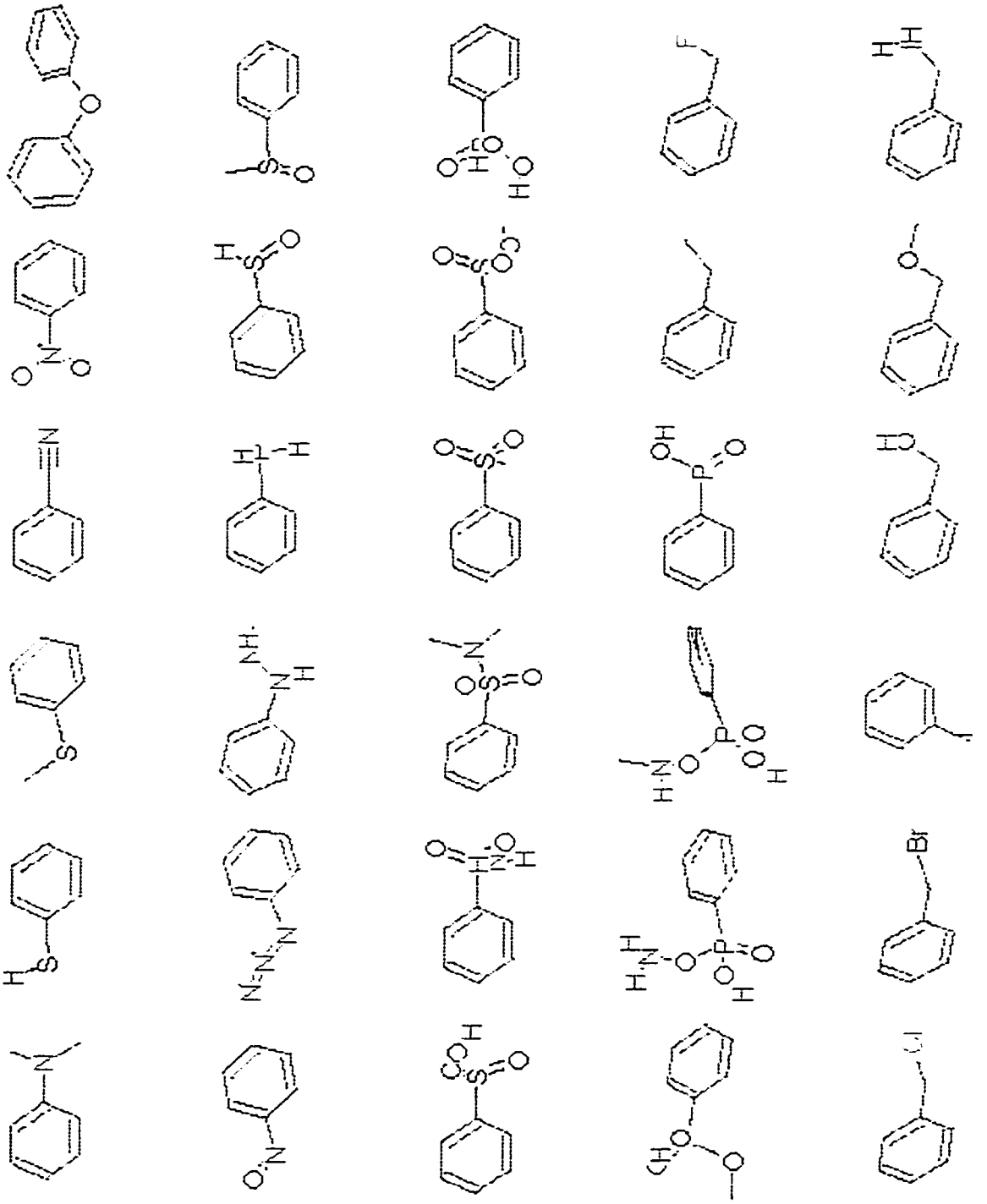


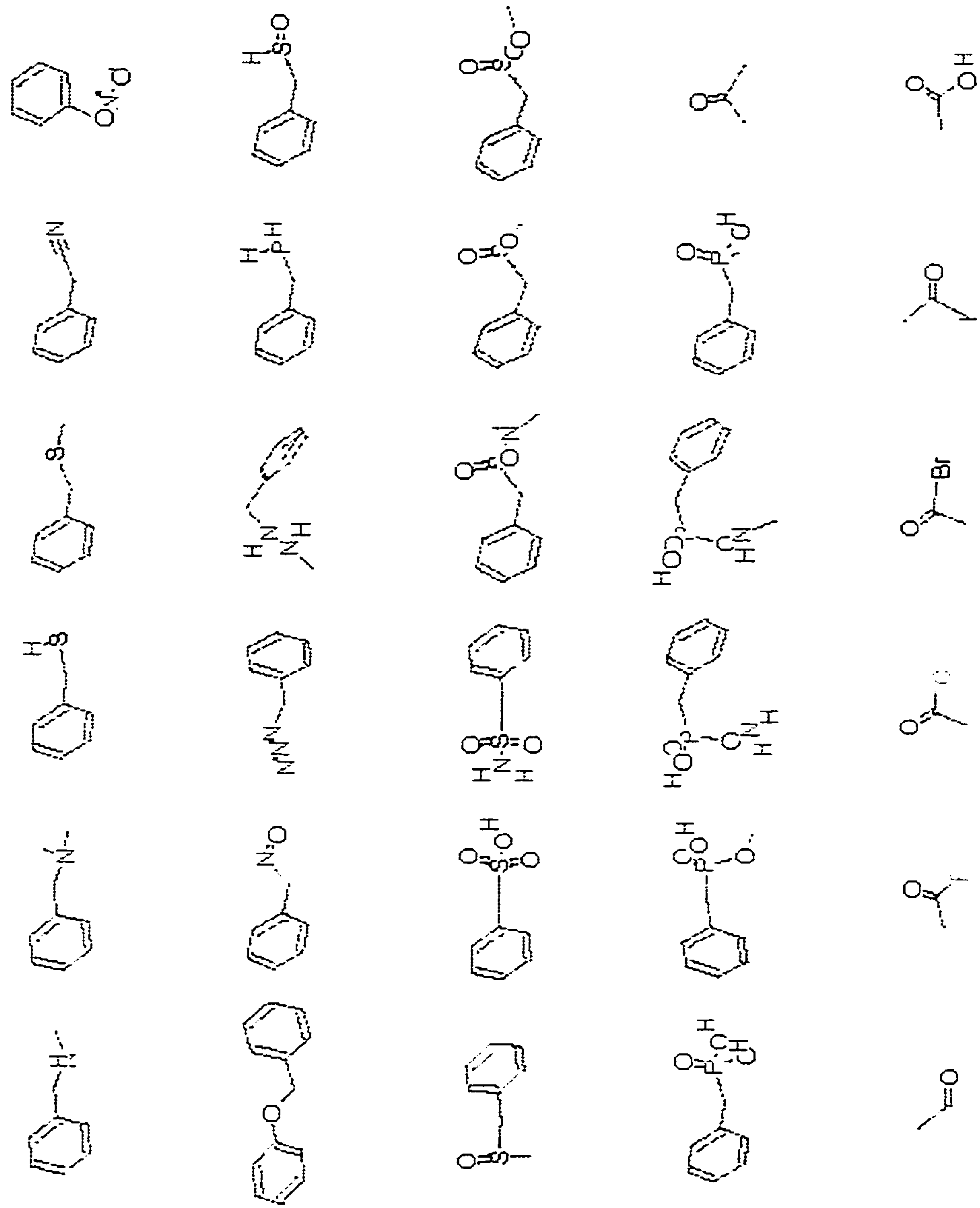


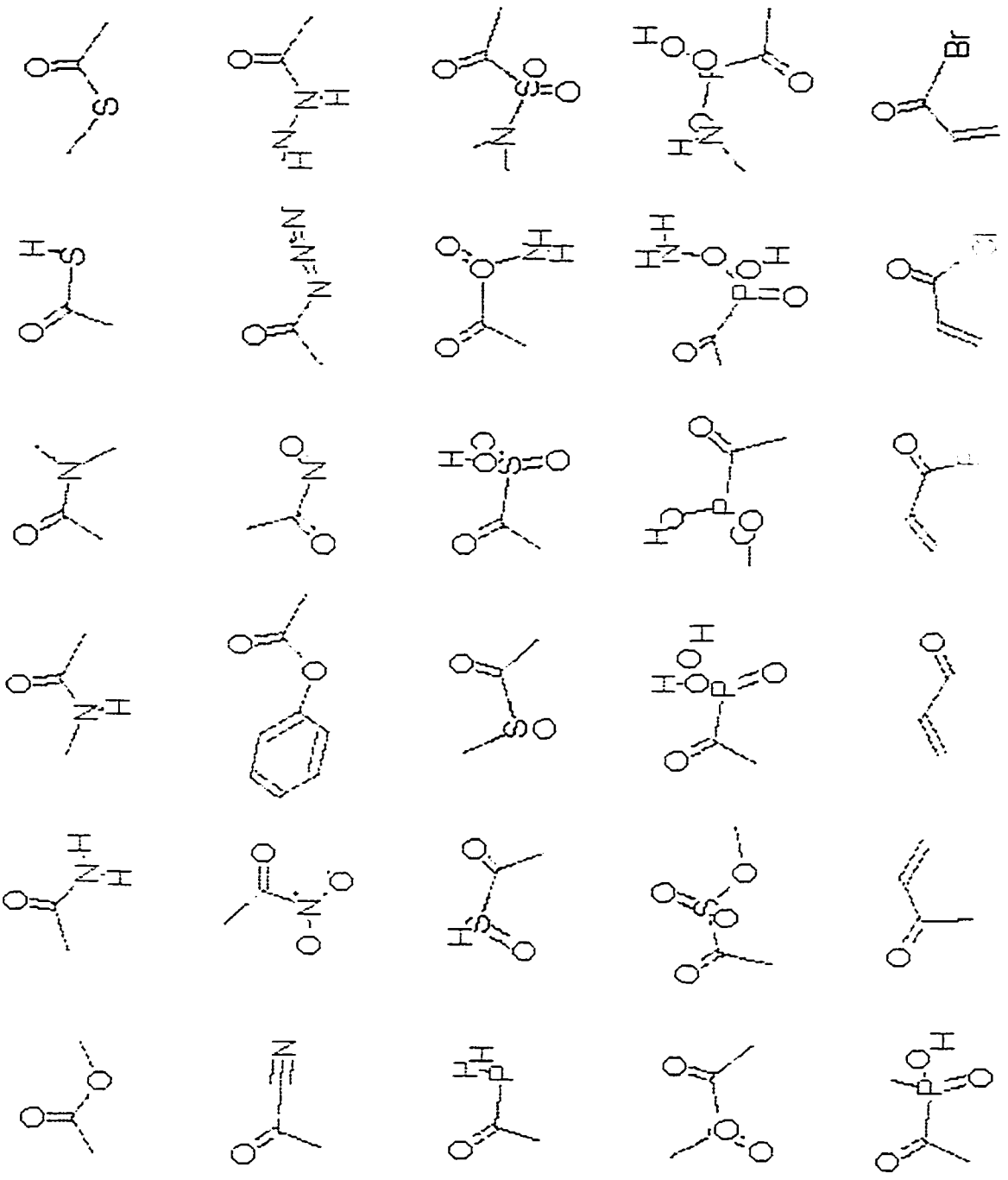


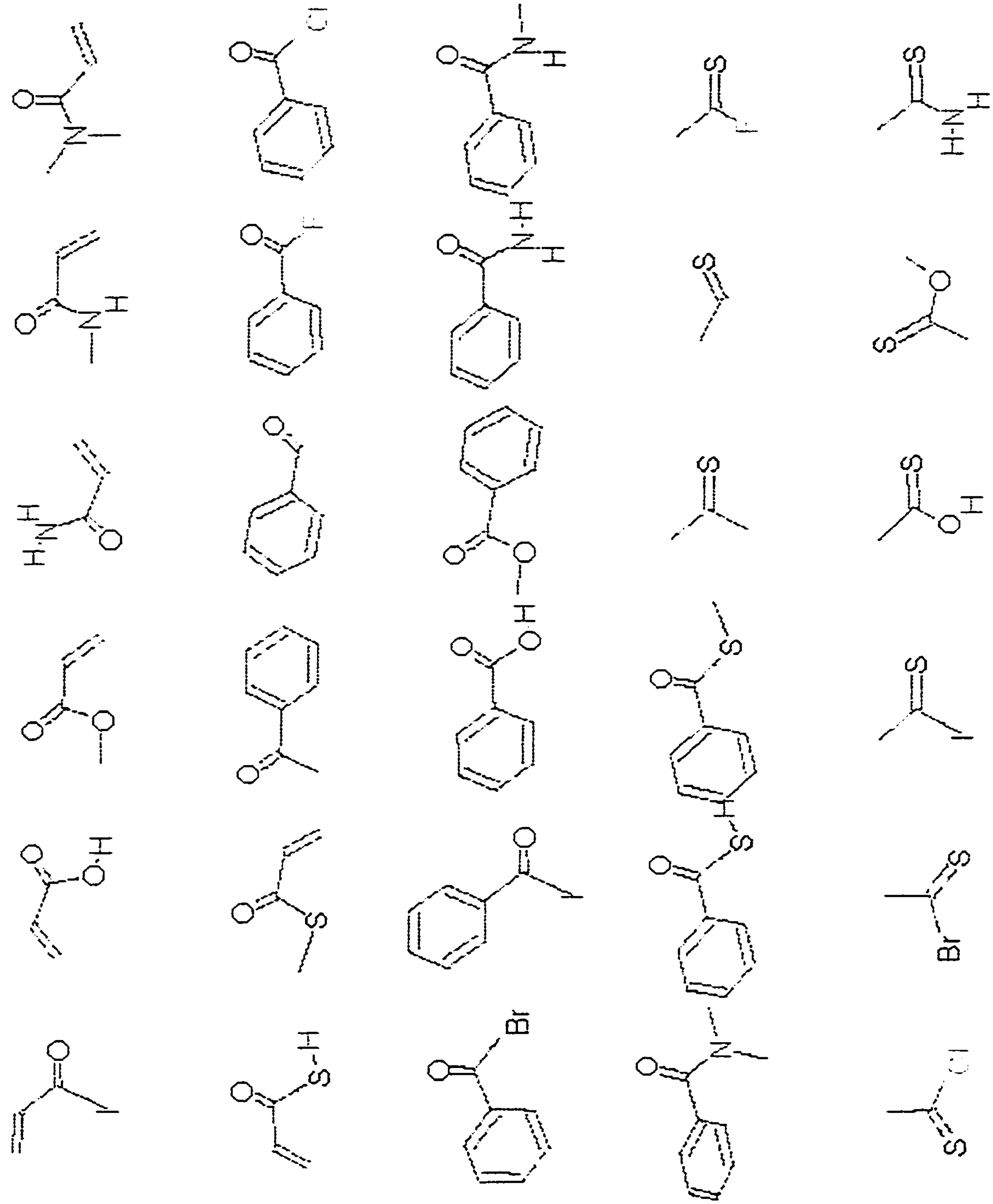


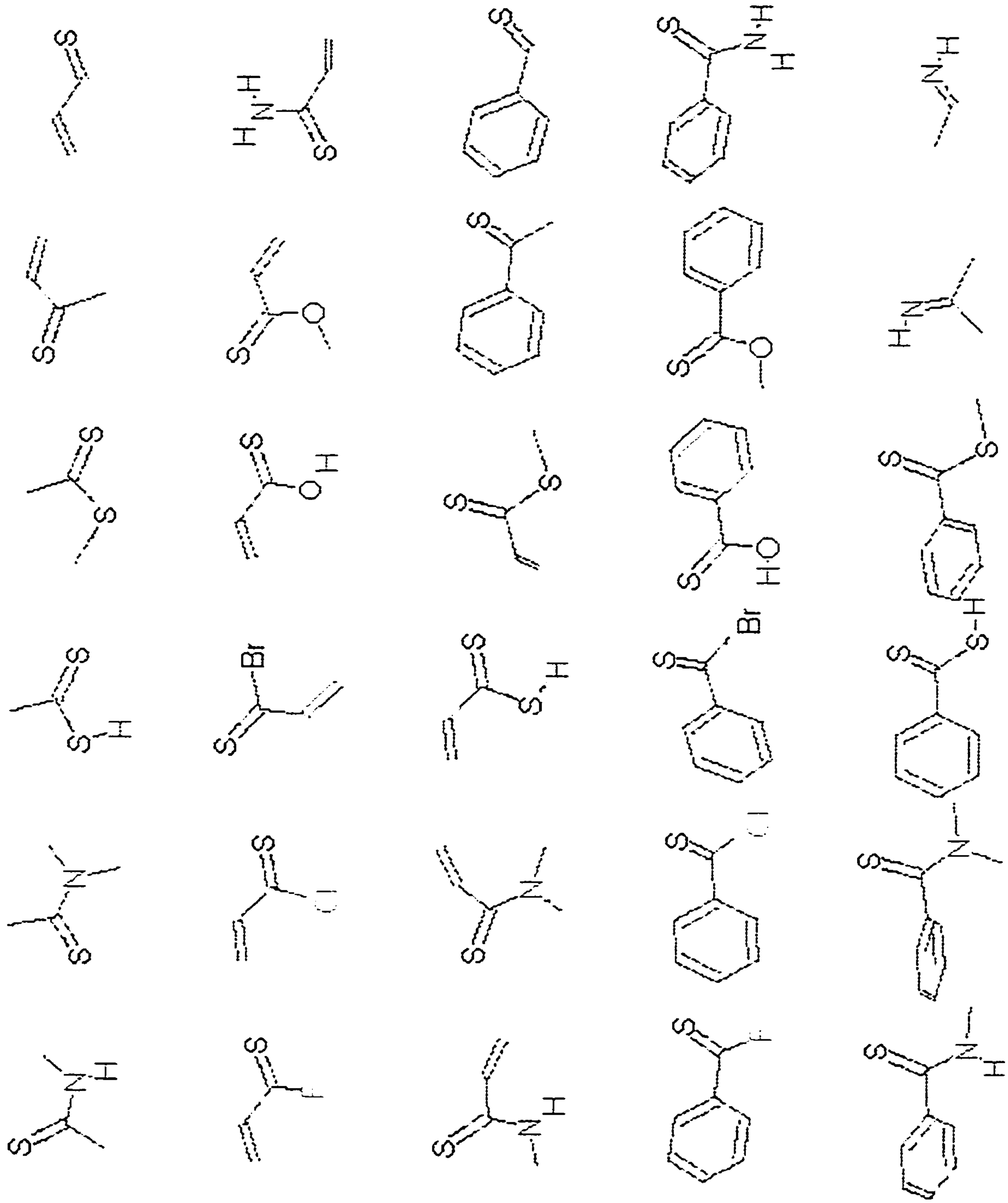


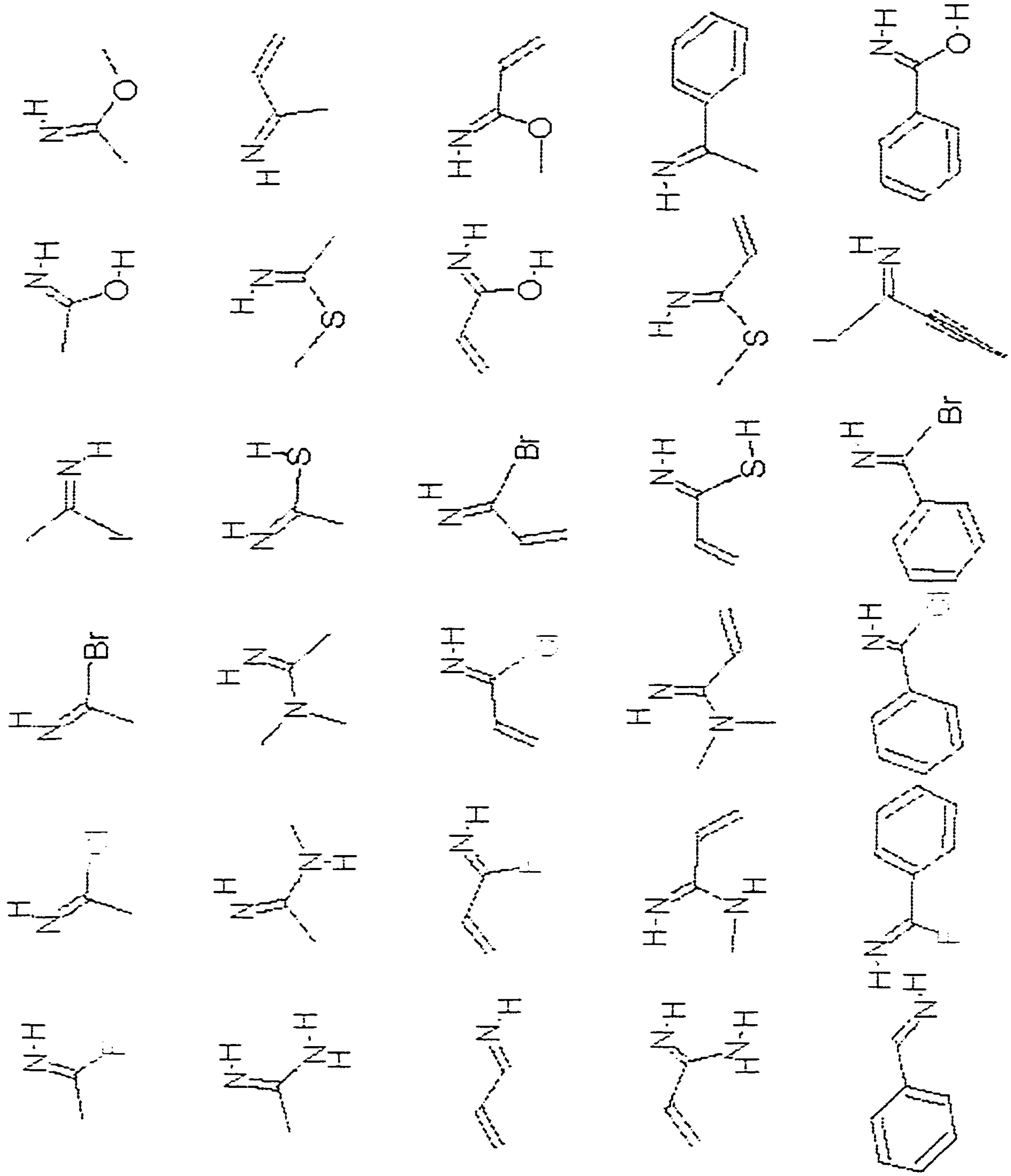


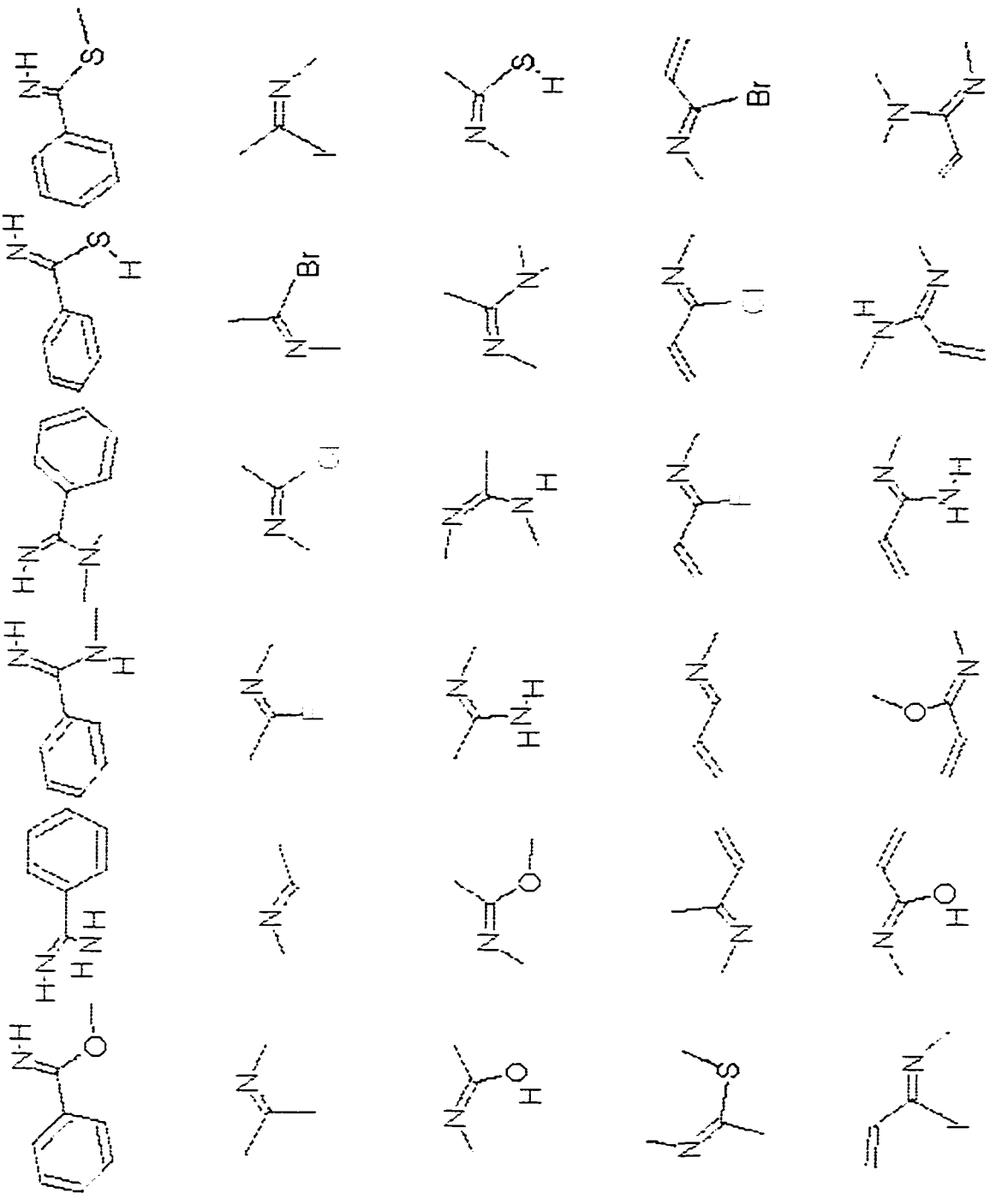


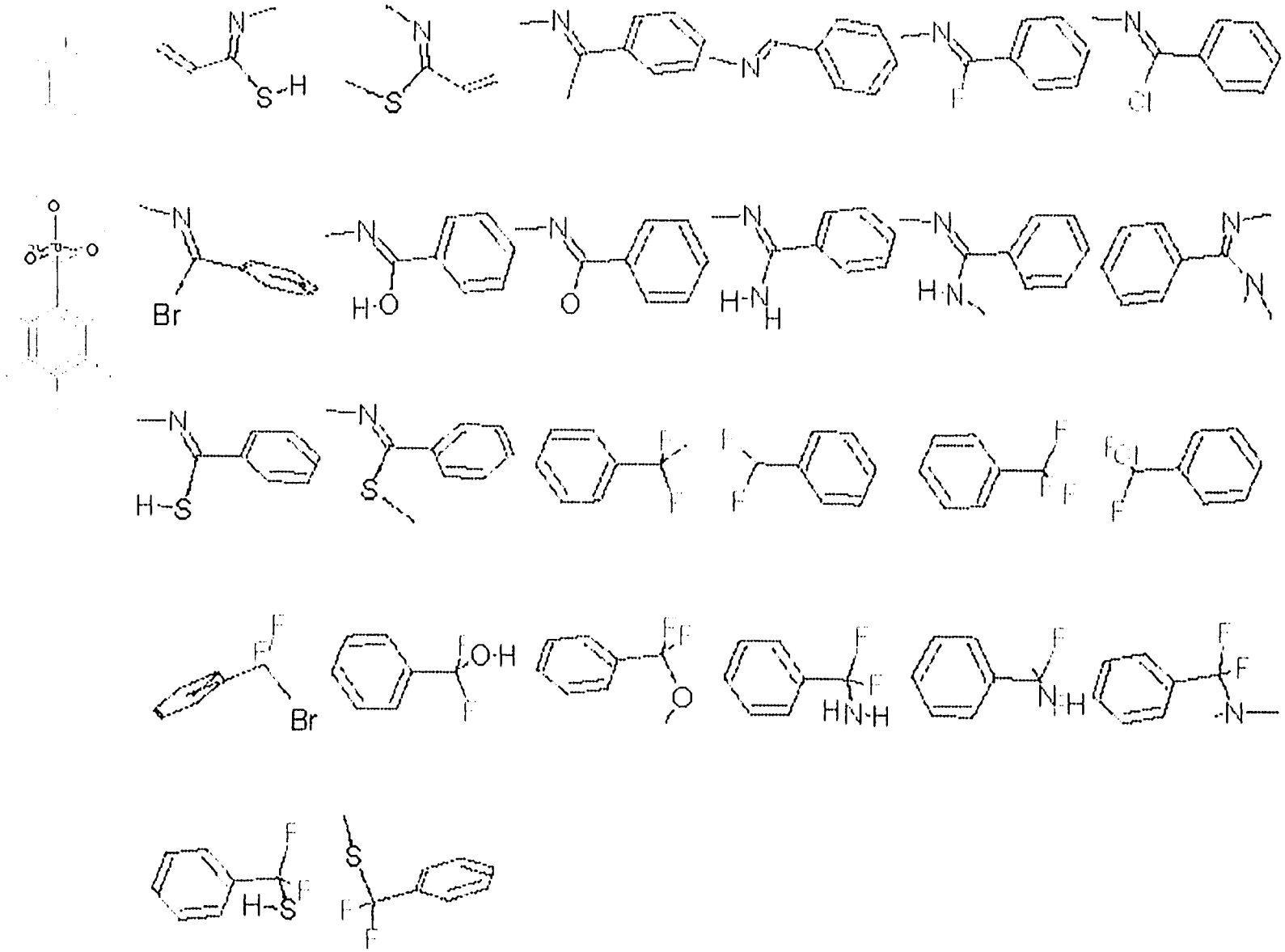




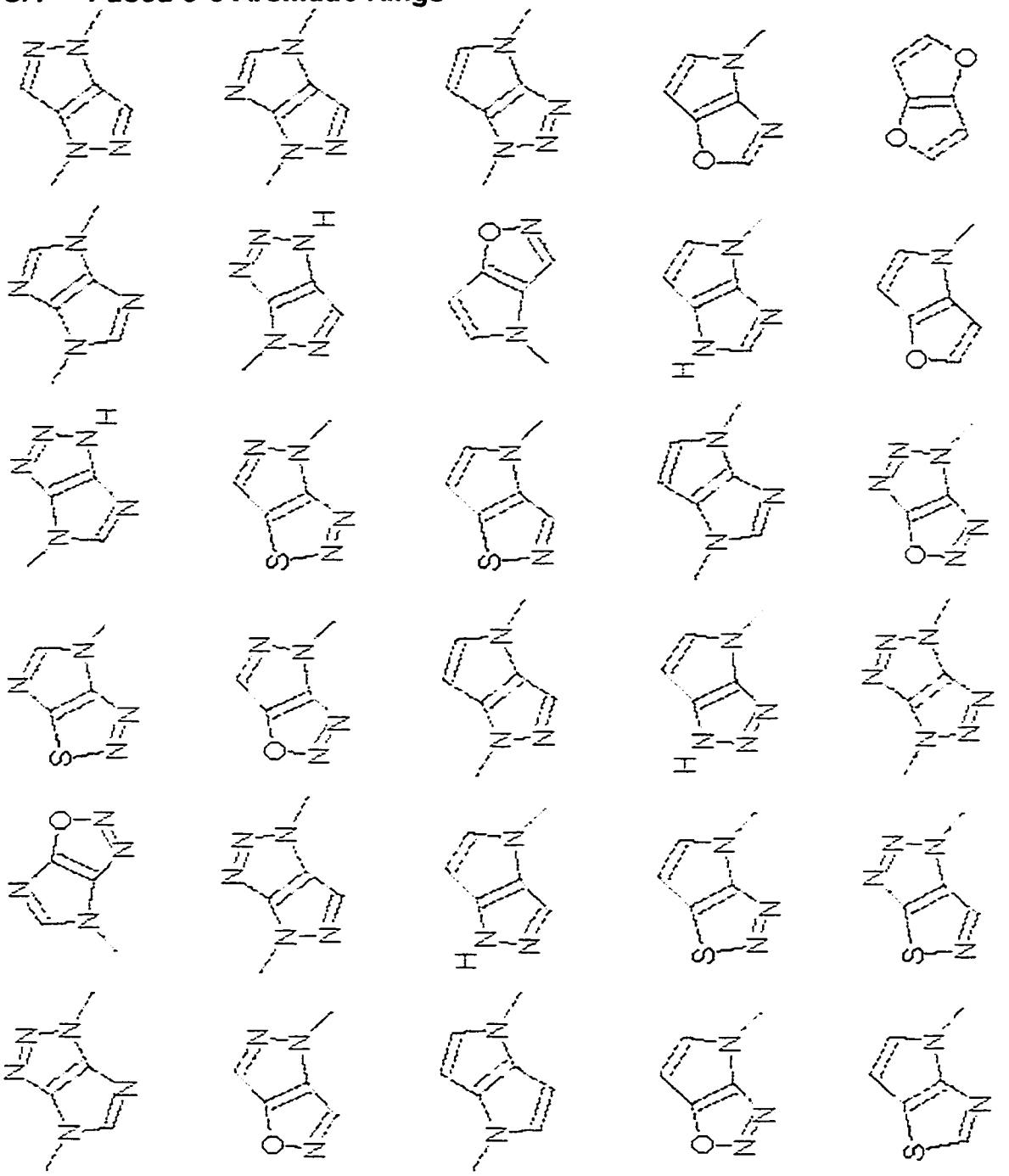


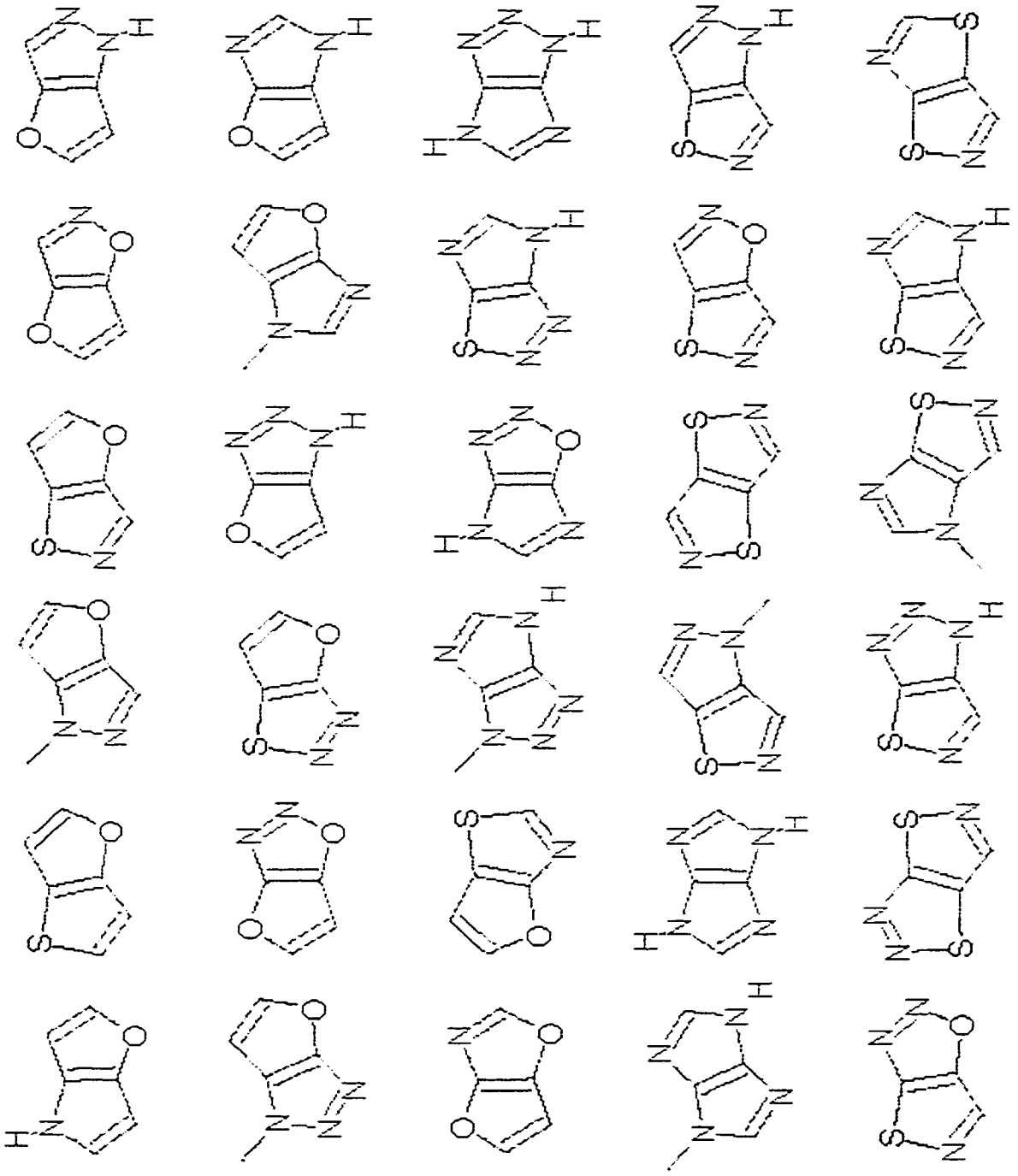


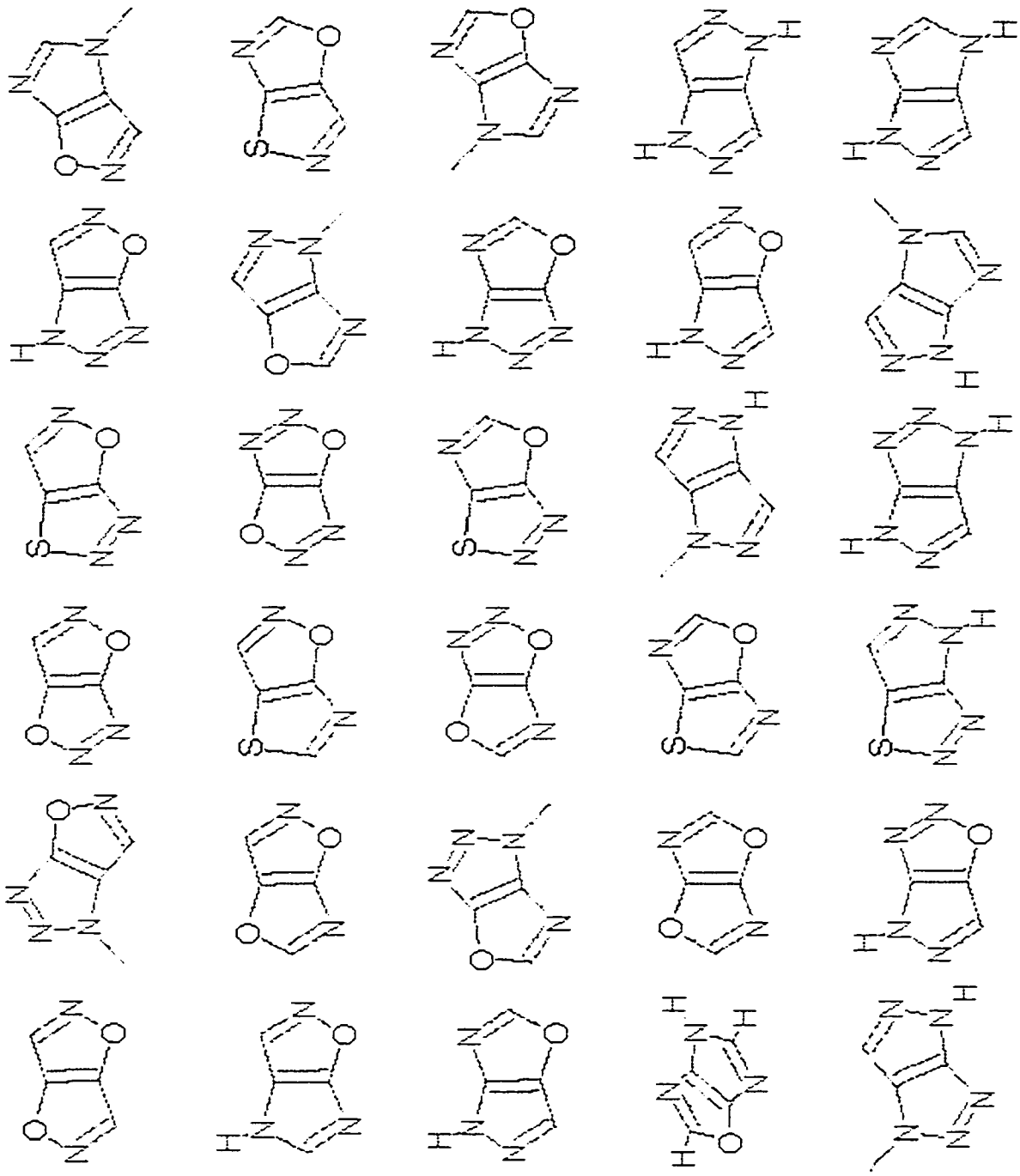


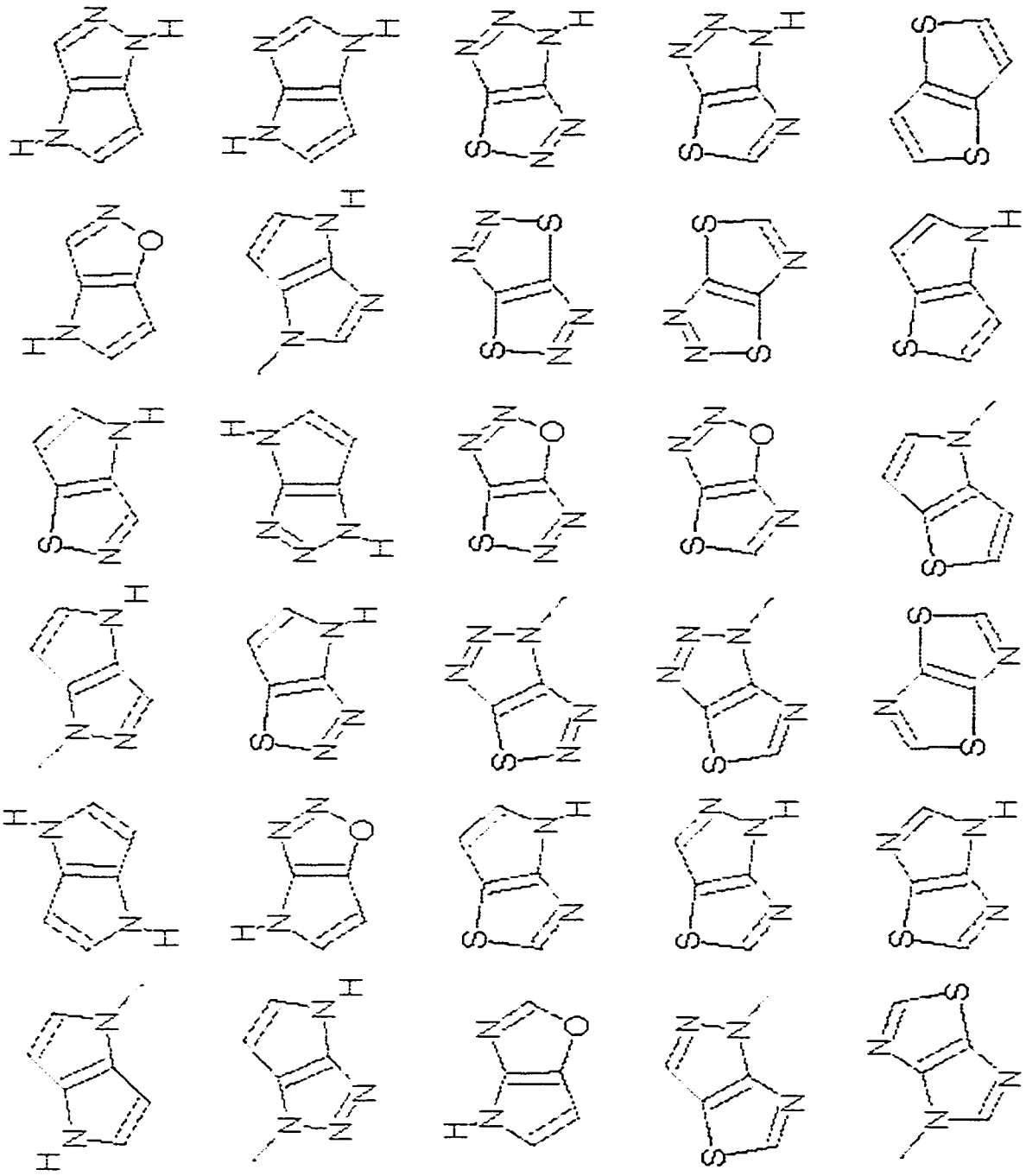


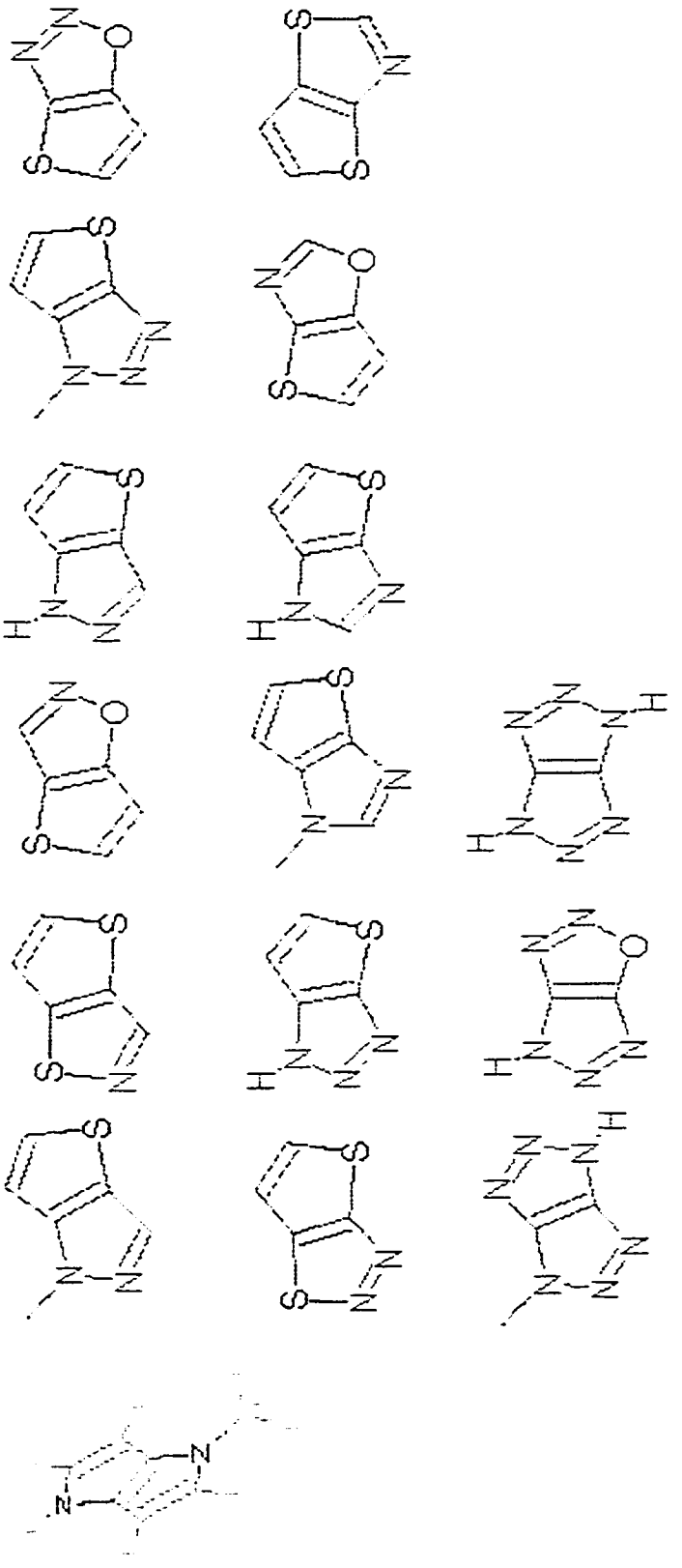
C.4 Fused 5-5 Aromatic Rings



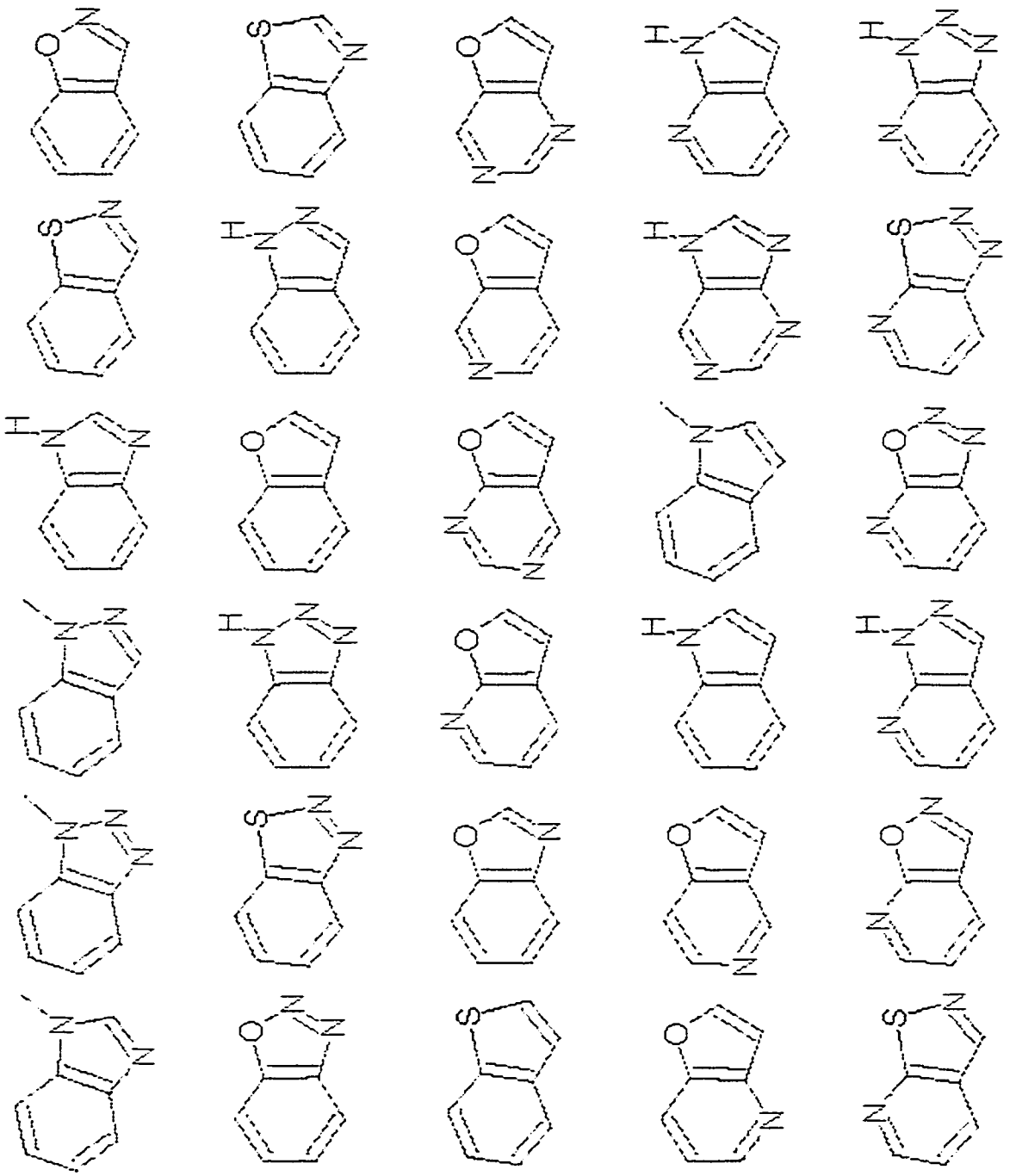


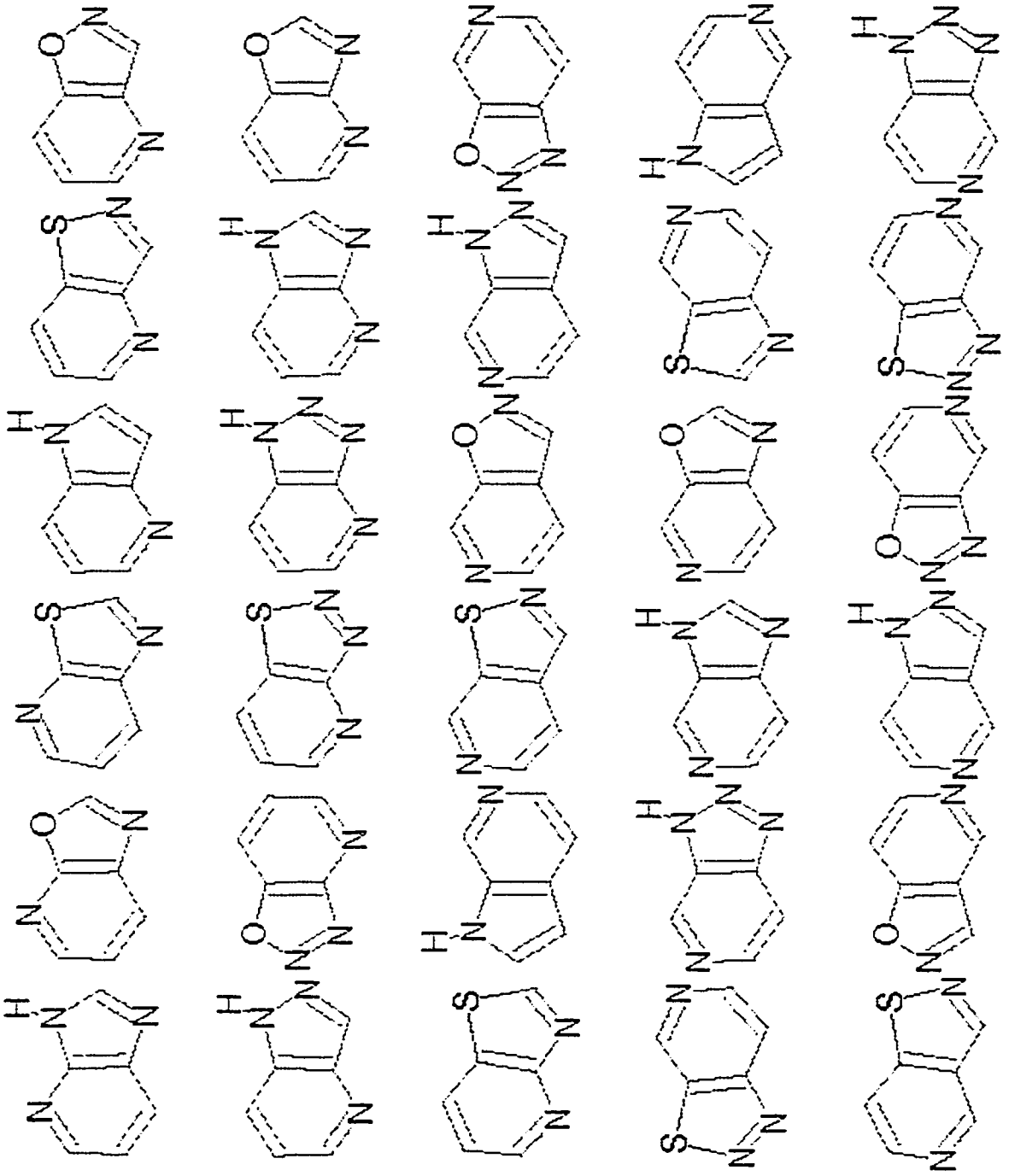


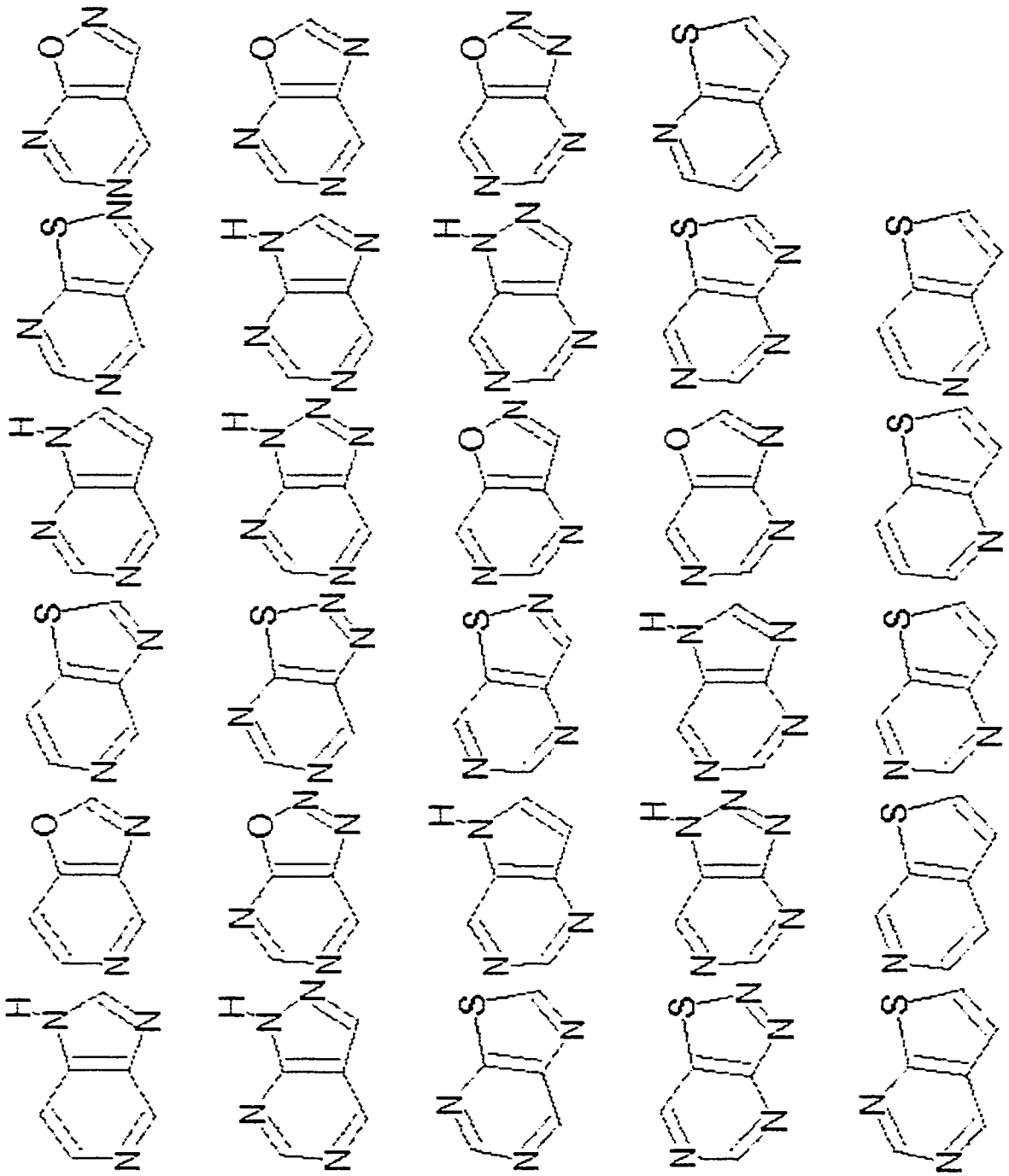




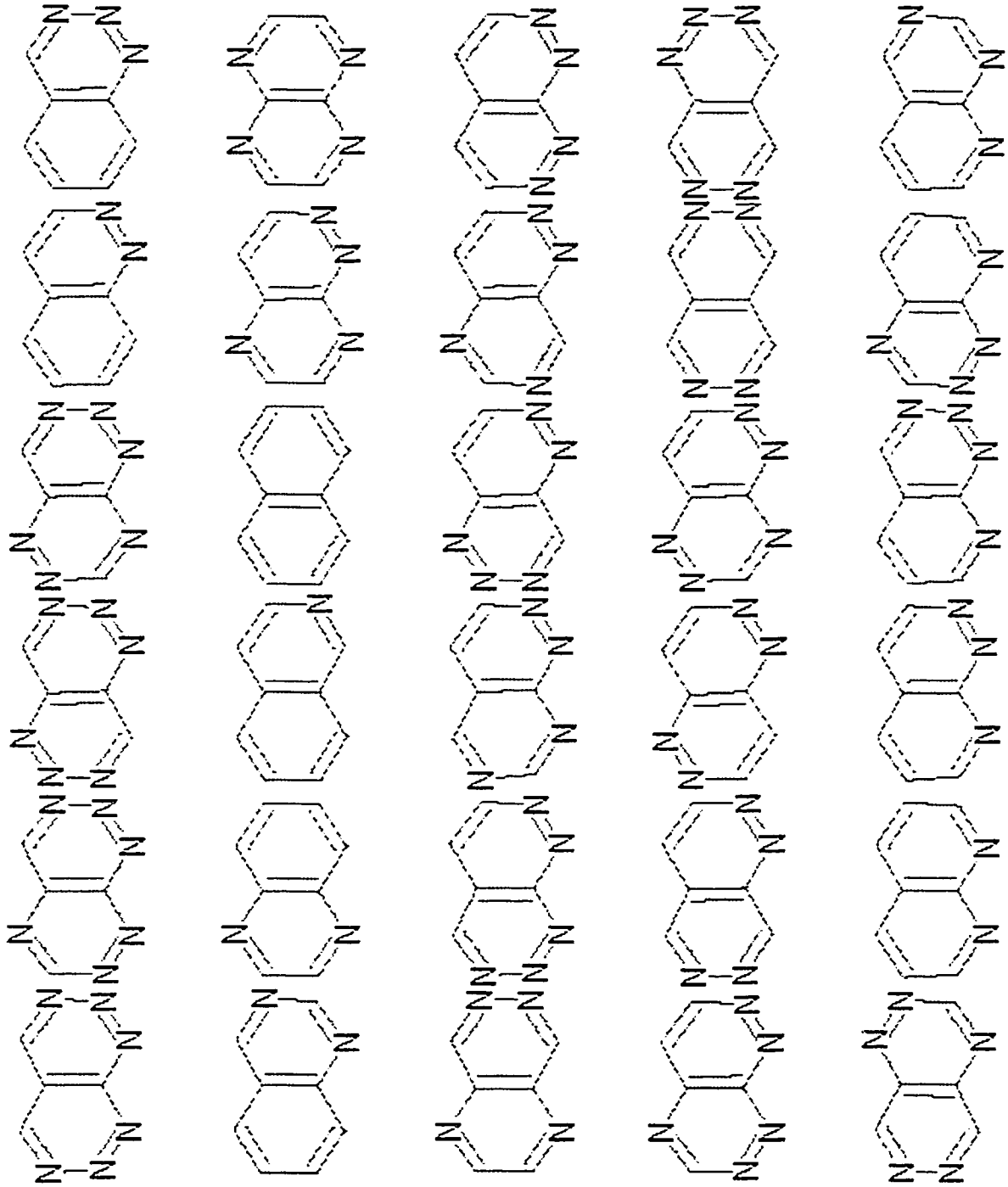
C.5 Fused 6-5 Aromatic Rings

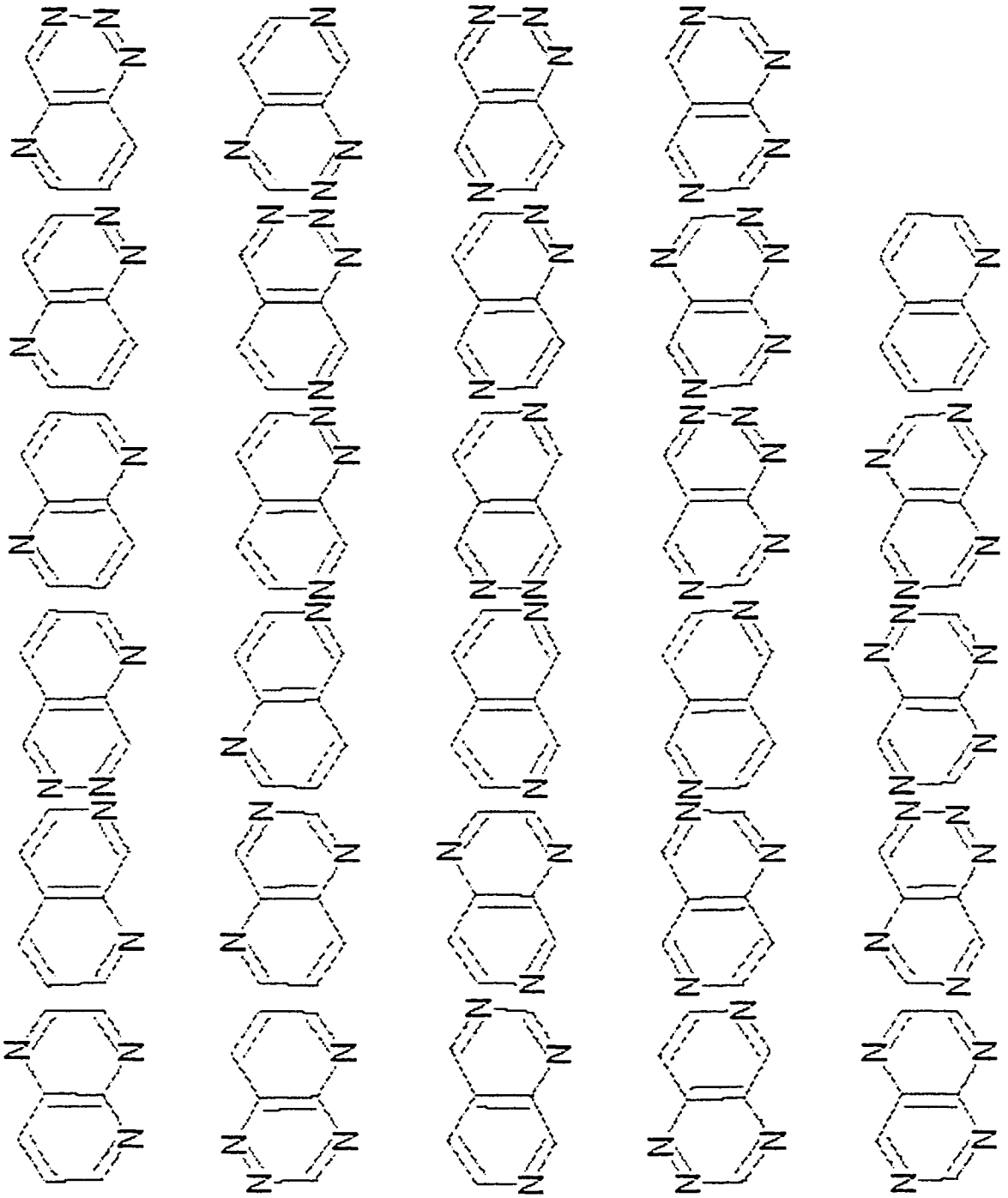




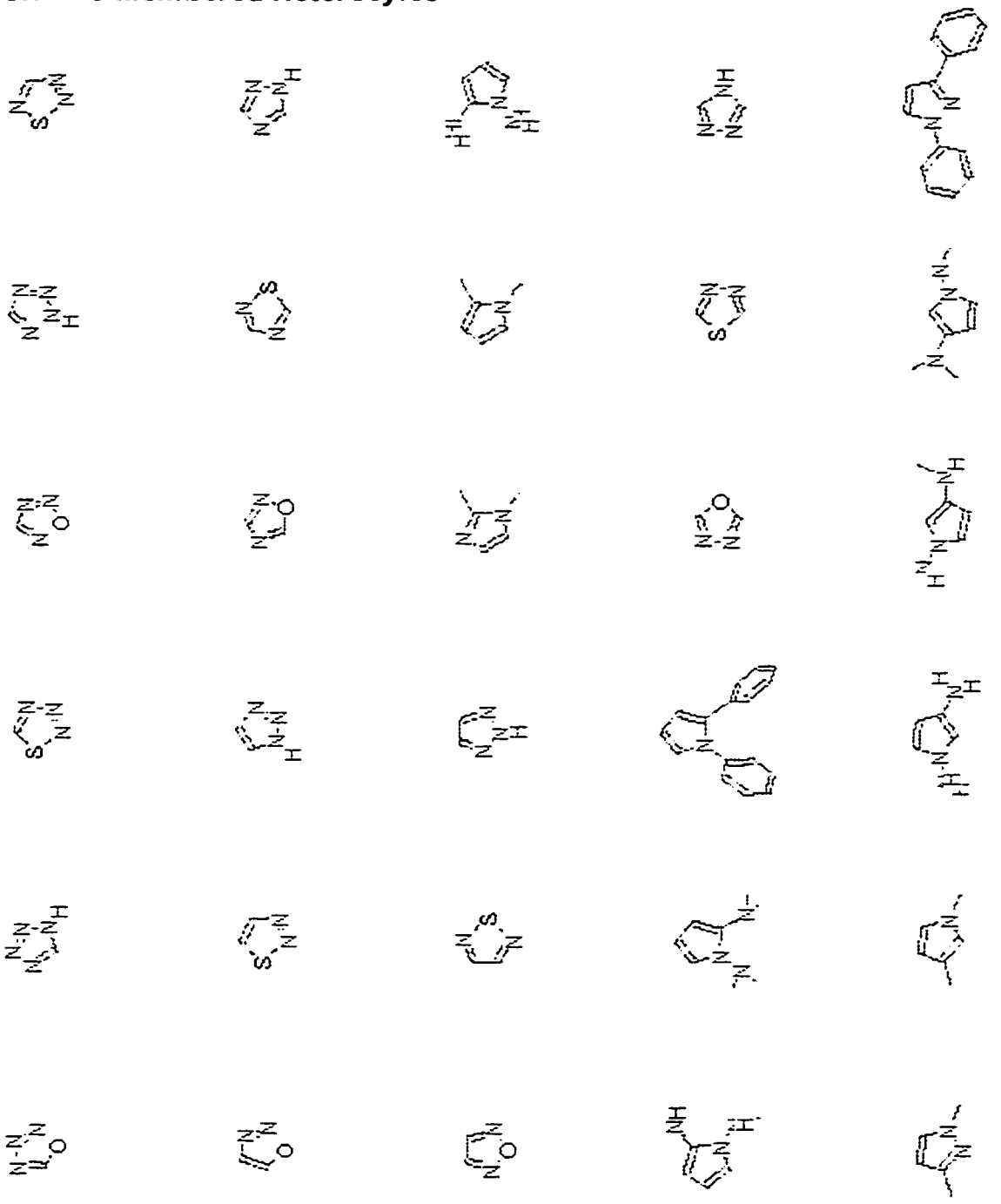


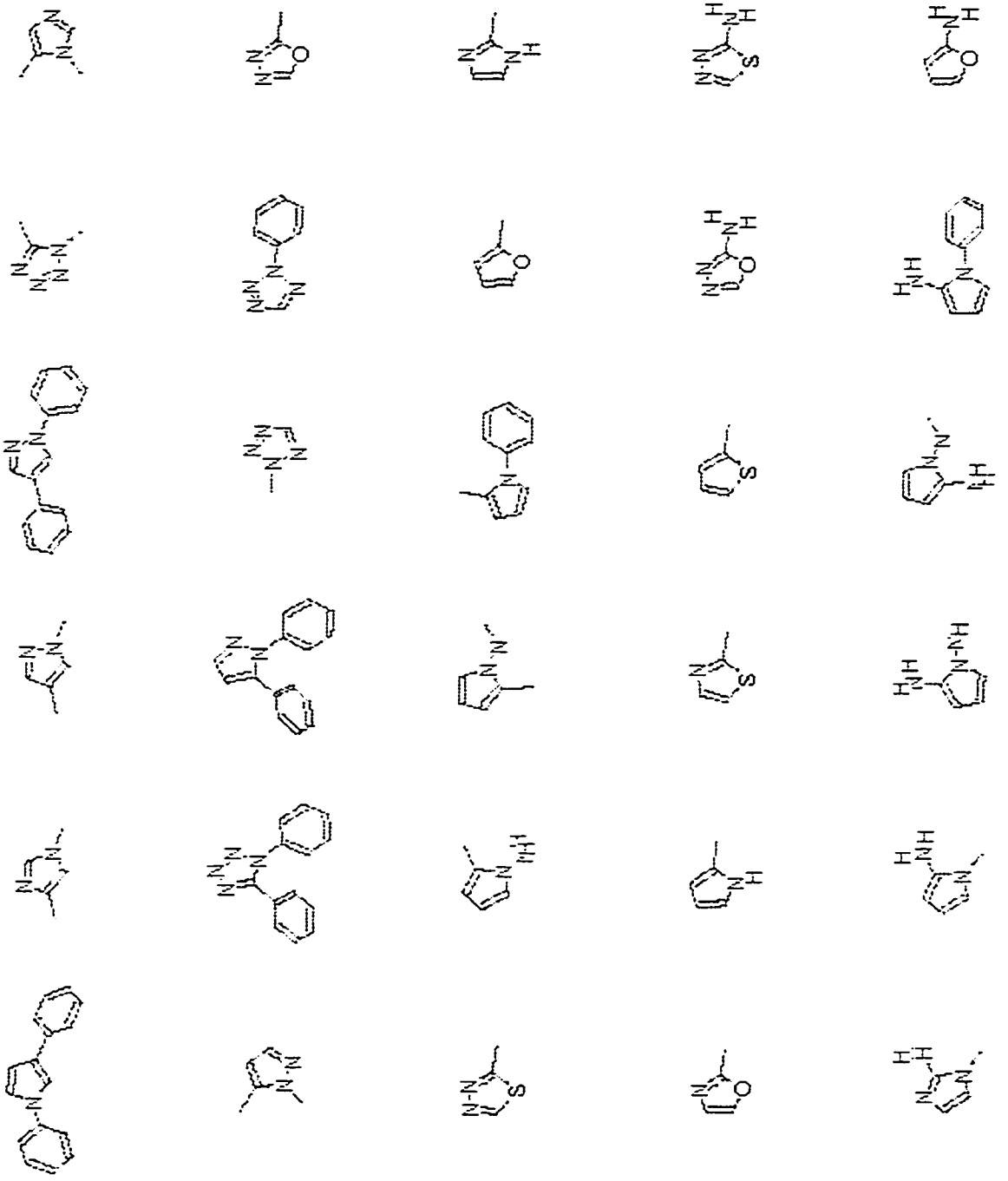
C.6 Fused 6-6 Aromatic Rings

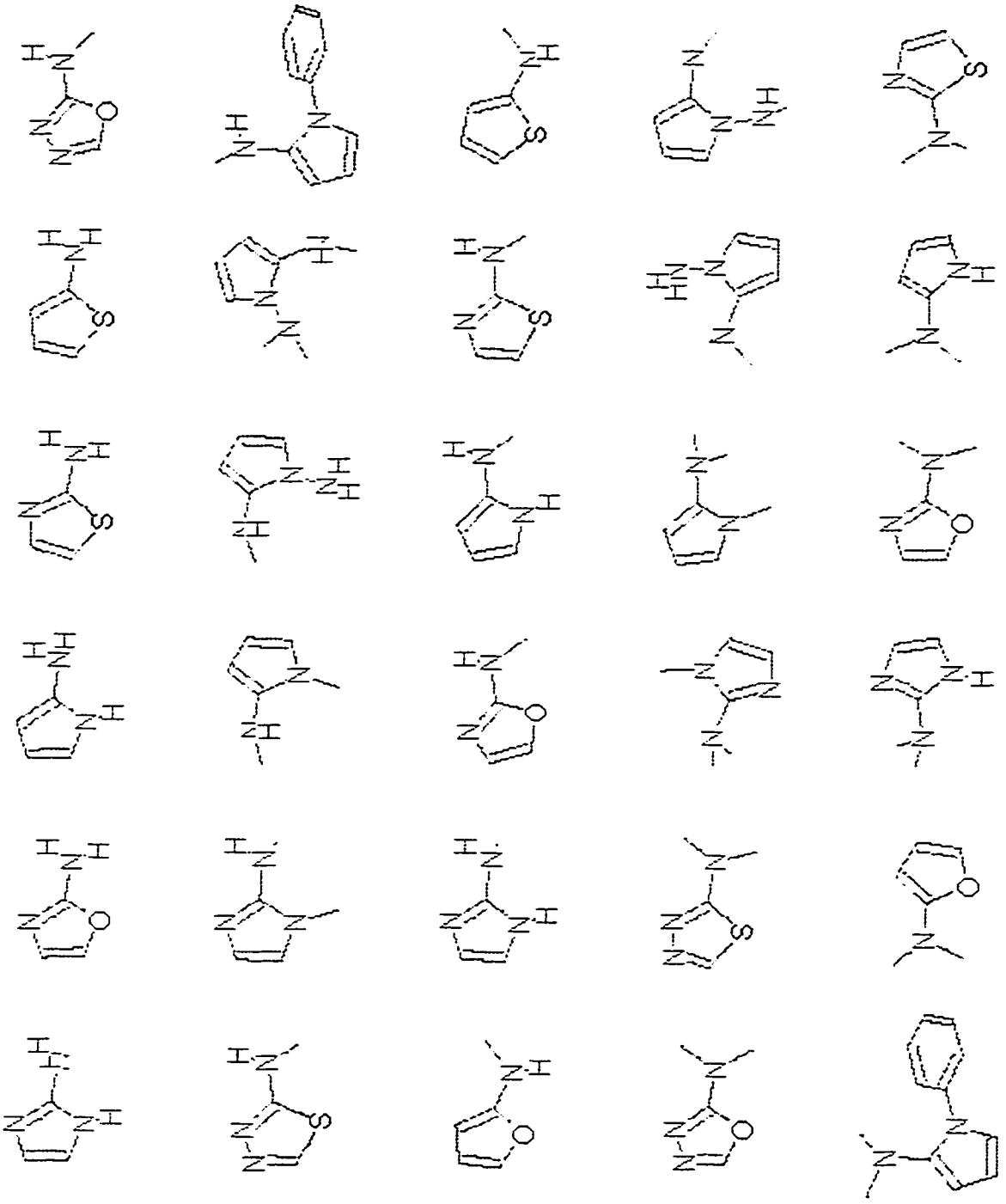


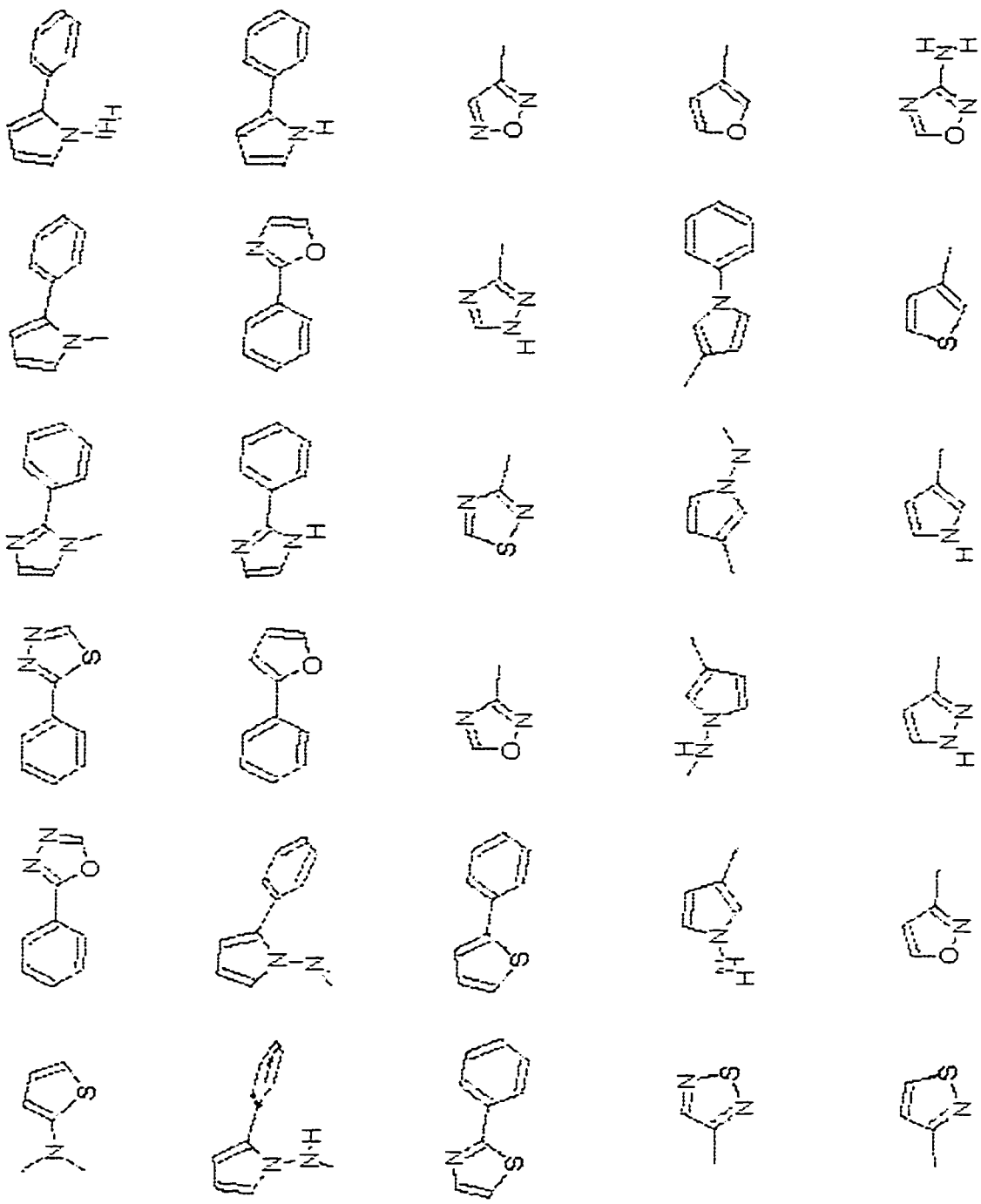


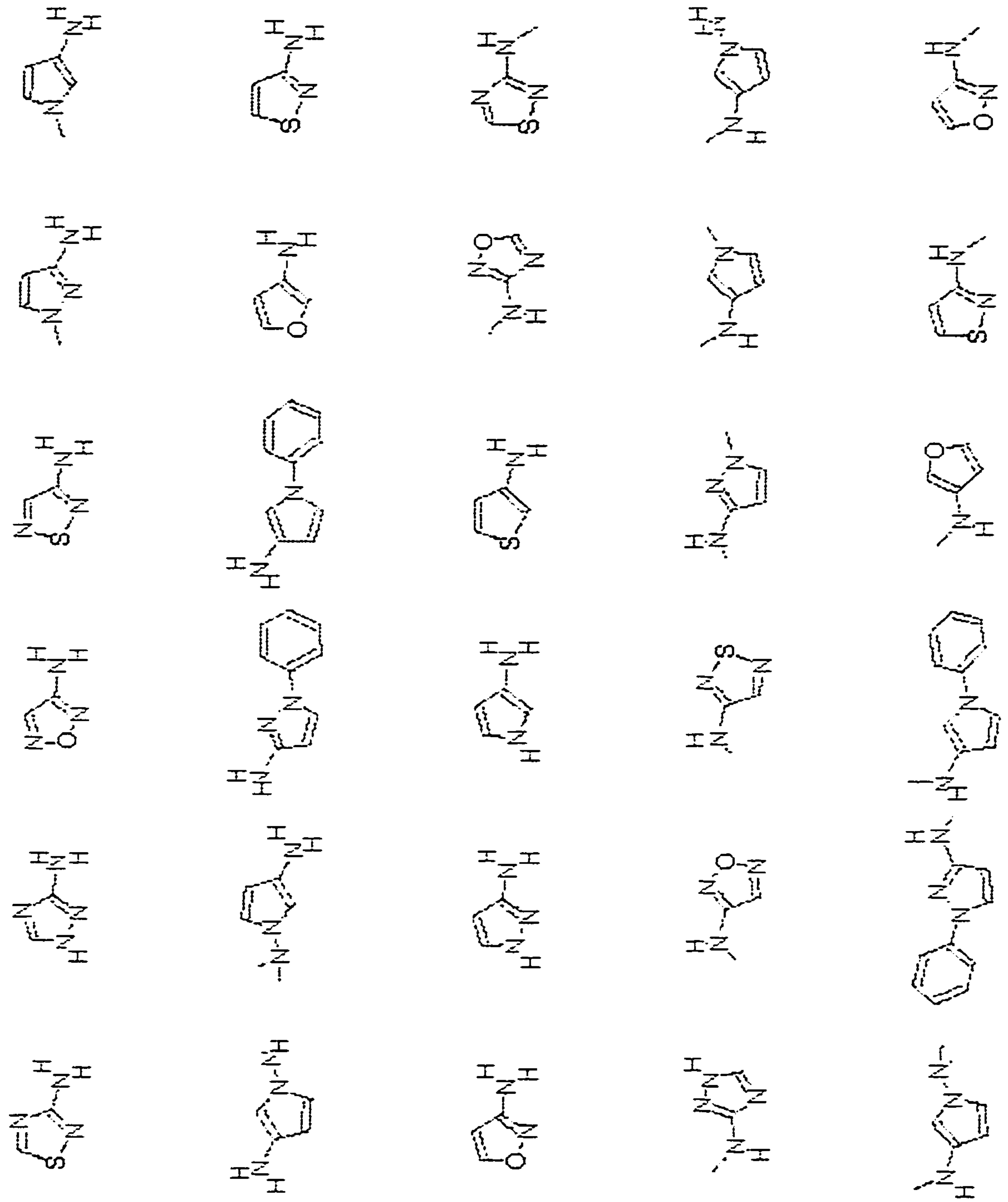
C.7 5-Membered Heterocyles

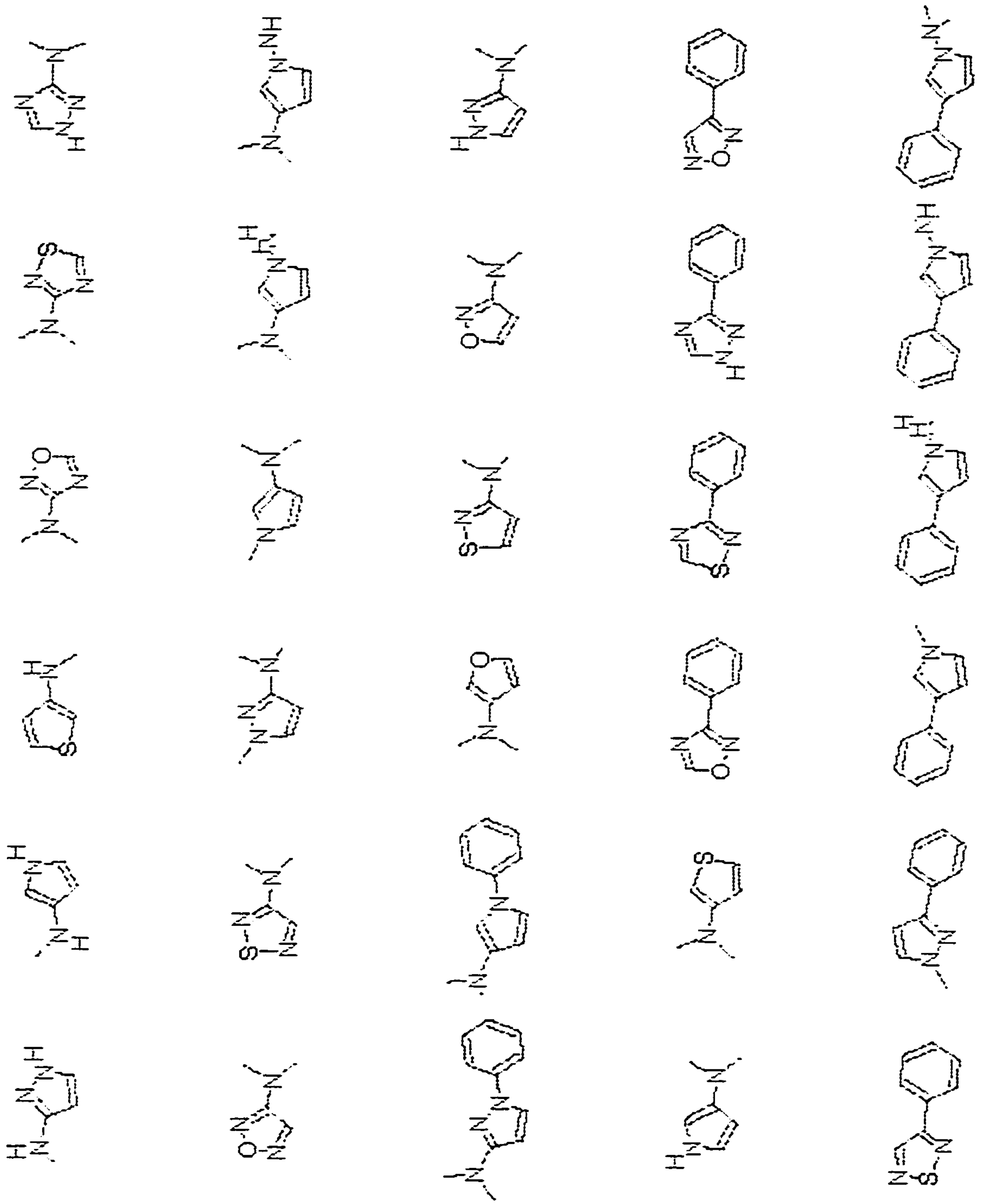


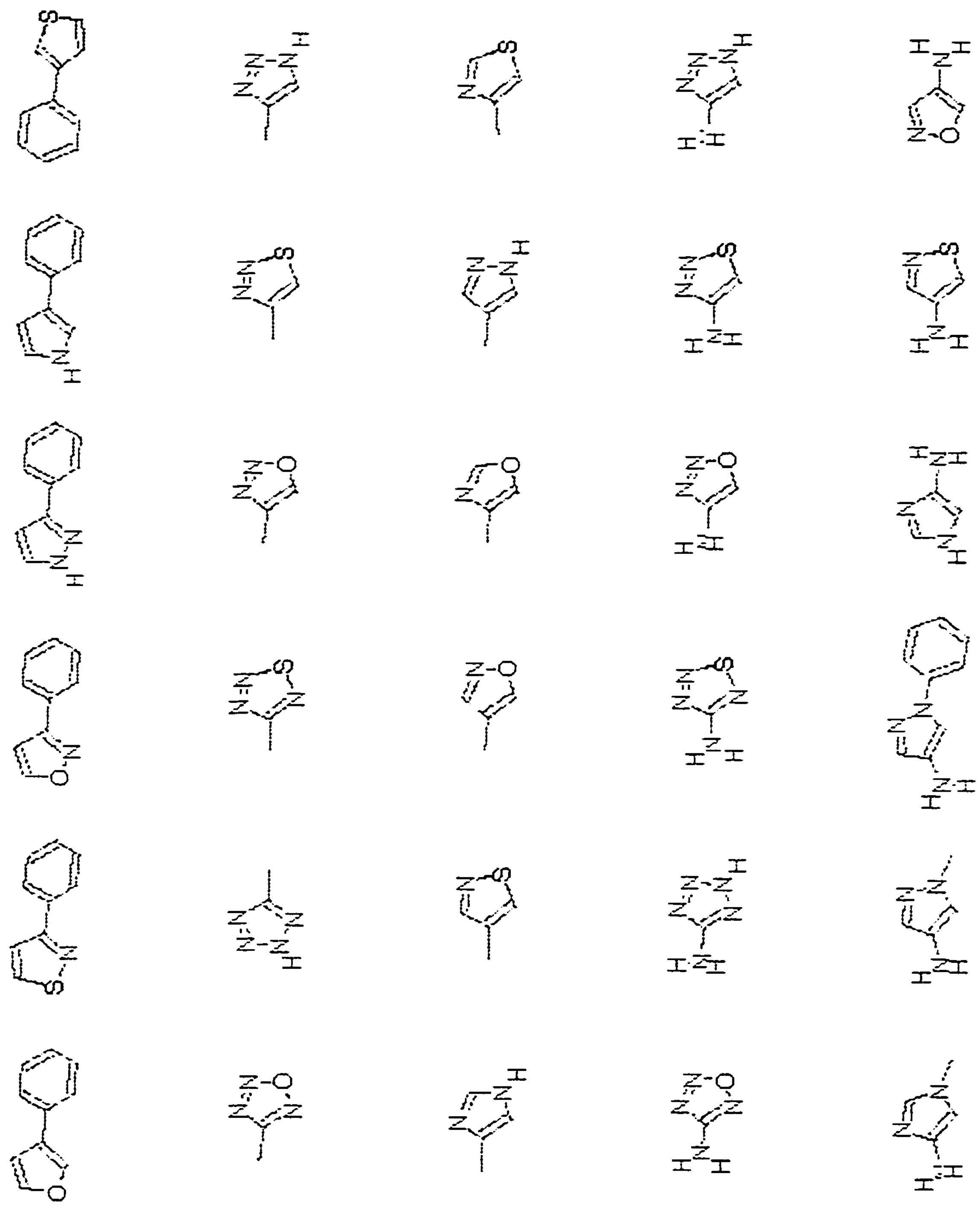


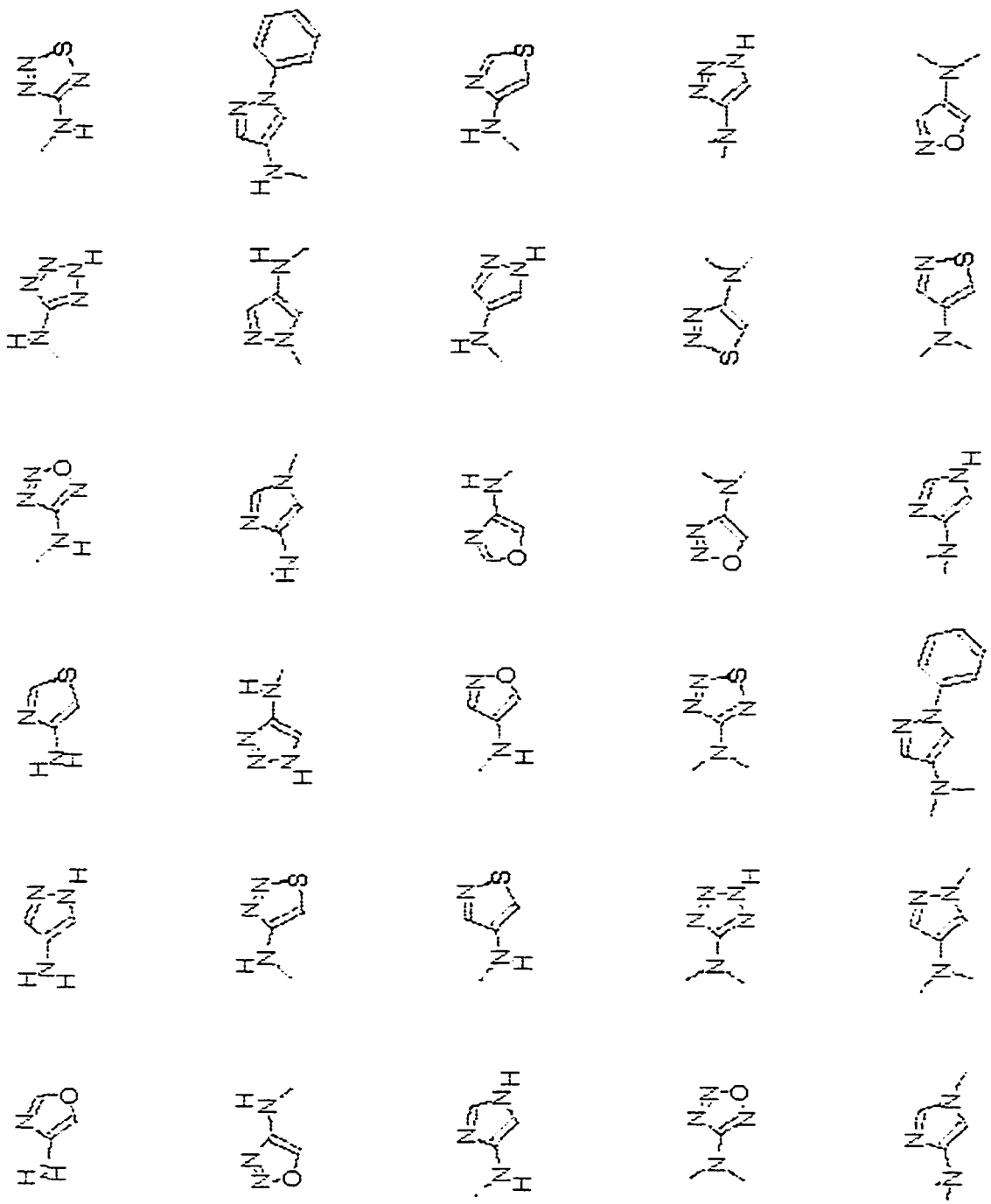


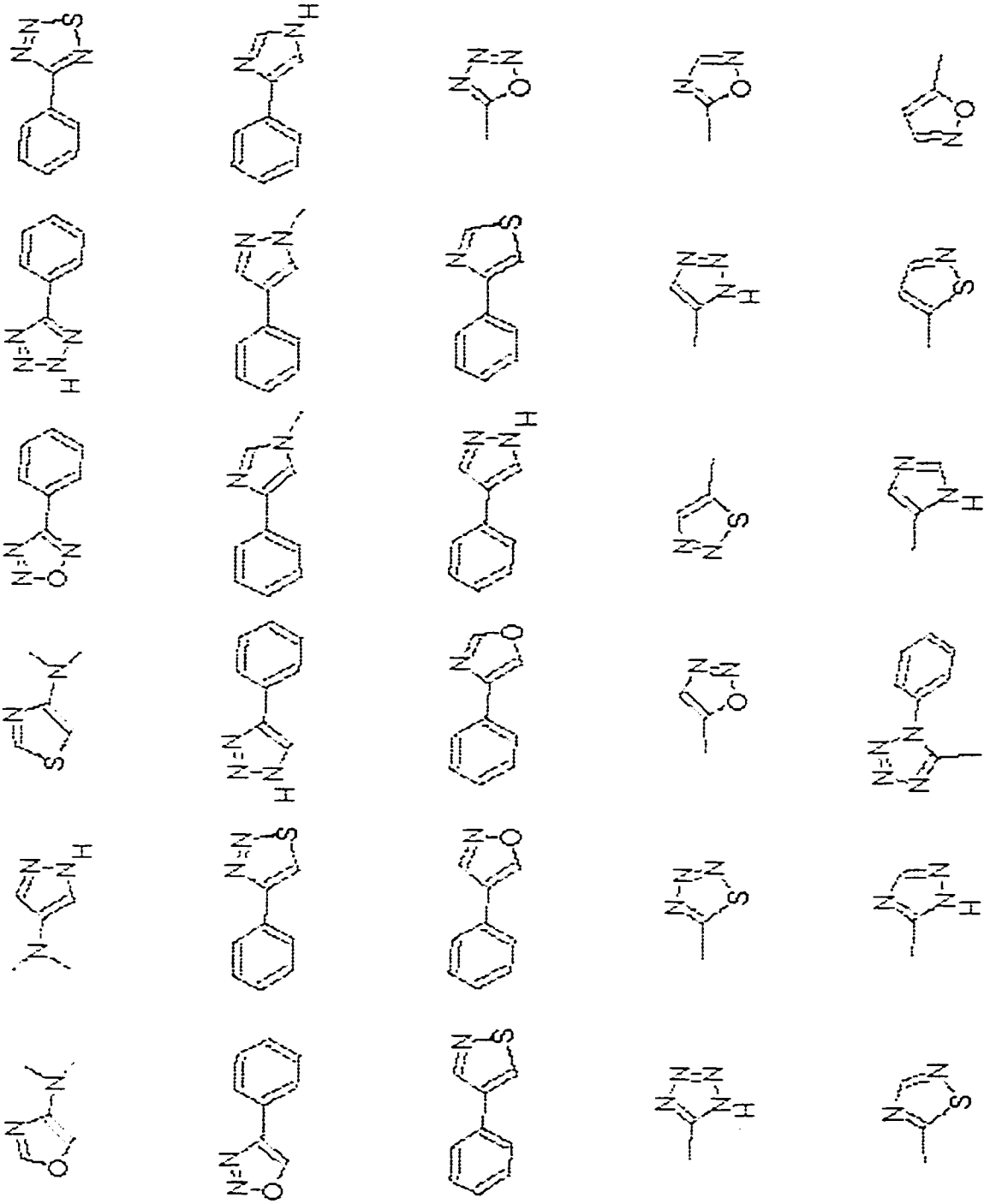


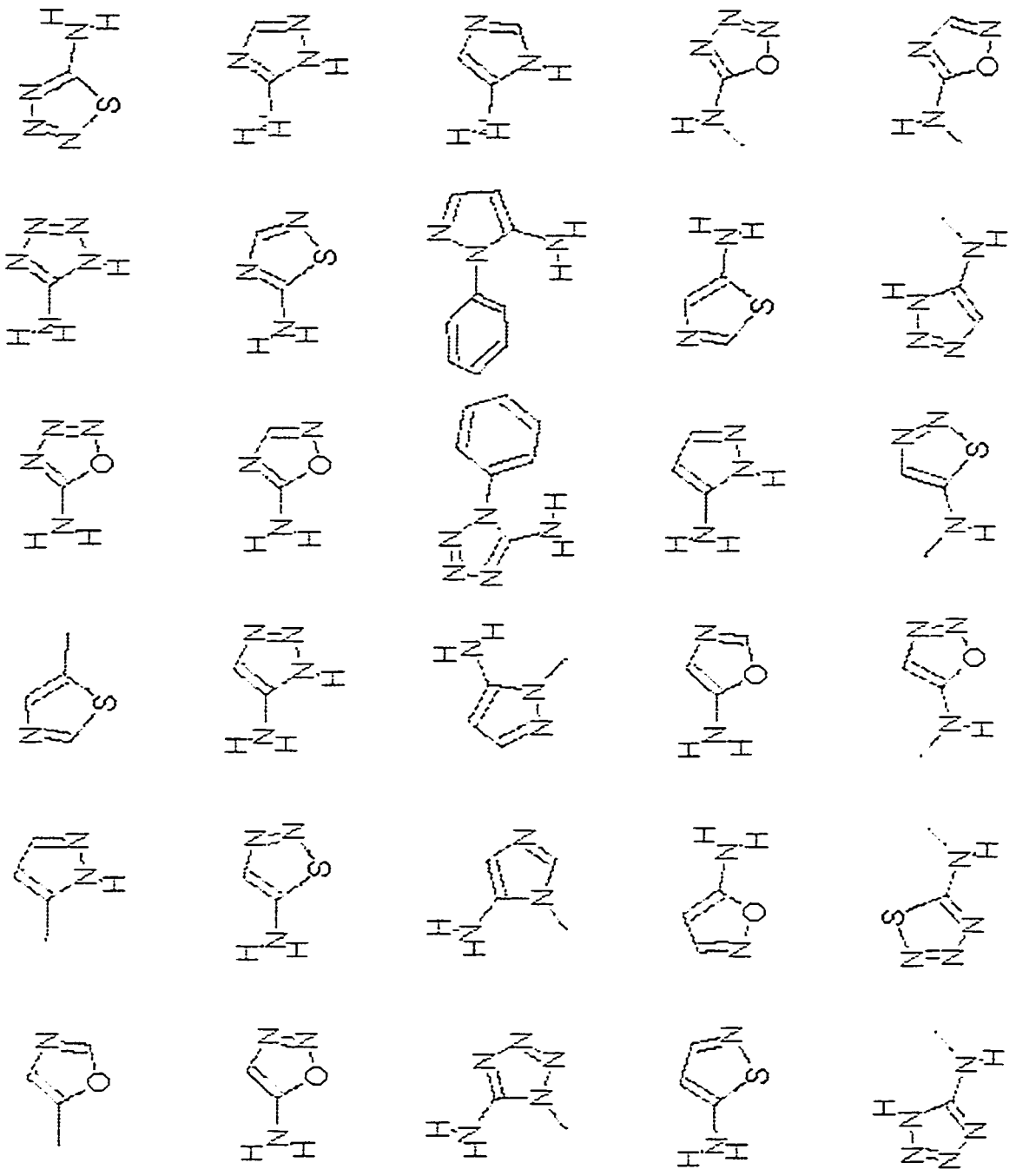


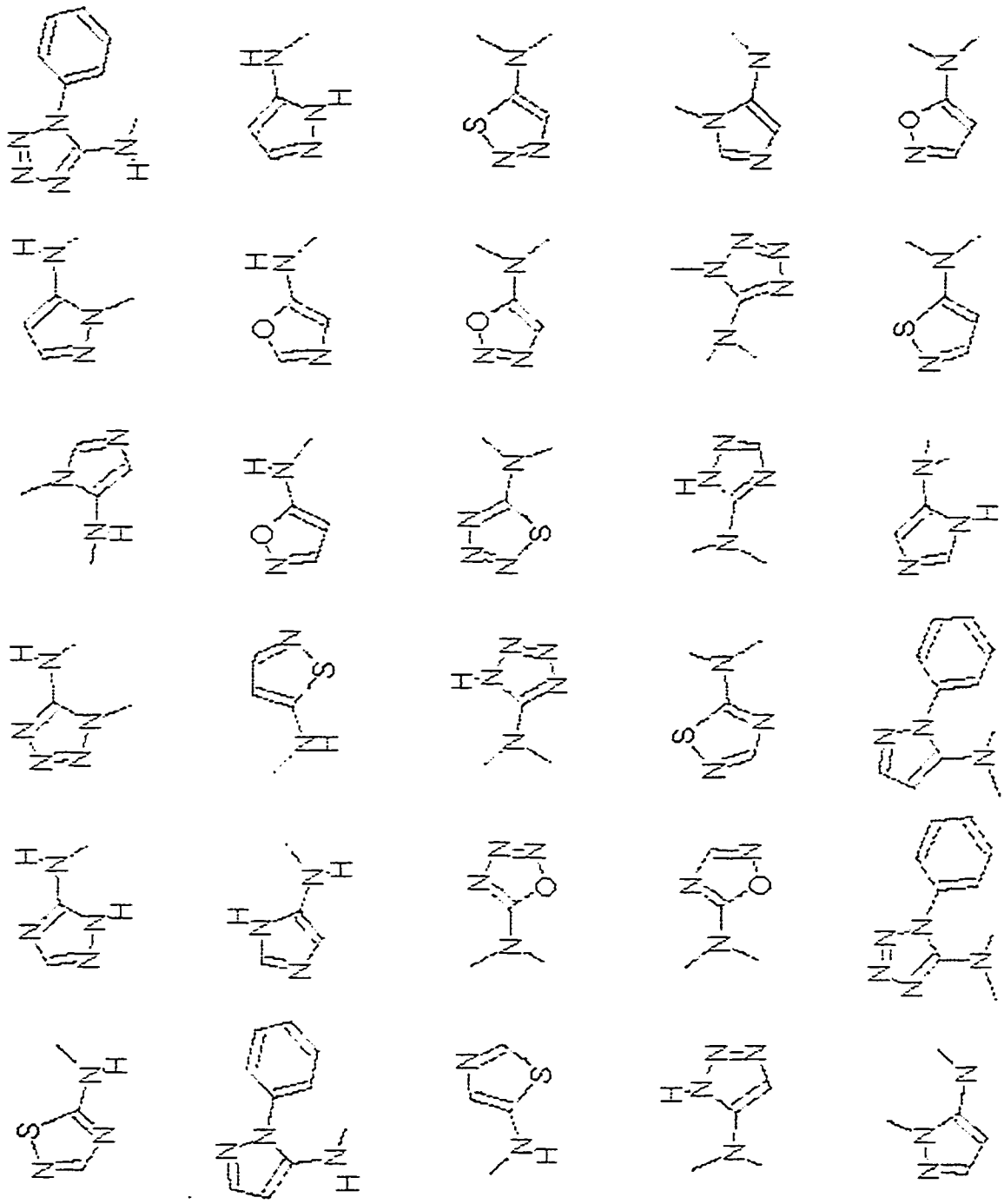


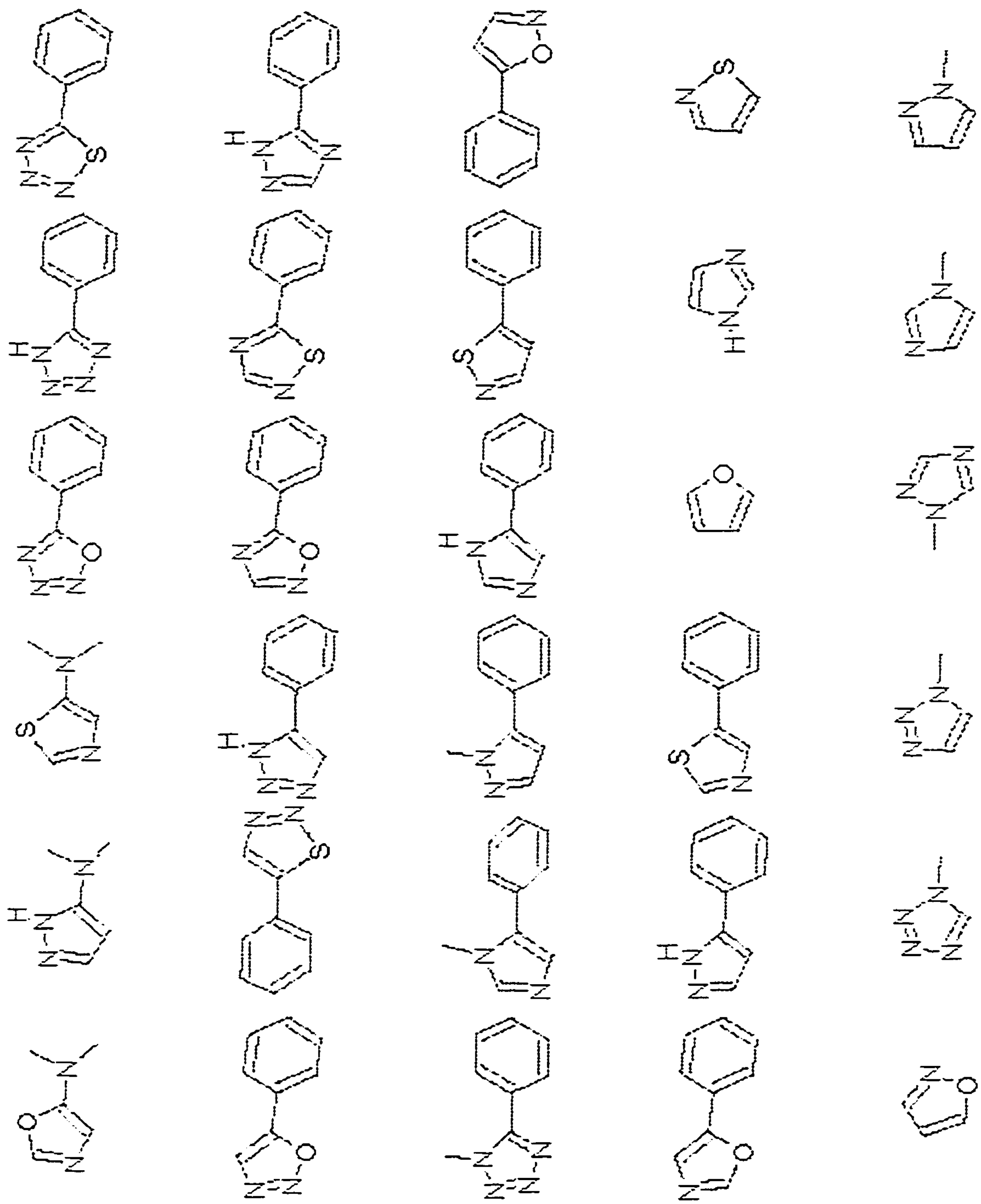


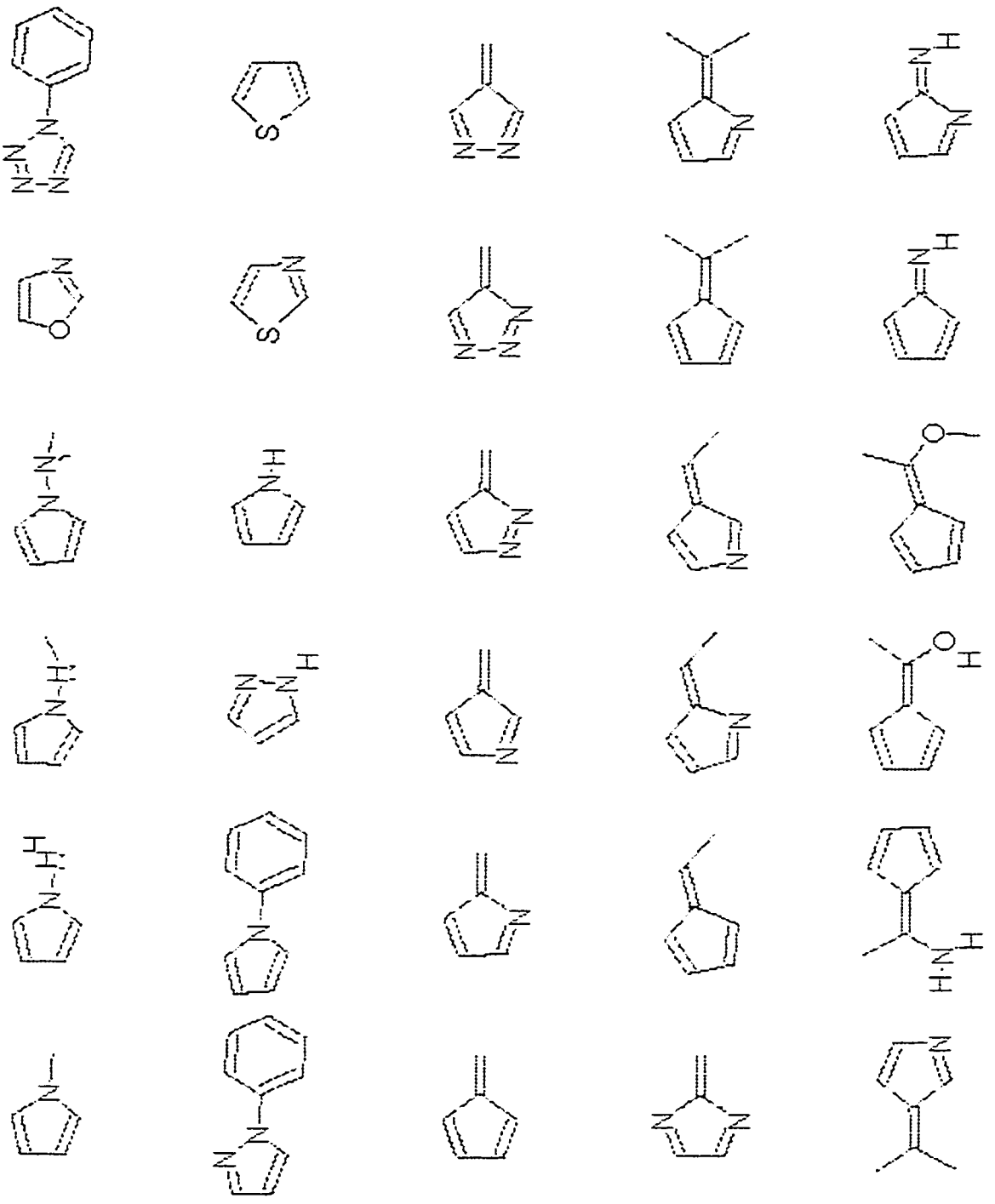


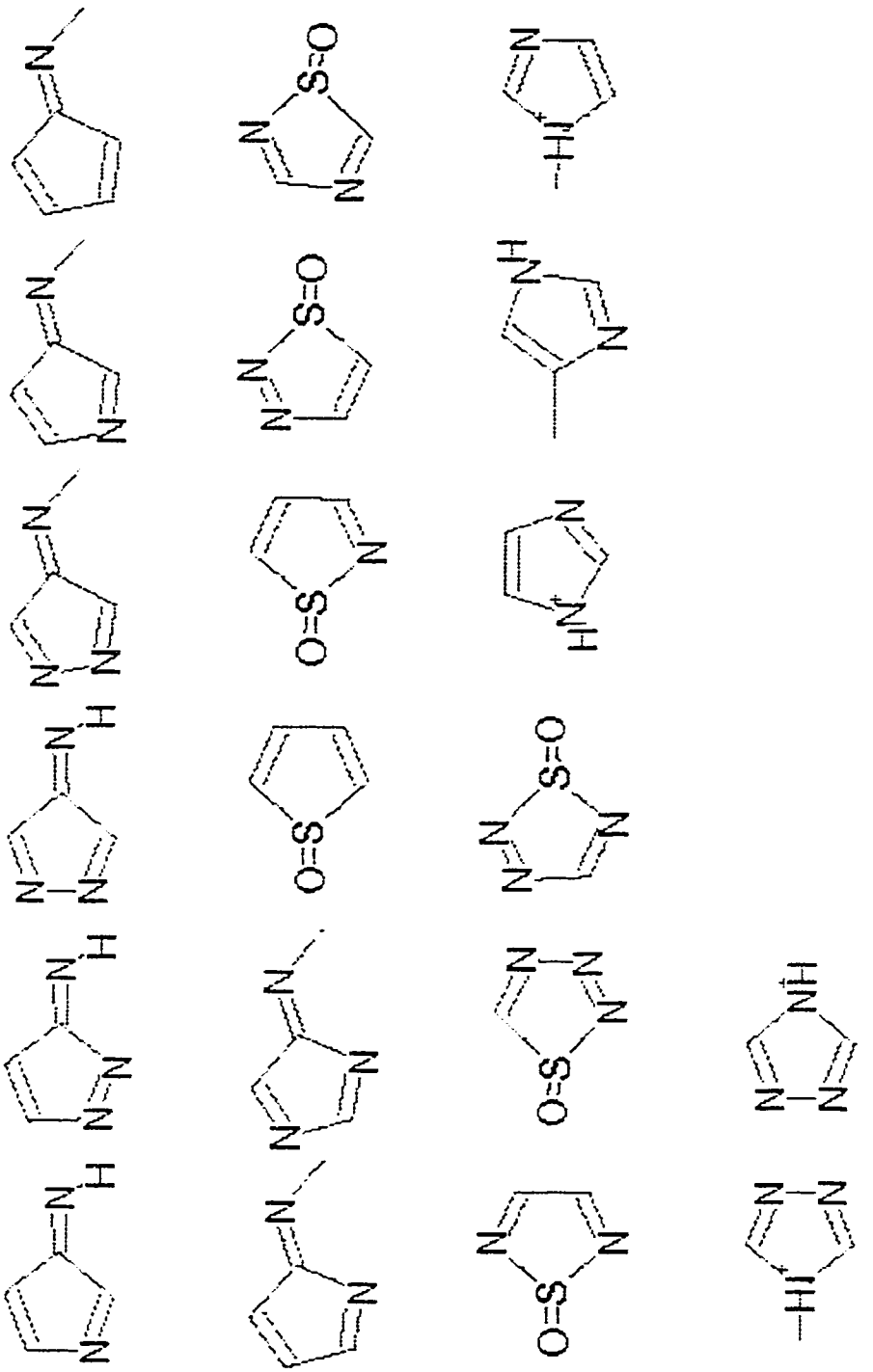




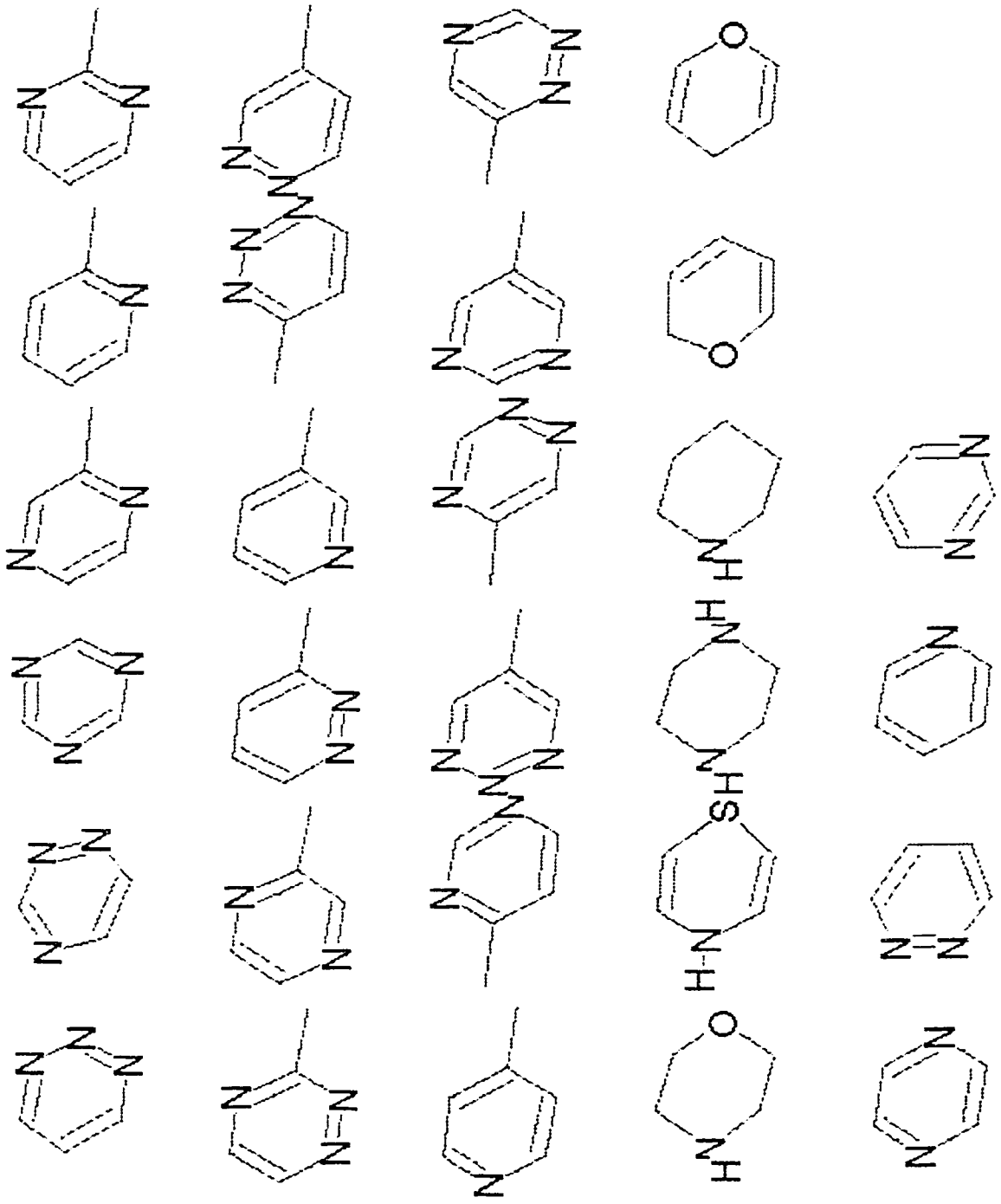




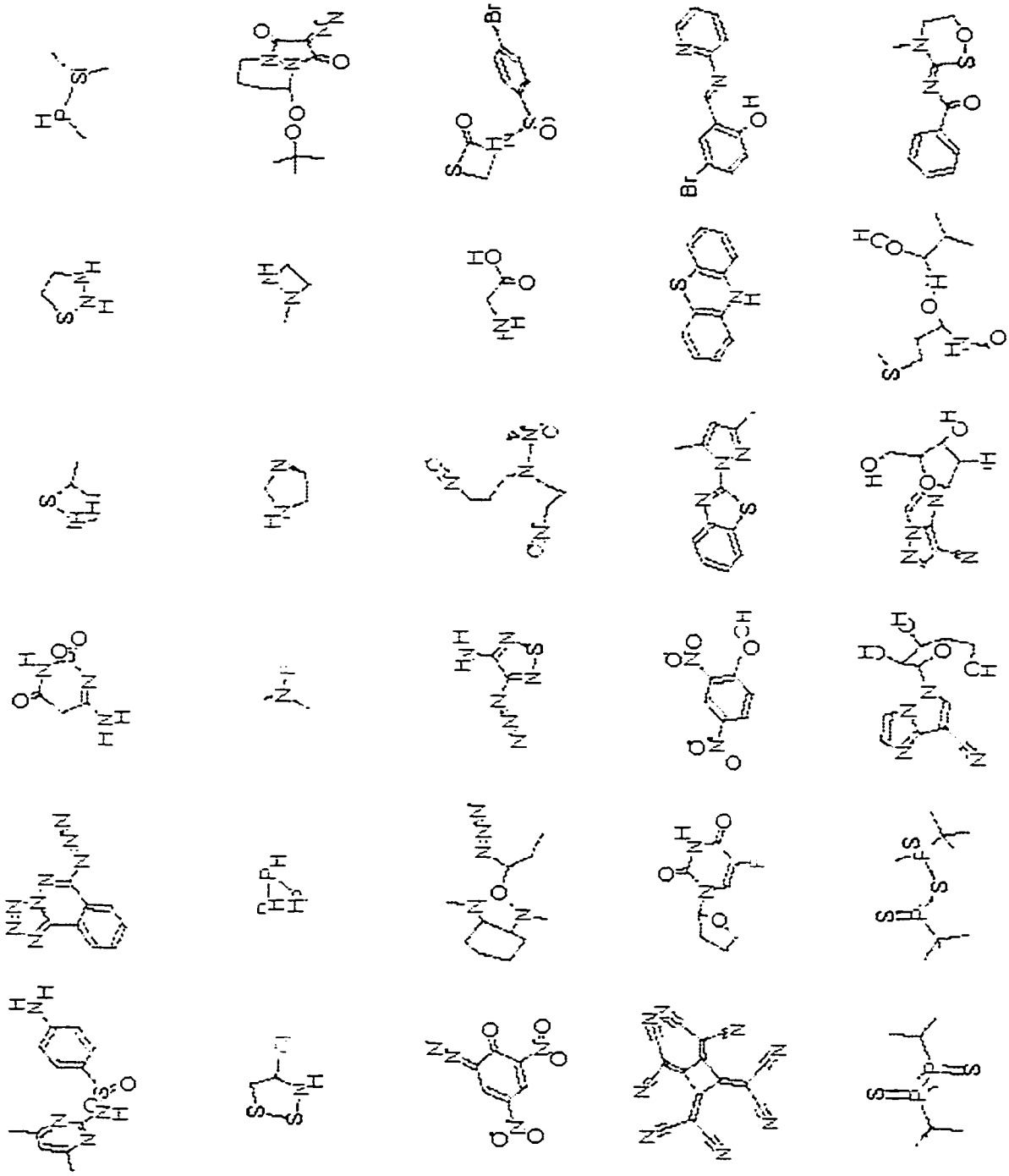


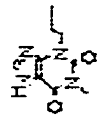
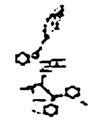
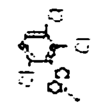
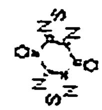
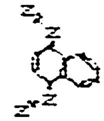


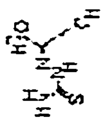
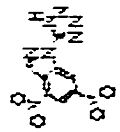
C.8 6-Membered Heterocyles

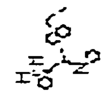
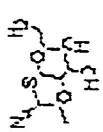
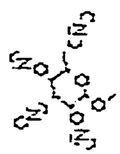


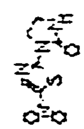
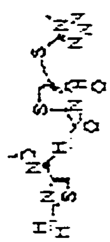
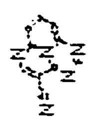
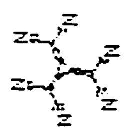
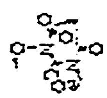
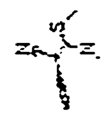
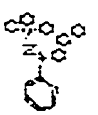
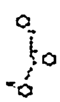
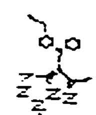
C.9 MMFF94 Validation Suite of Neutral Molecules

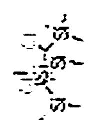


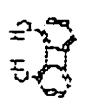
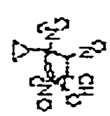
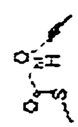
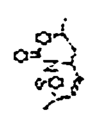






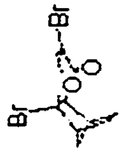
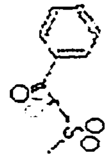
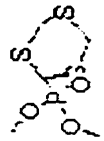
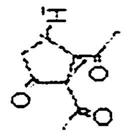
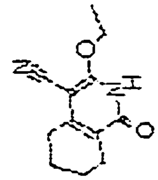



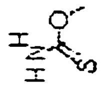
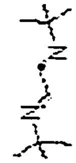
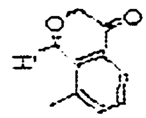
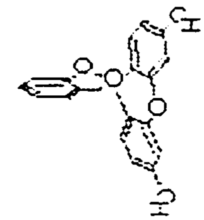
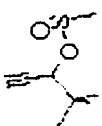
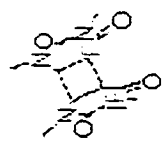
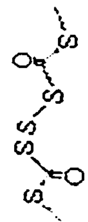
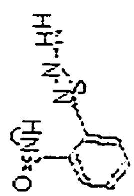
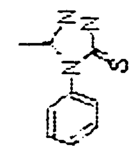
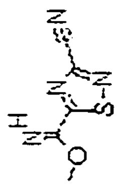




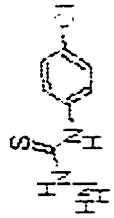
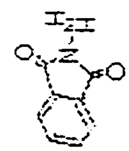
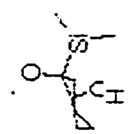











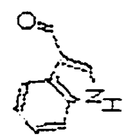
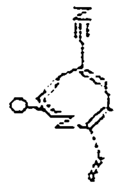
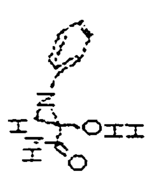
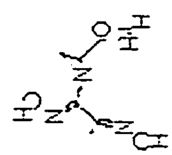


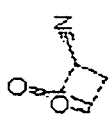





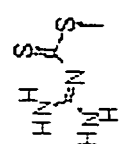
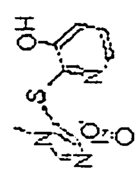






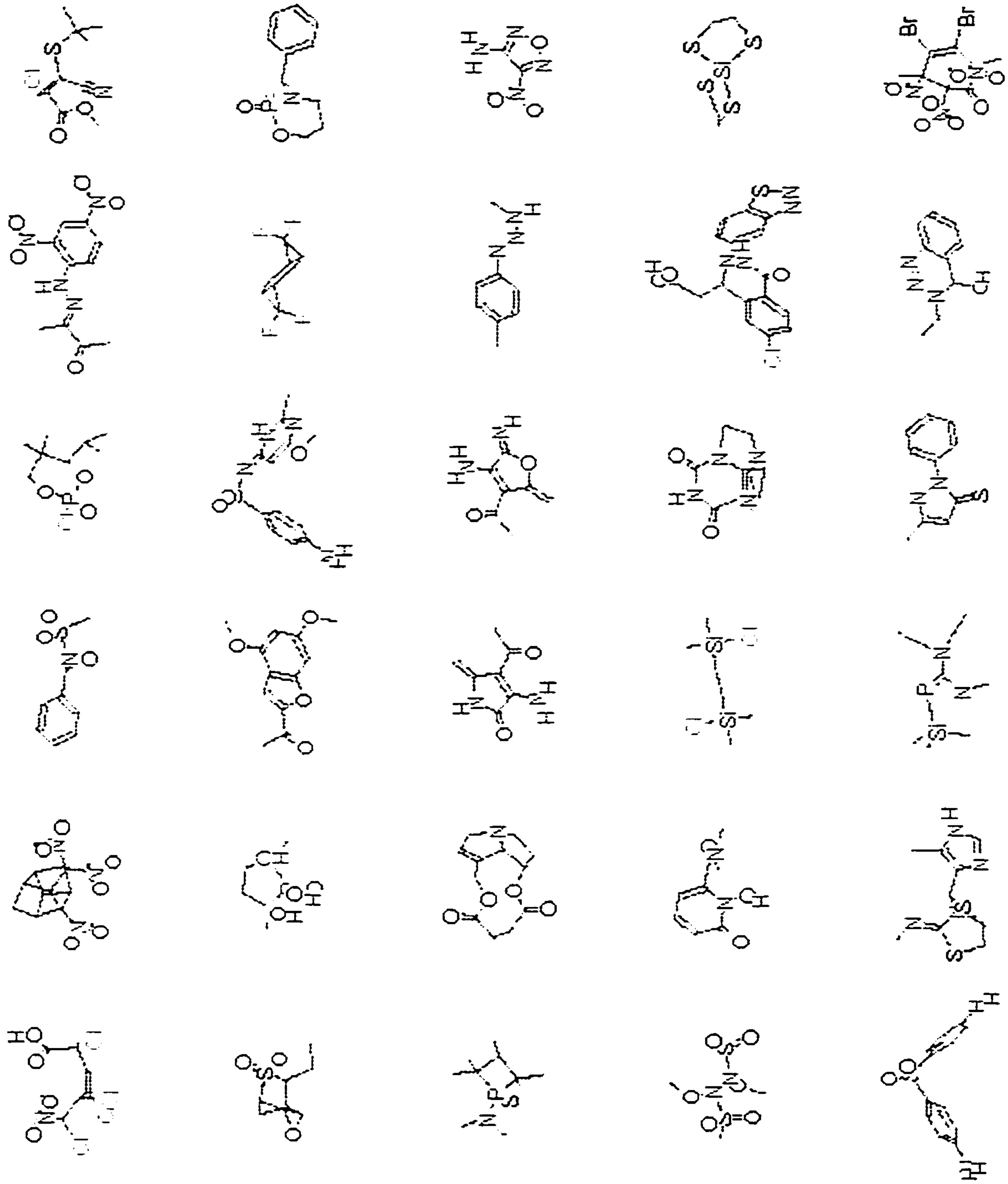











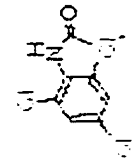
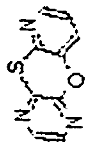




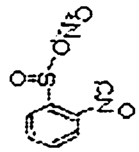
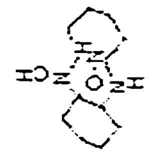
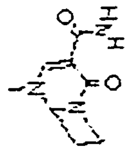
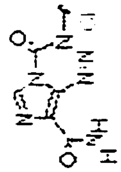





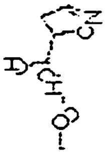
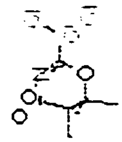
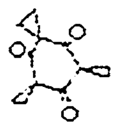


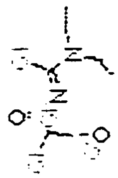
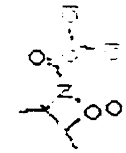
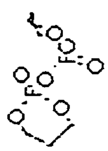
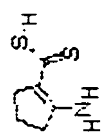
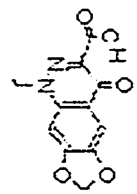
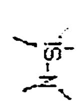
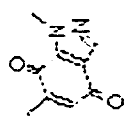
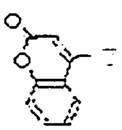
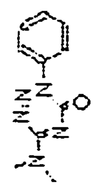
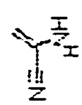
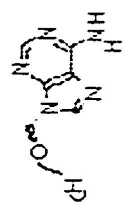
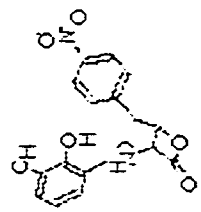
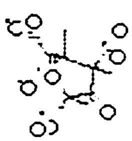
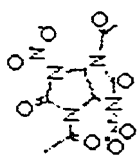
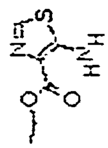
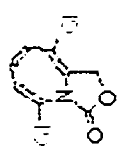
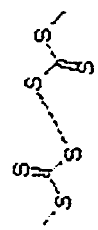



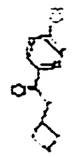
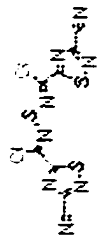


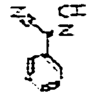
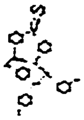



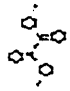
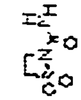
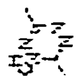





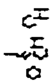
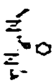




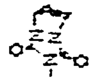
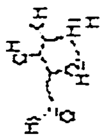



















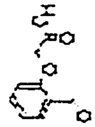
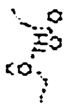



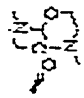
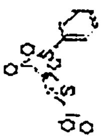
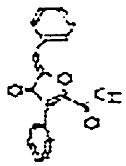
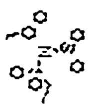
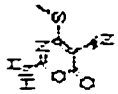
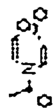



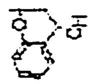
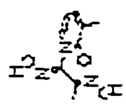
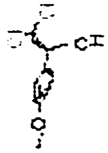
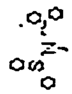




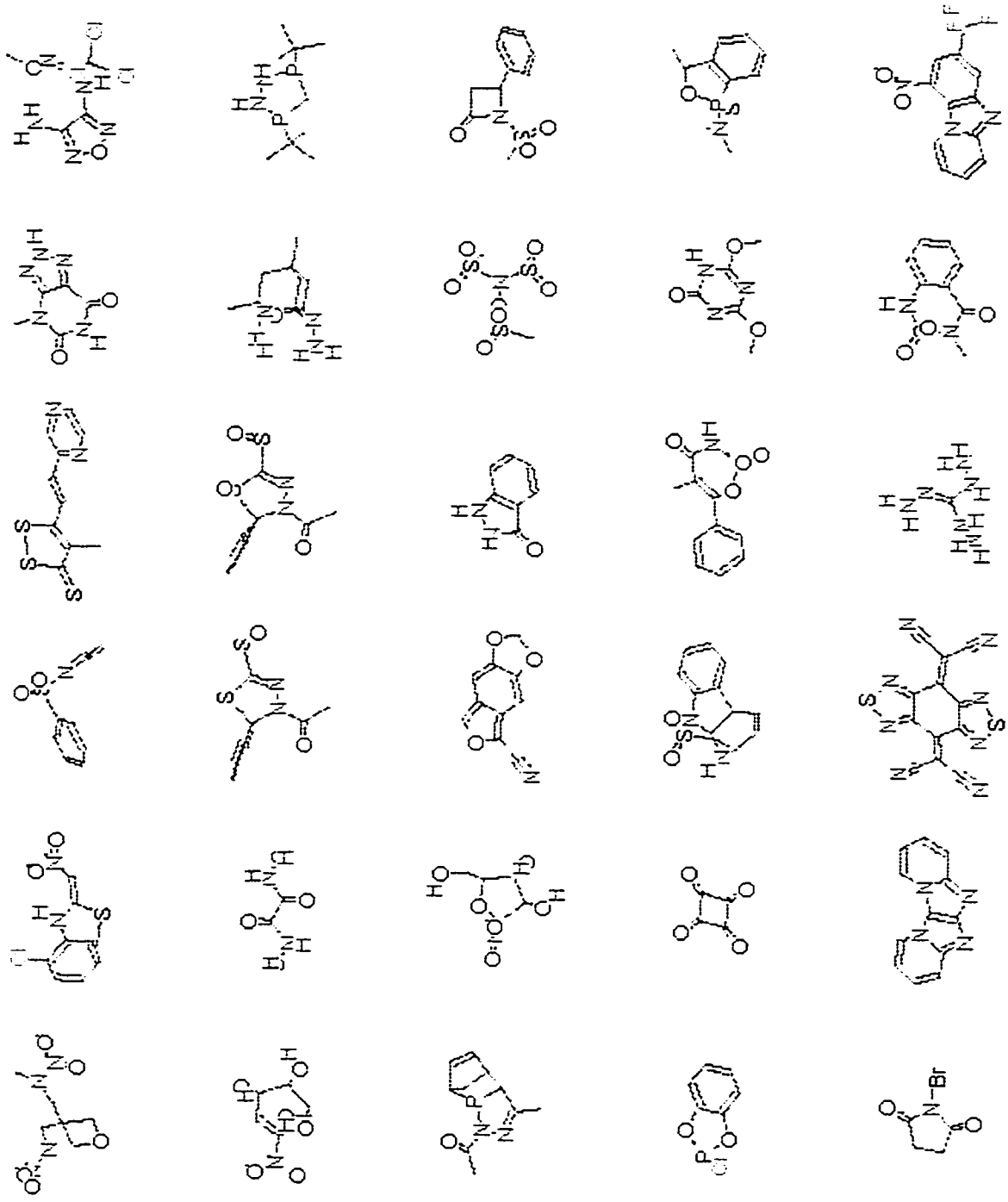



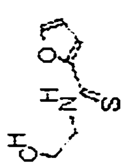
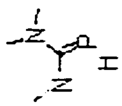
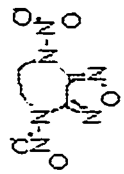
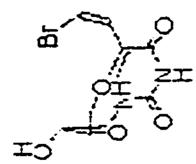
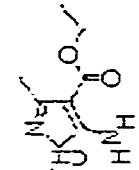



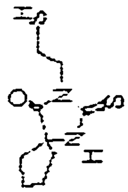
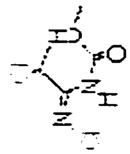



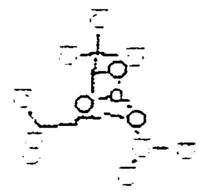
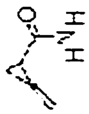
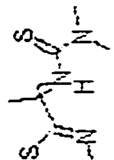
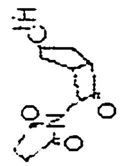
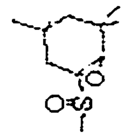
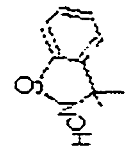
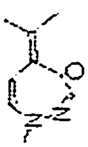
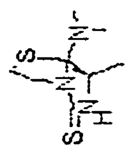
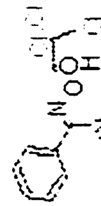
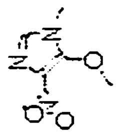
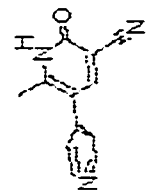
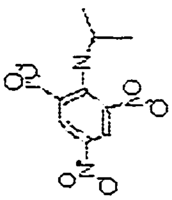
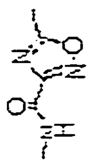
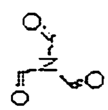
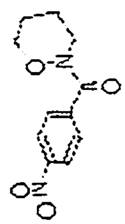







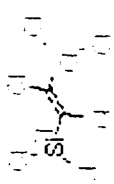
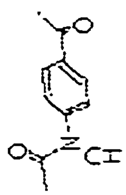
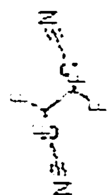
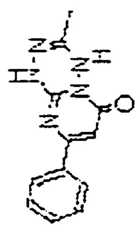
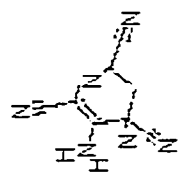





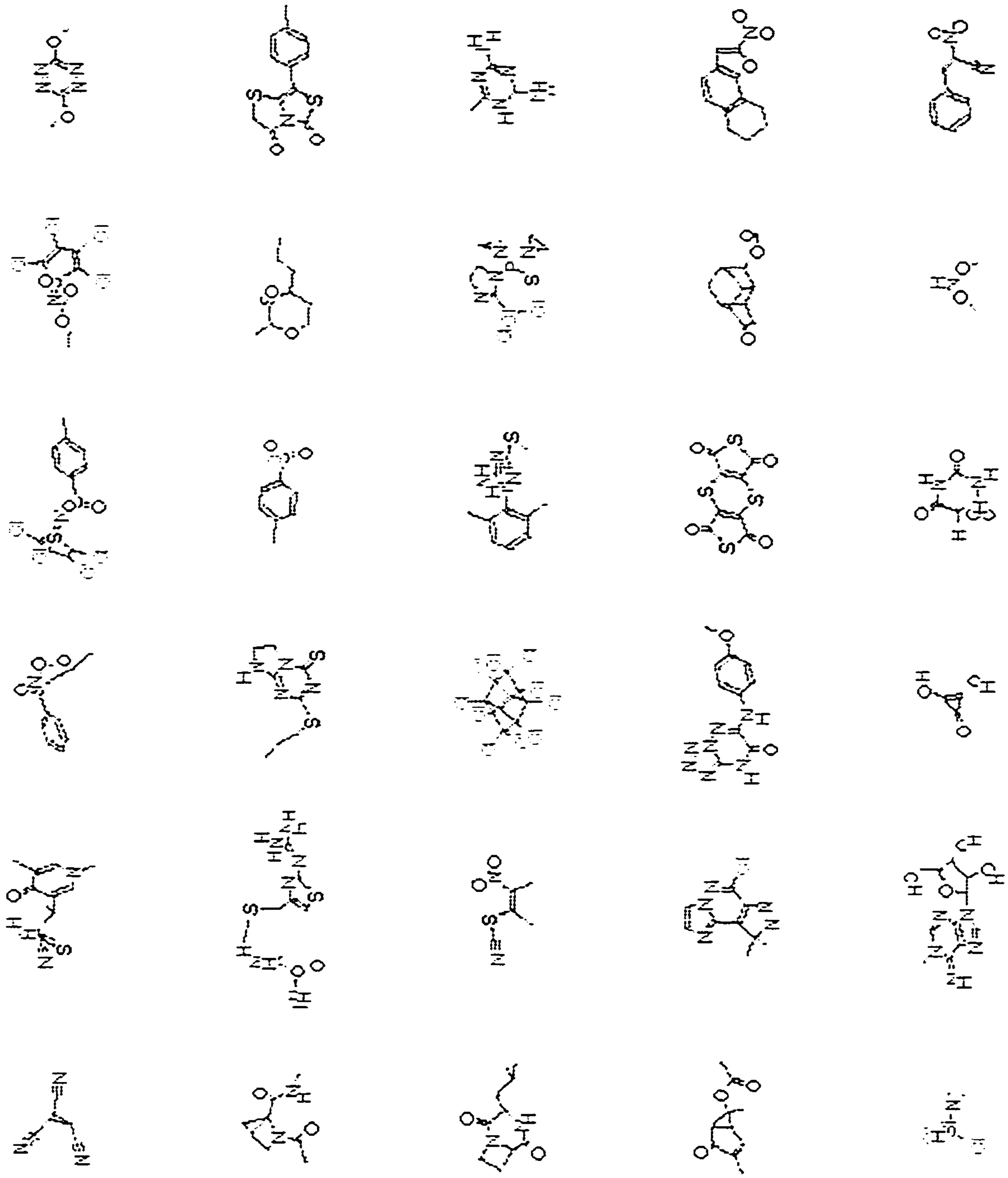


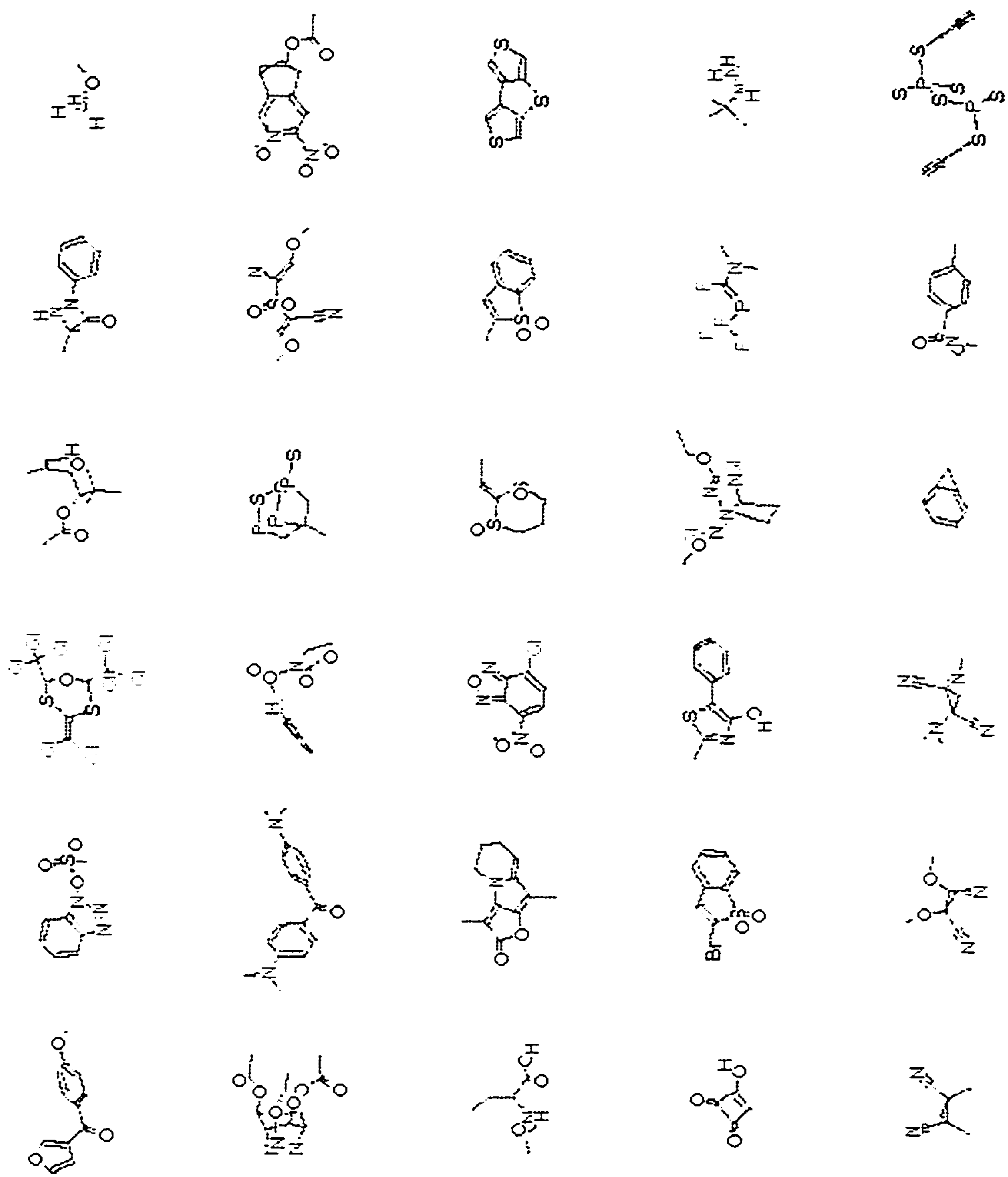






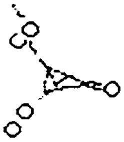
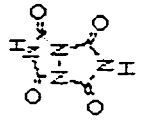
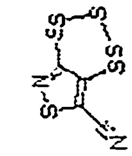



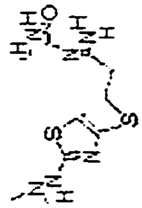
















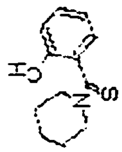
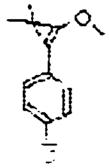
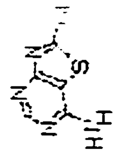






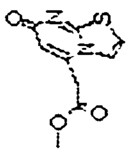


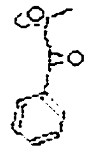
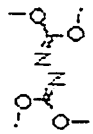
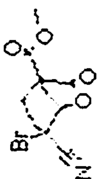
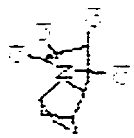


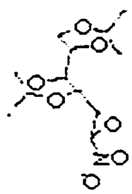
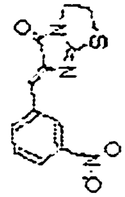
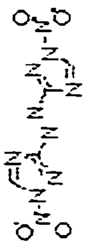
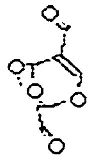
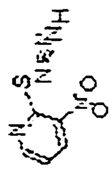
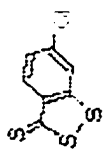
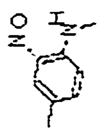




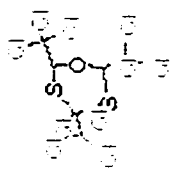
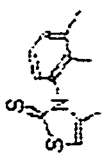


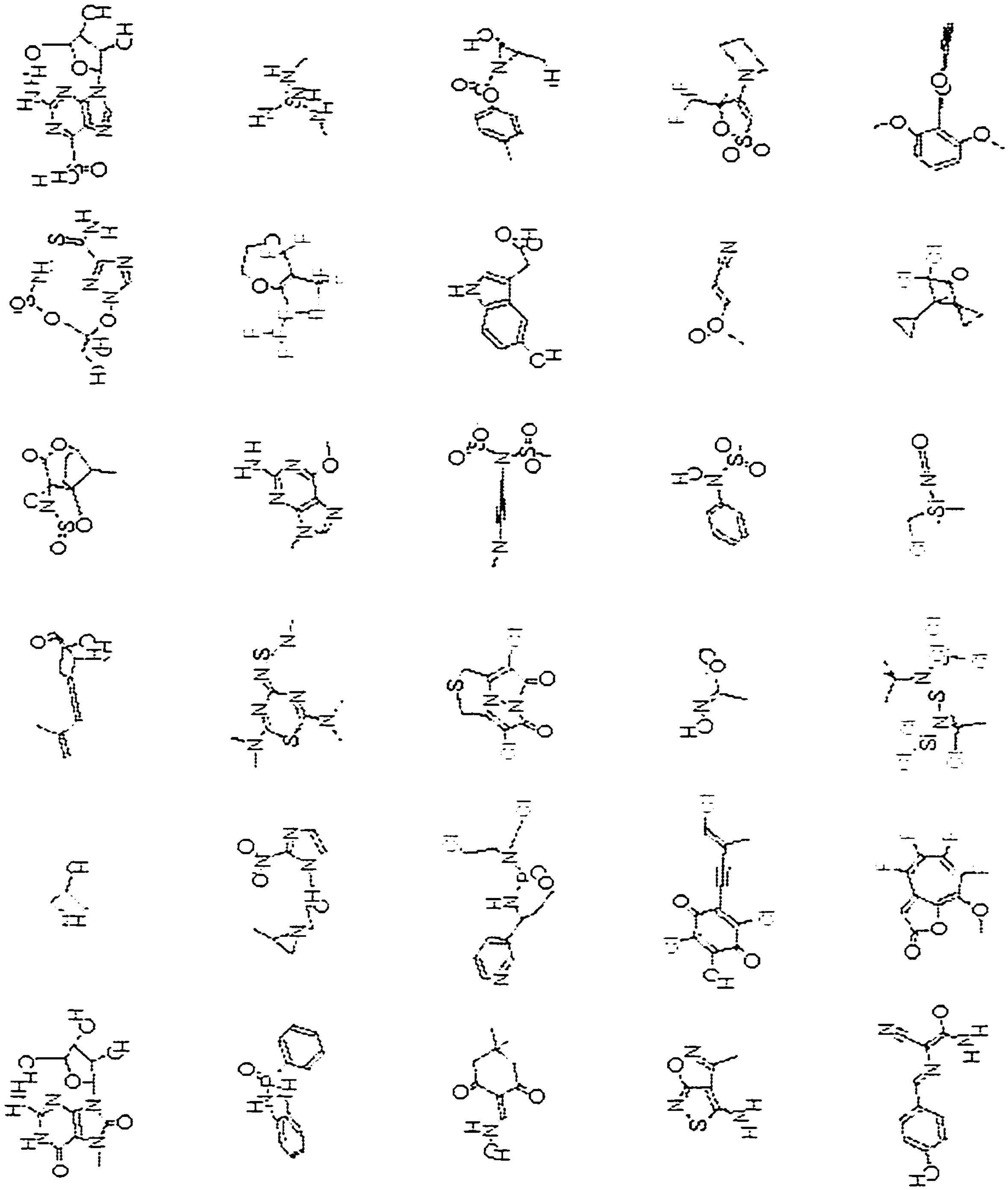





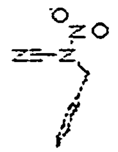
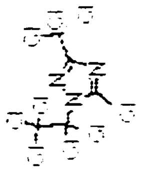
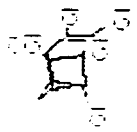
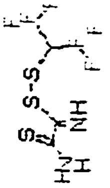
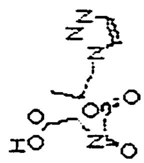


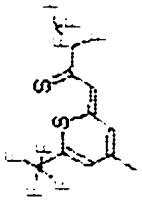





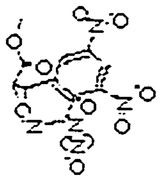
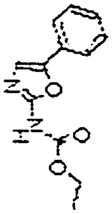
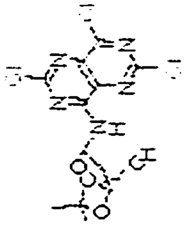
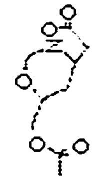
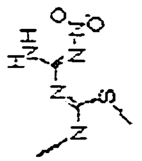
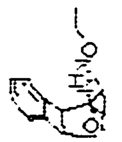
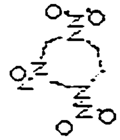
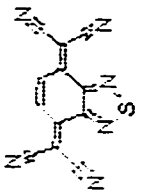
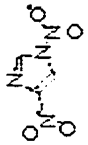
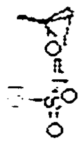








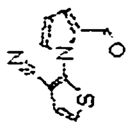
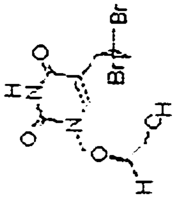
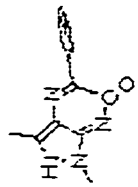
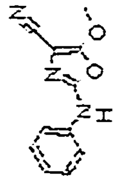
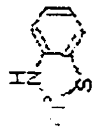
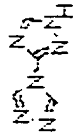
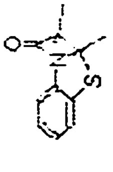




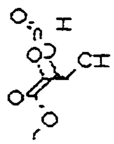
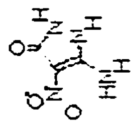



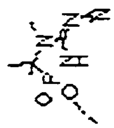
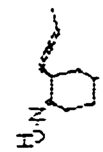






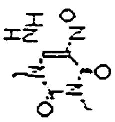
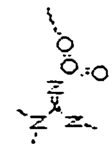
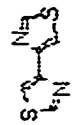
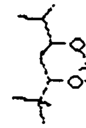
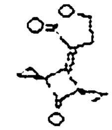


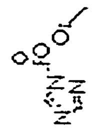
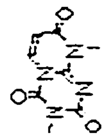
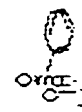











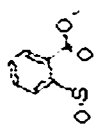
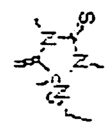
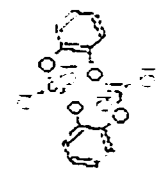








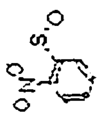
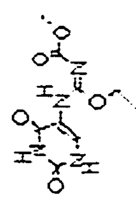
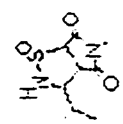
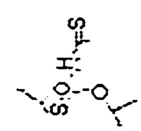




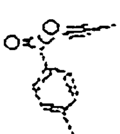
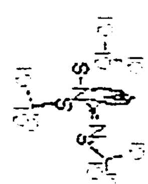
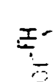
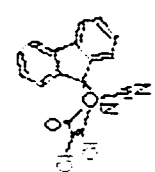
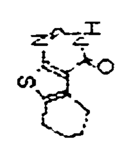
    

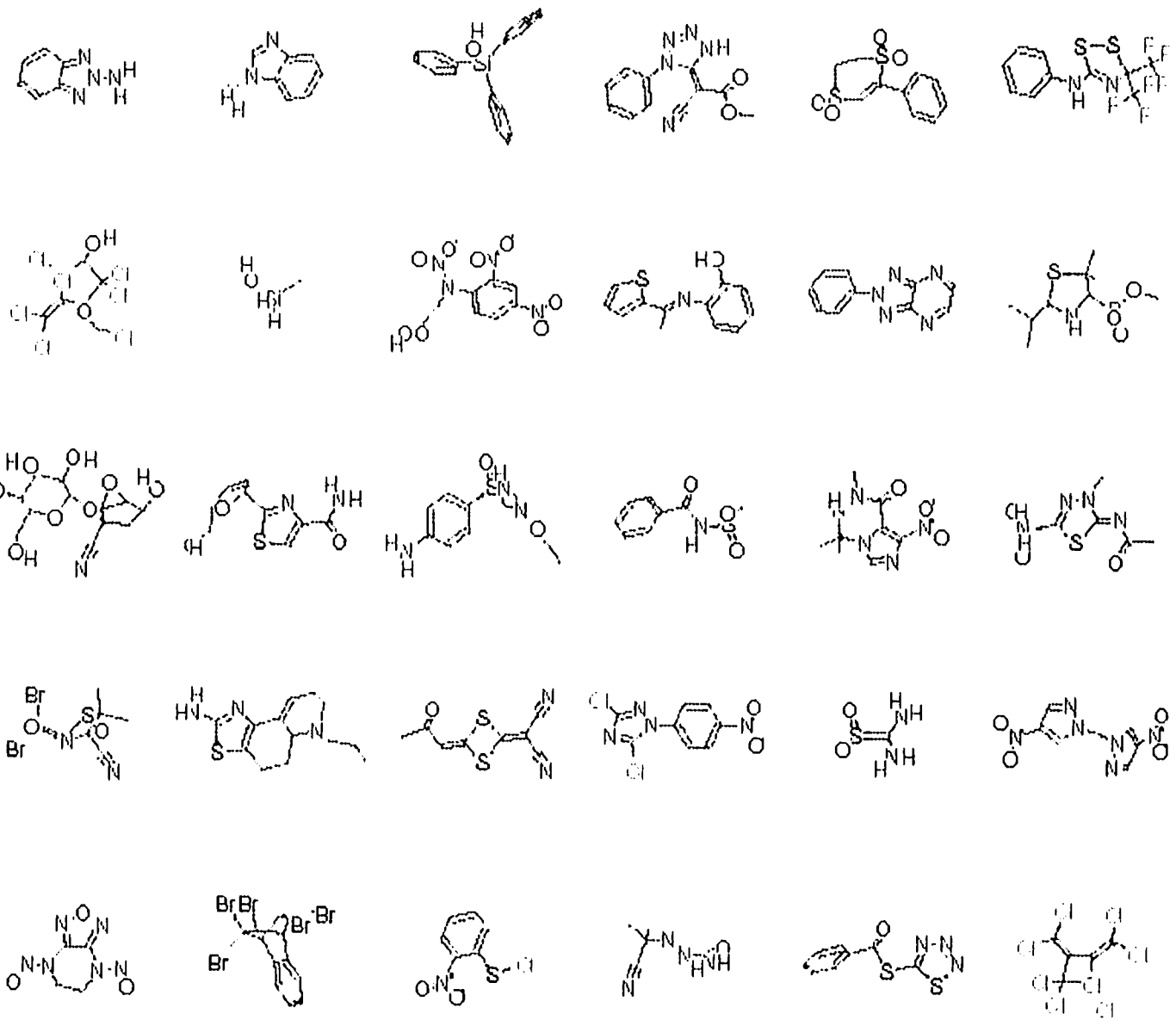
    

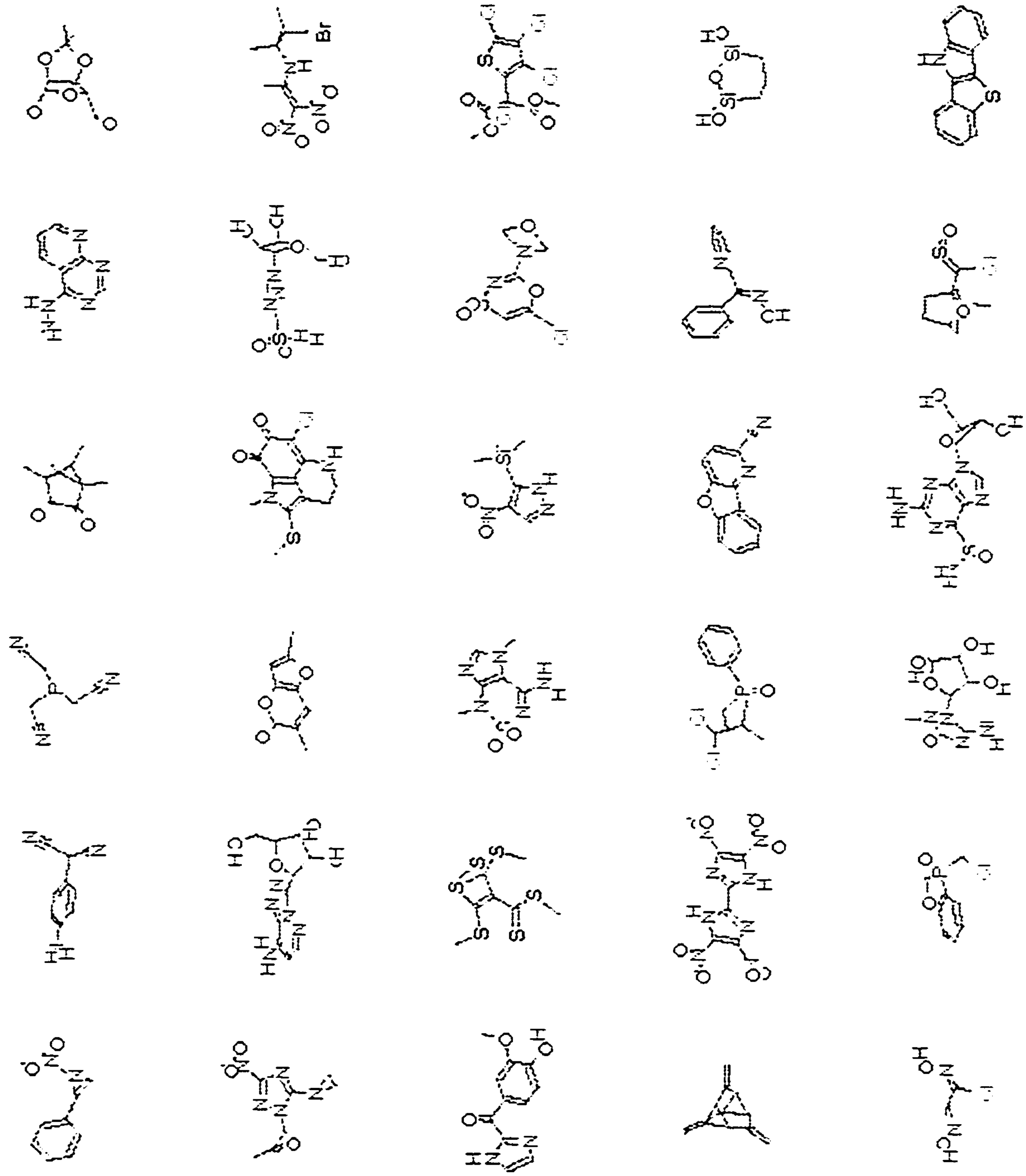
    

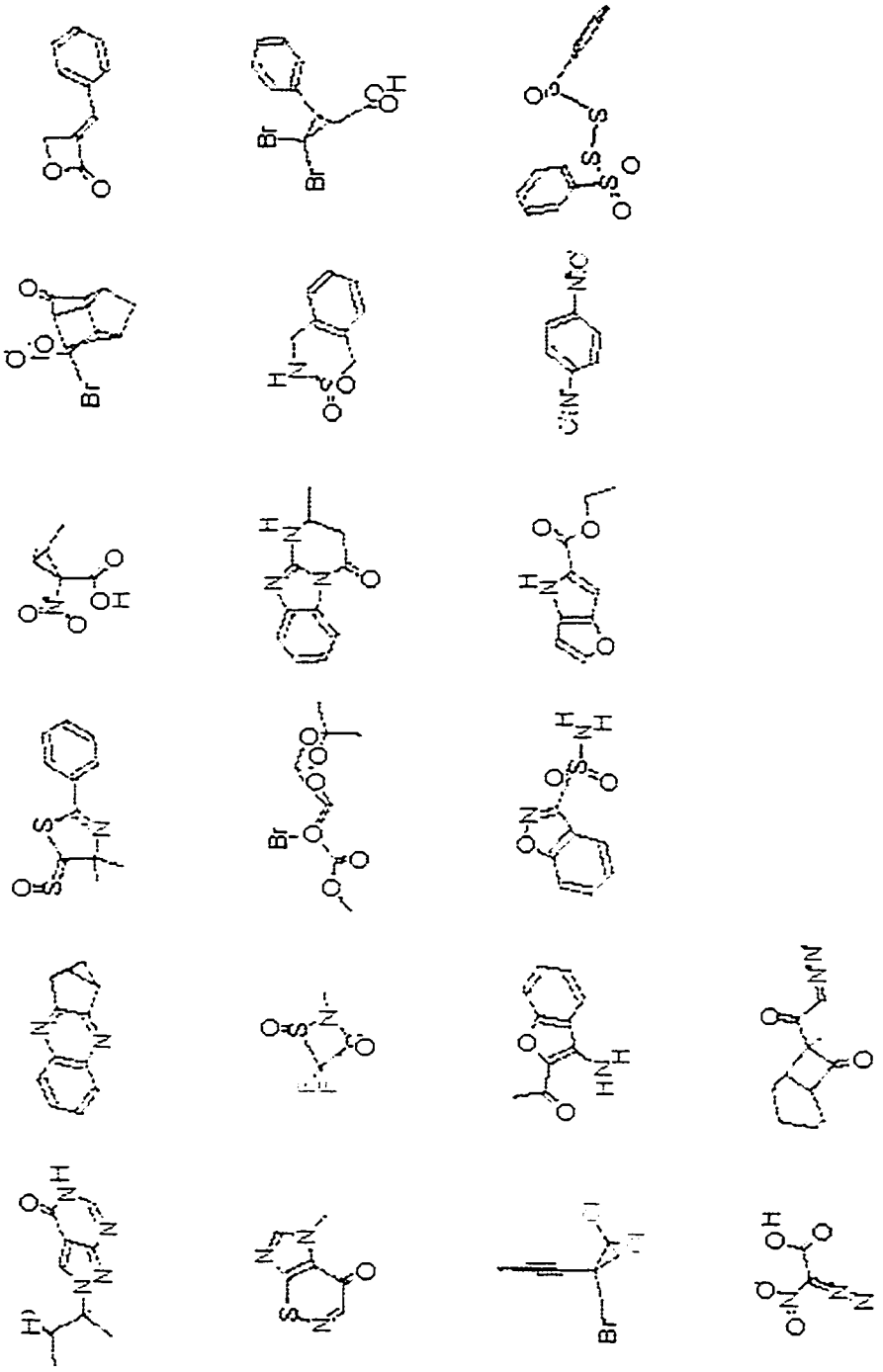
    

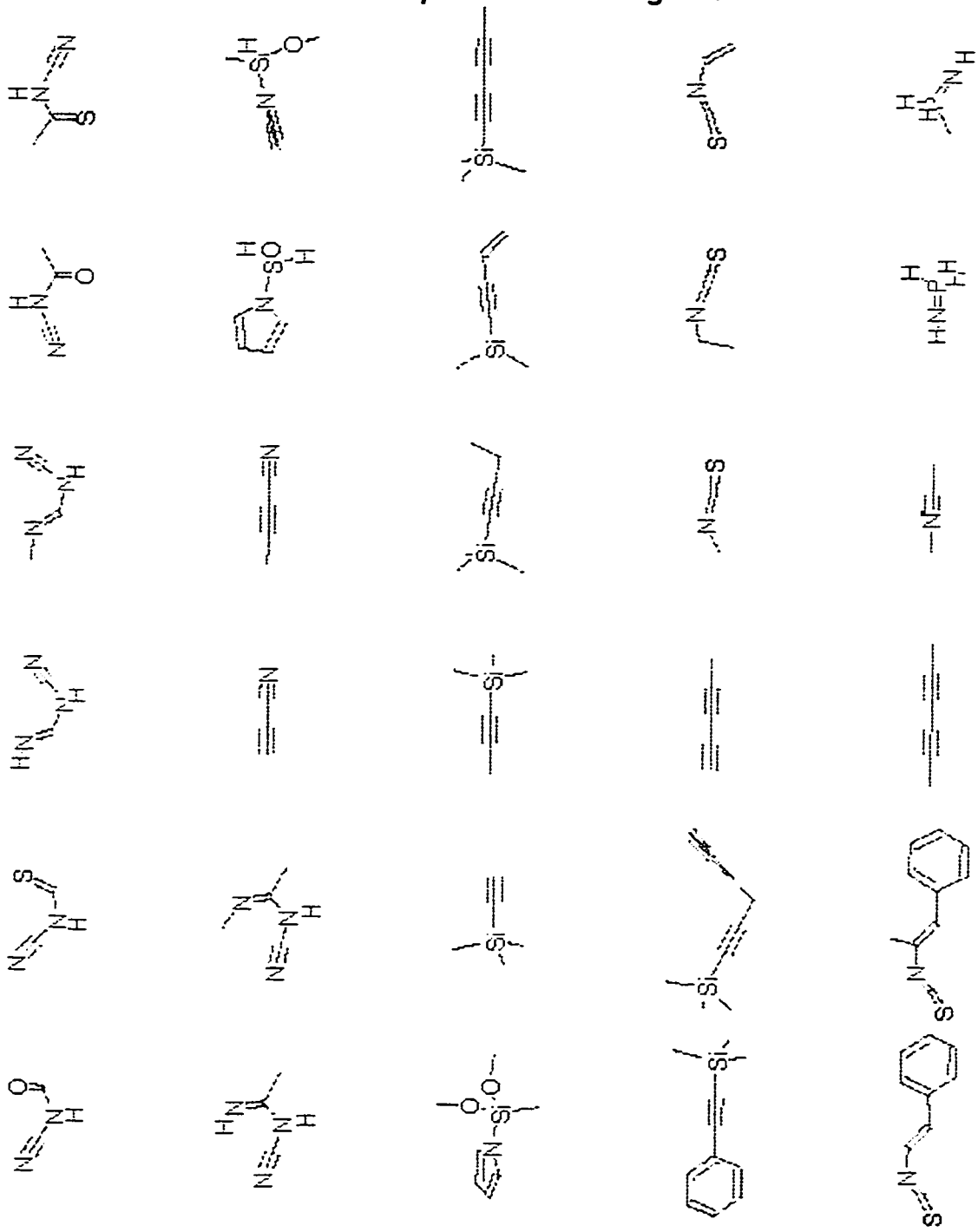
    

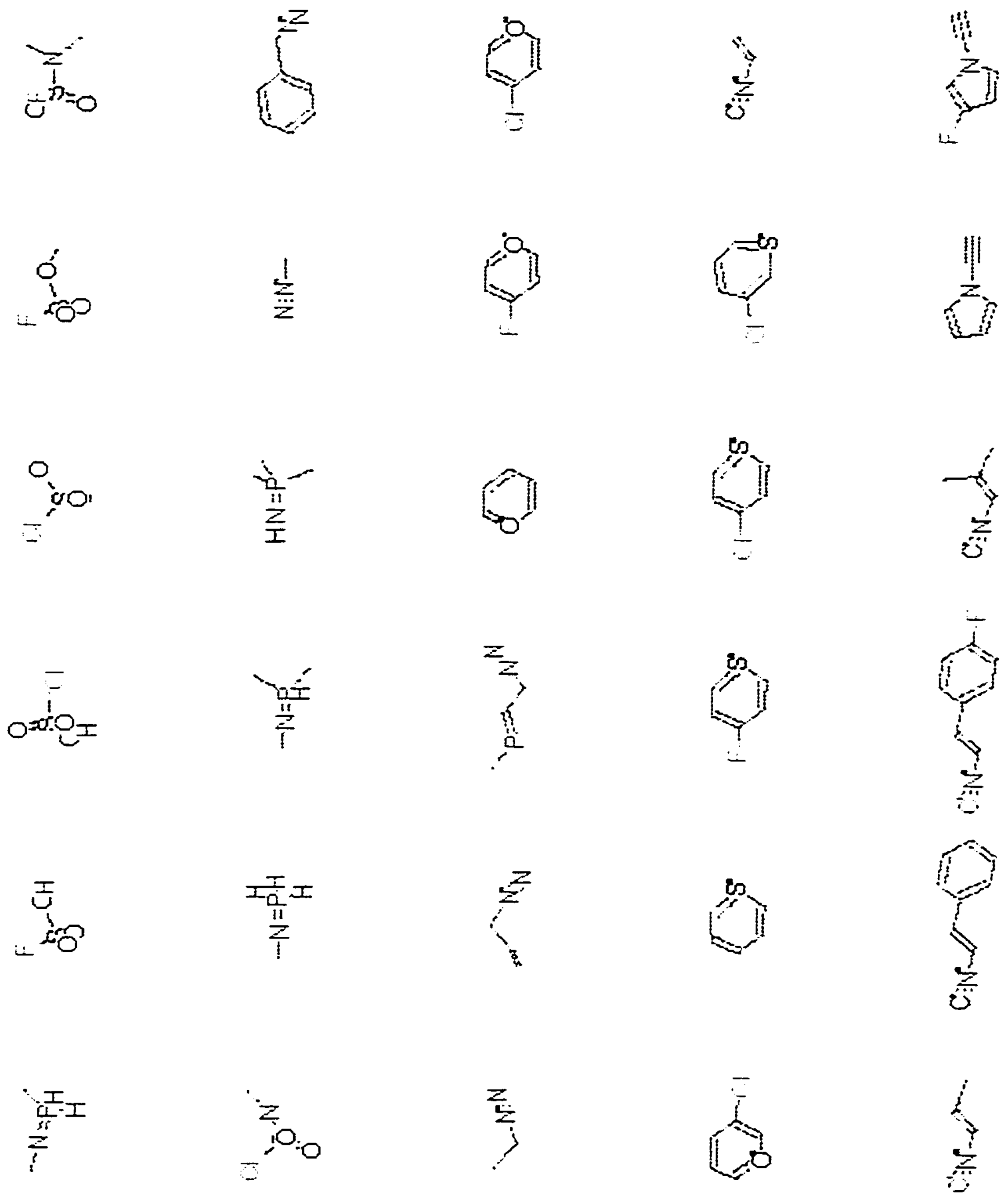


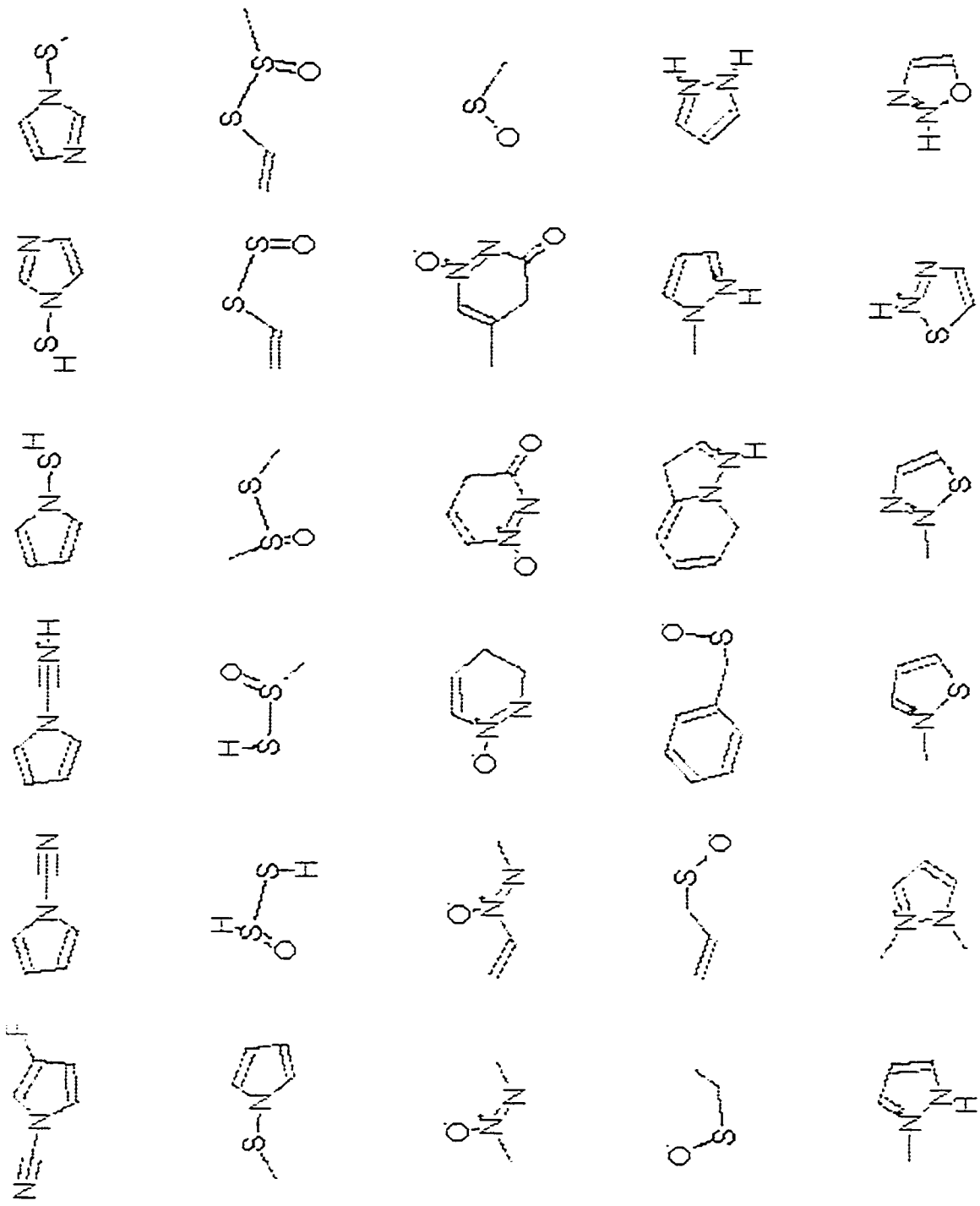


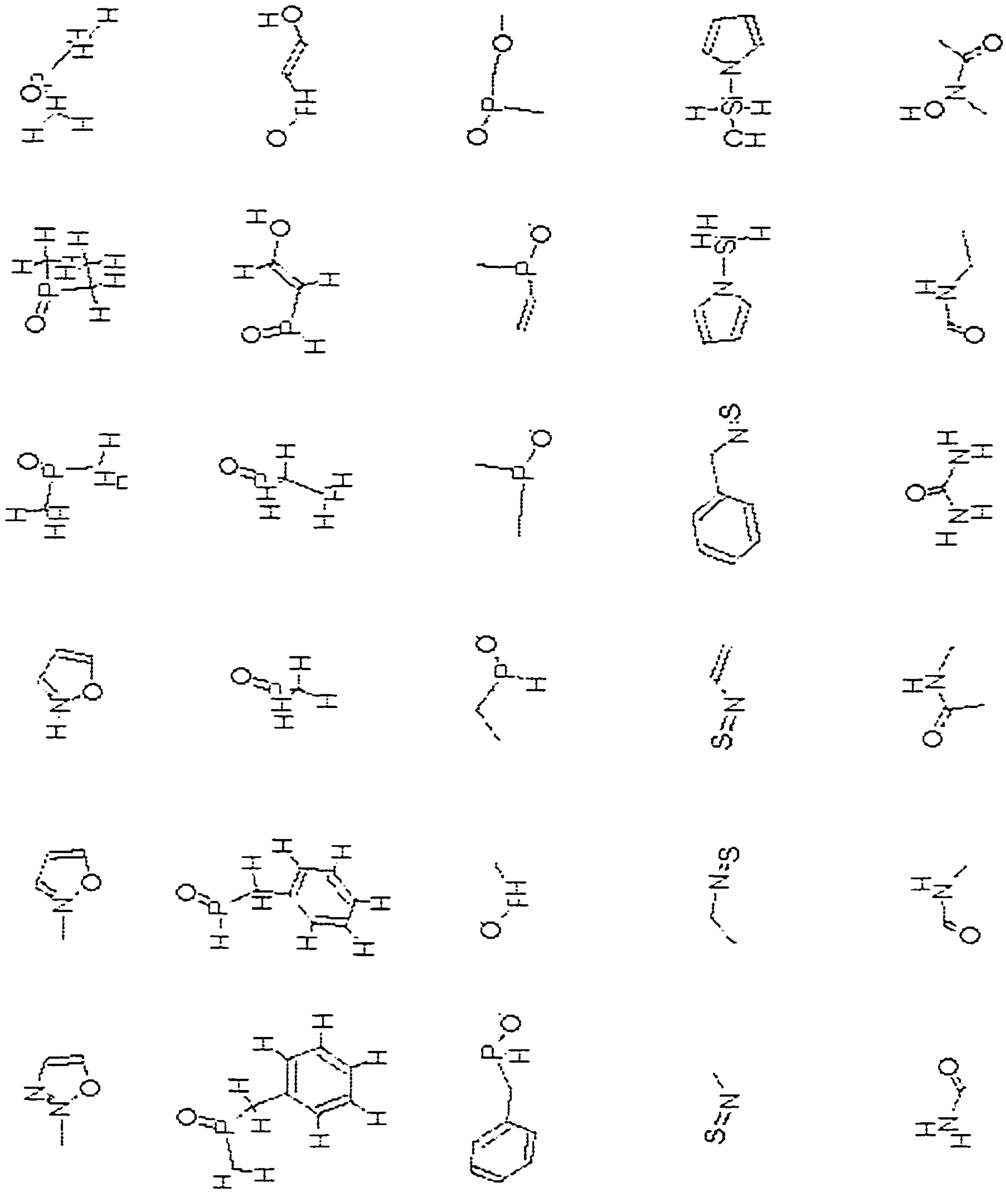


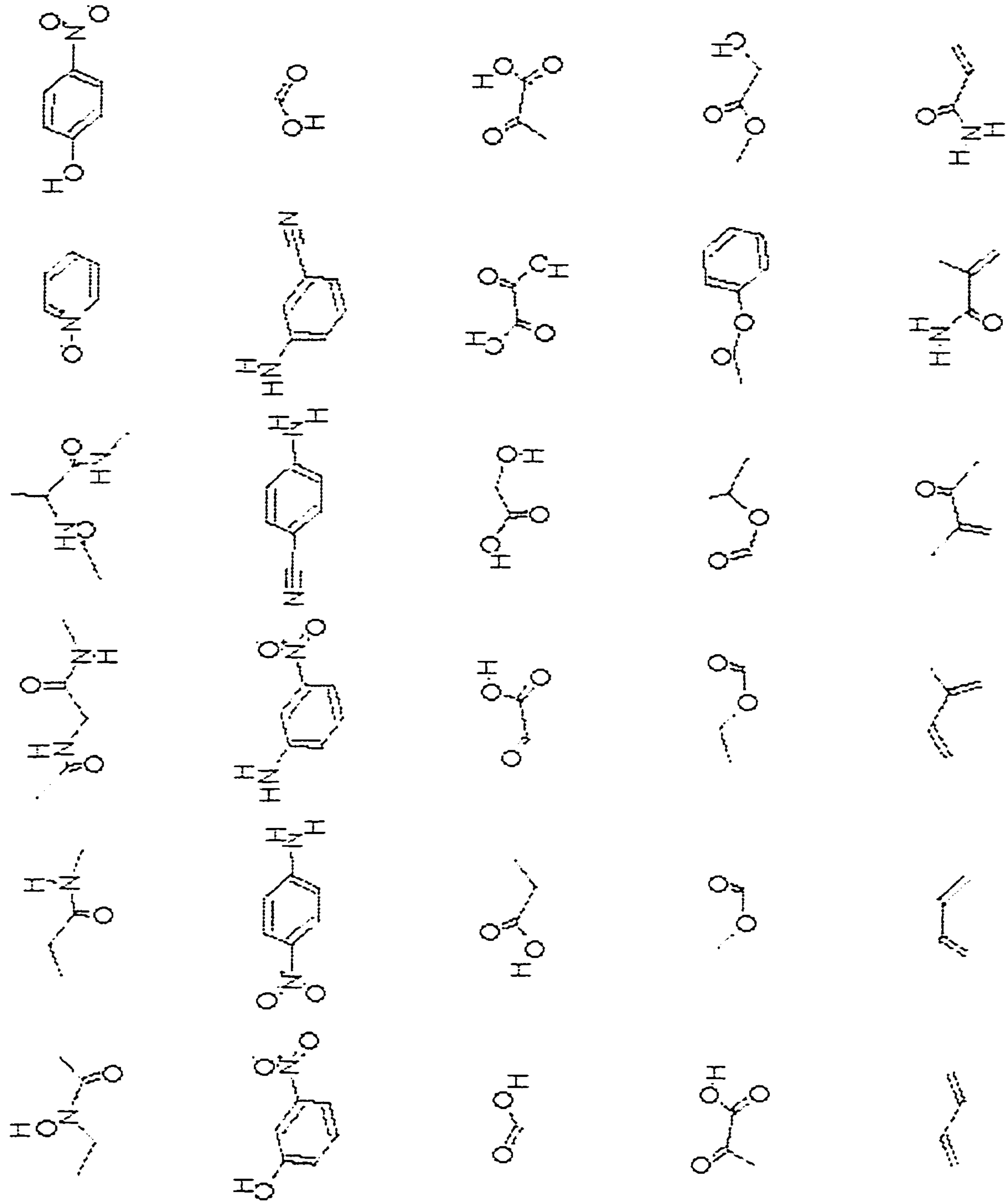
C.10 Extra Molecules to Complete the Training Set

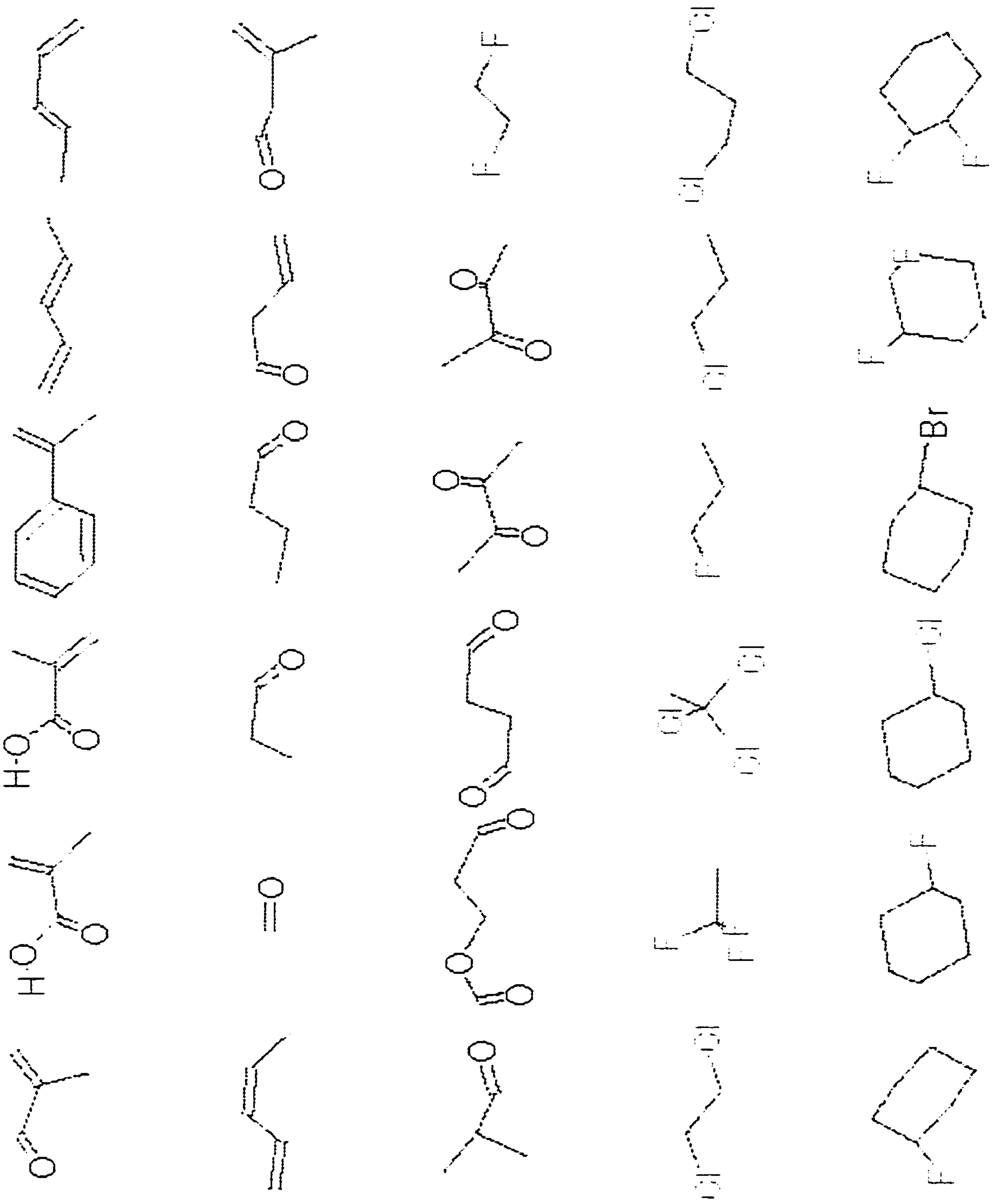


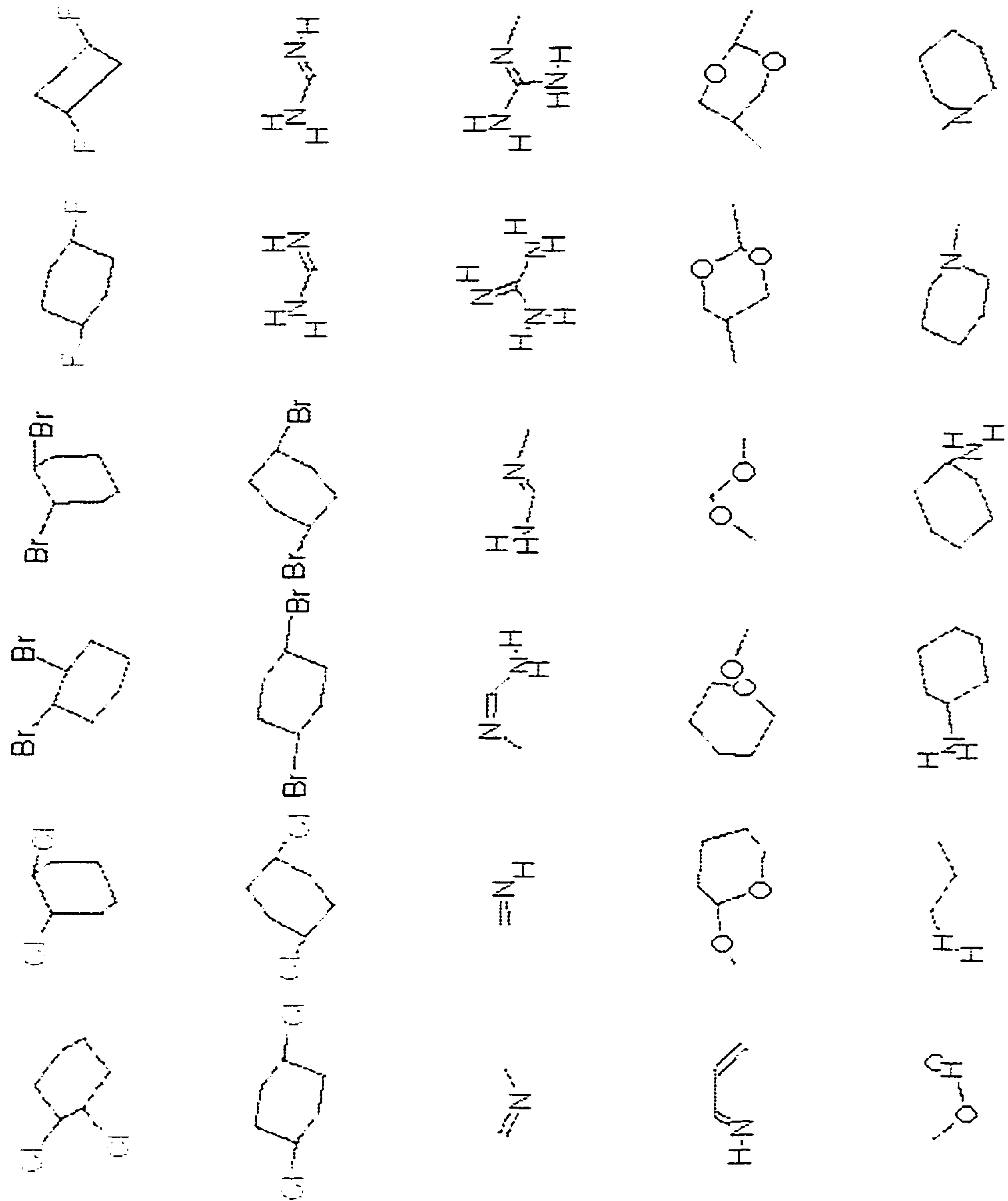


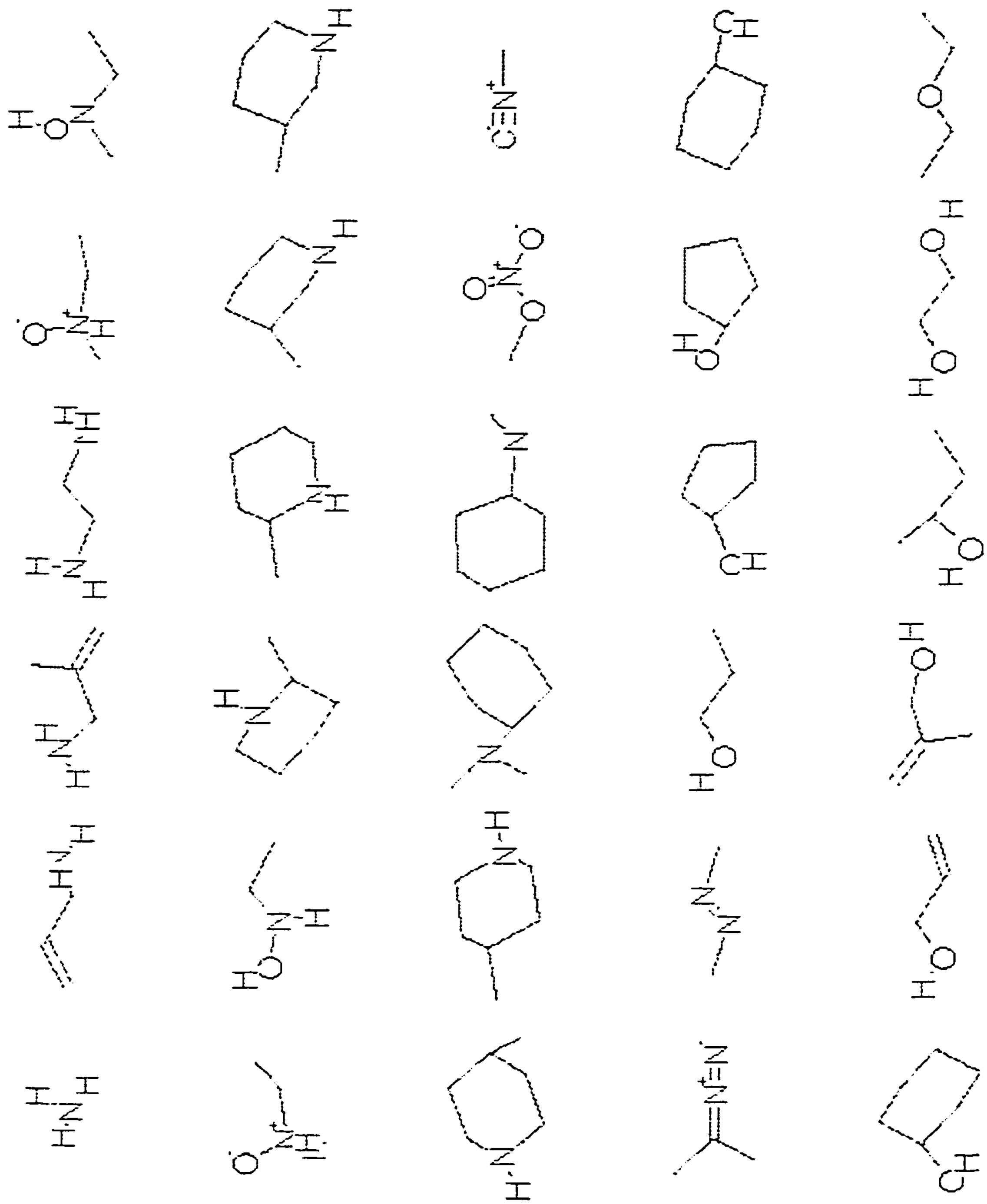


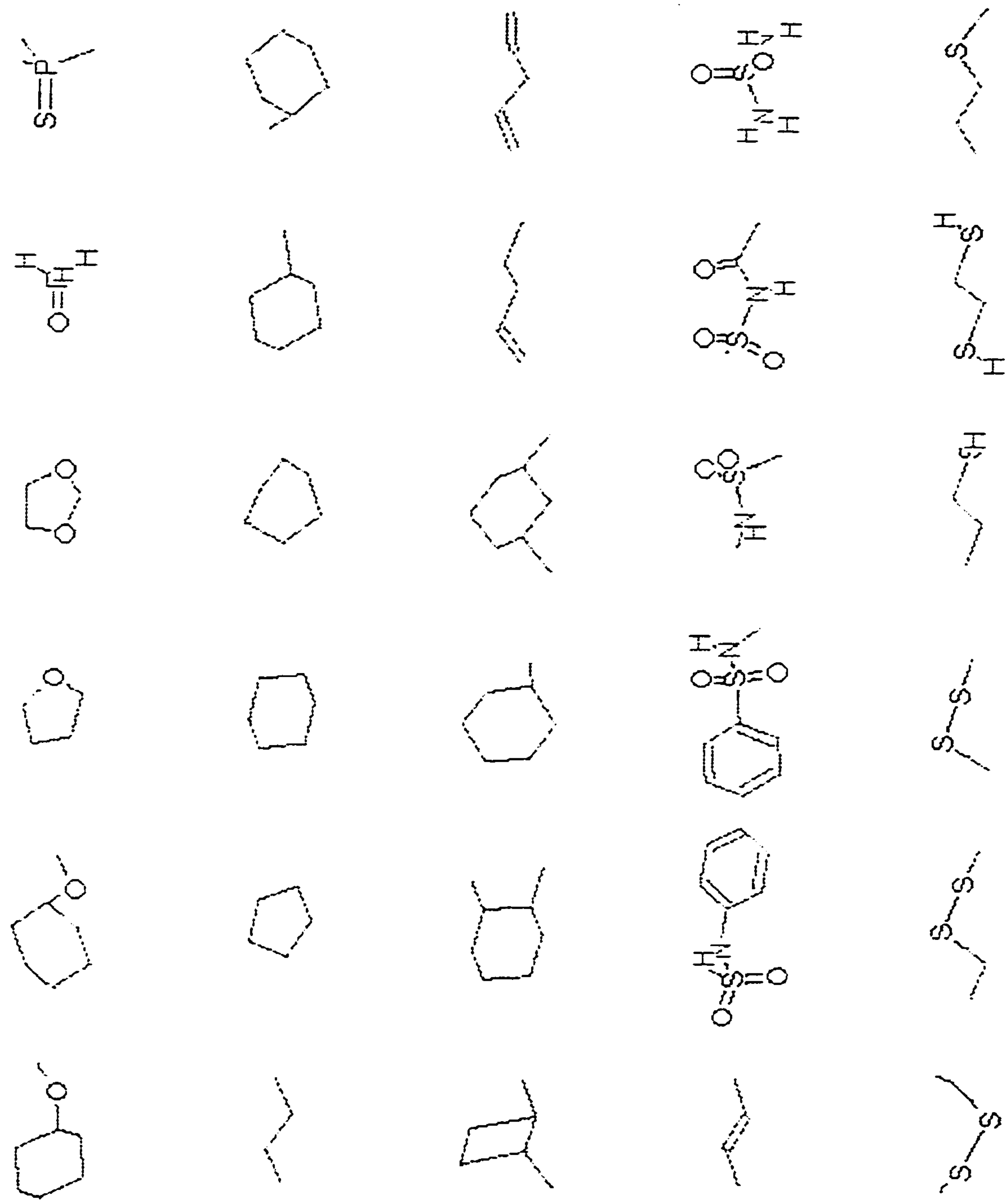


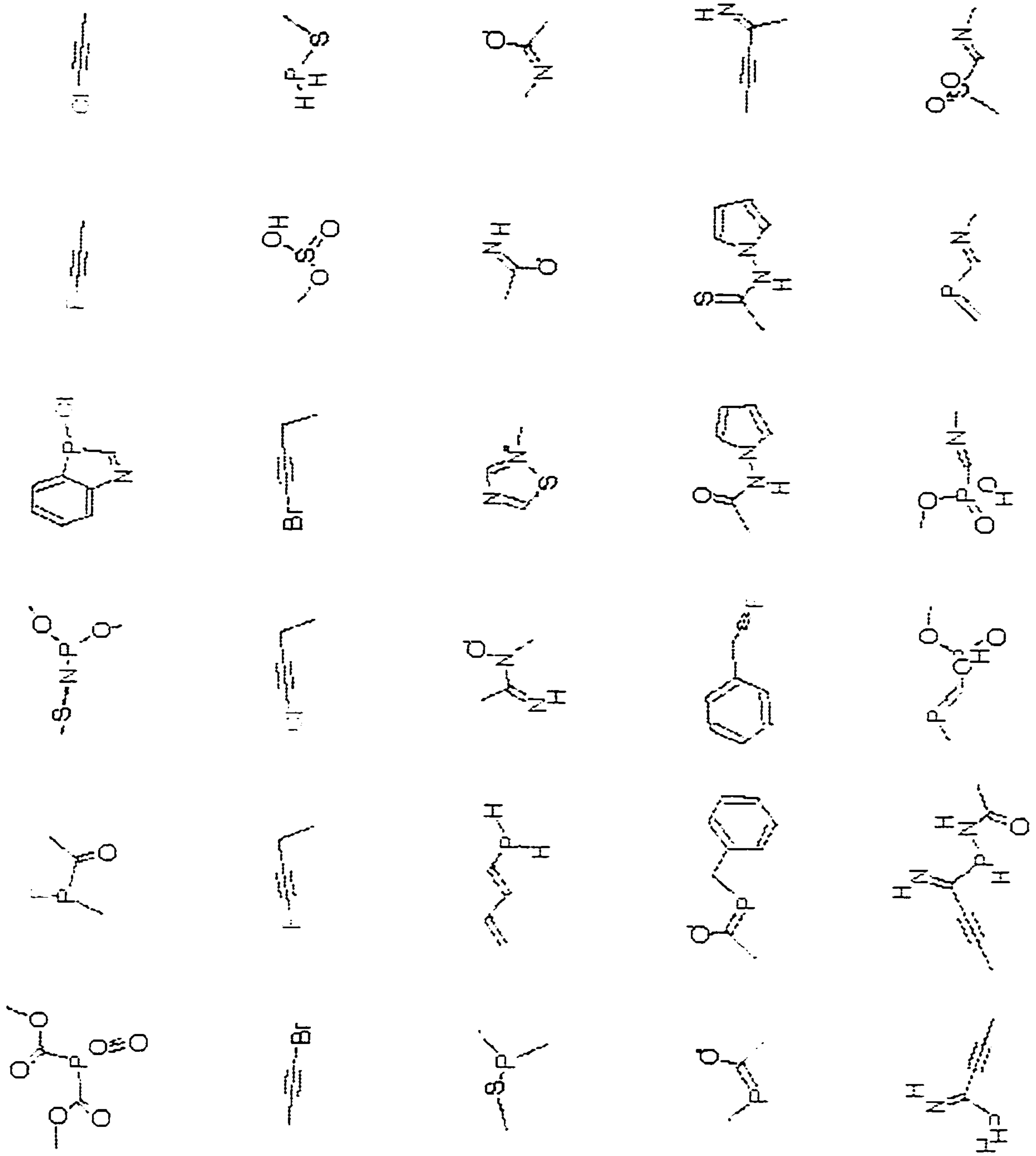


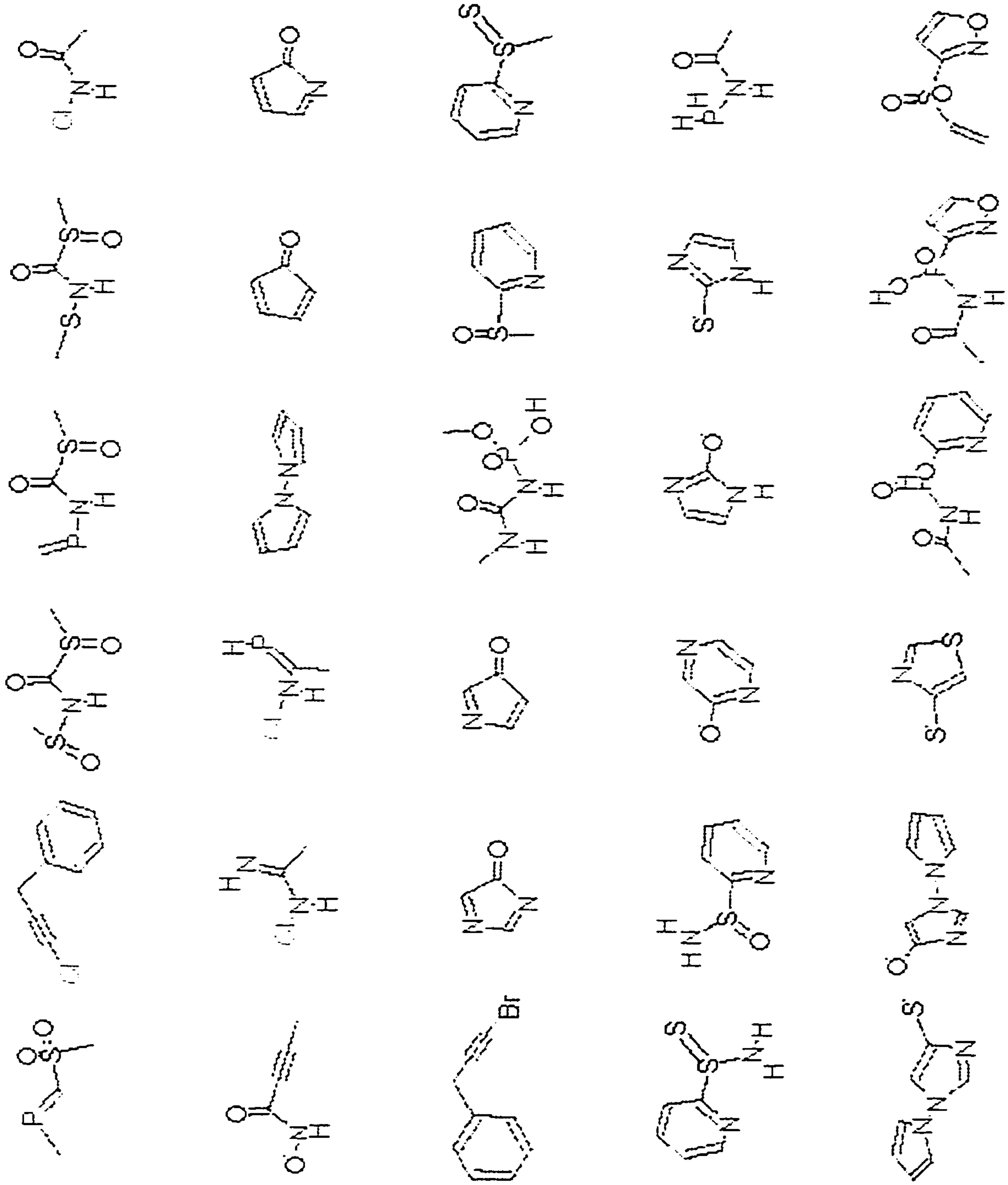


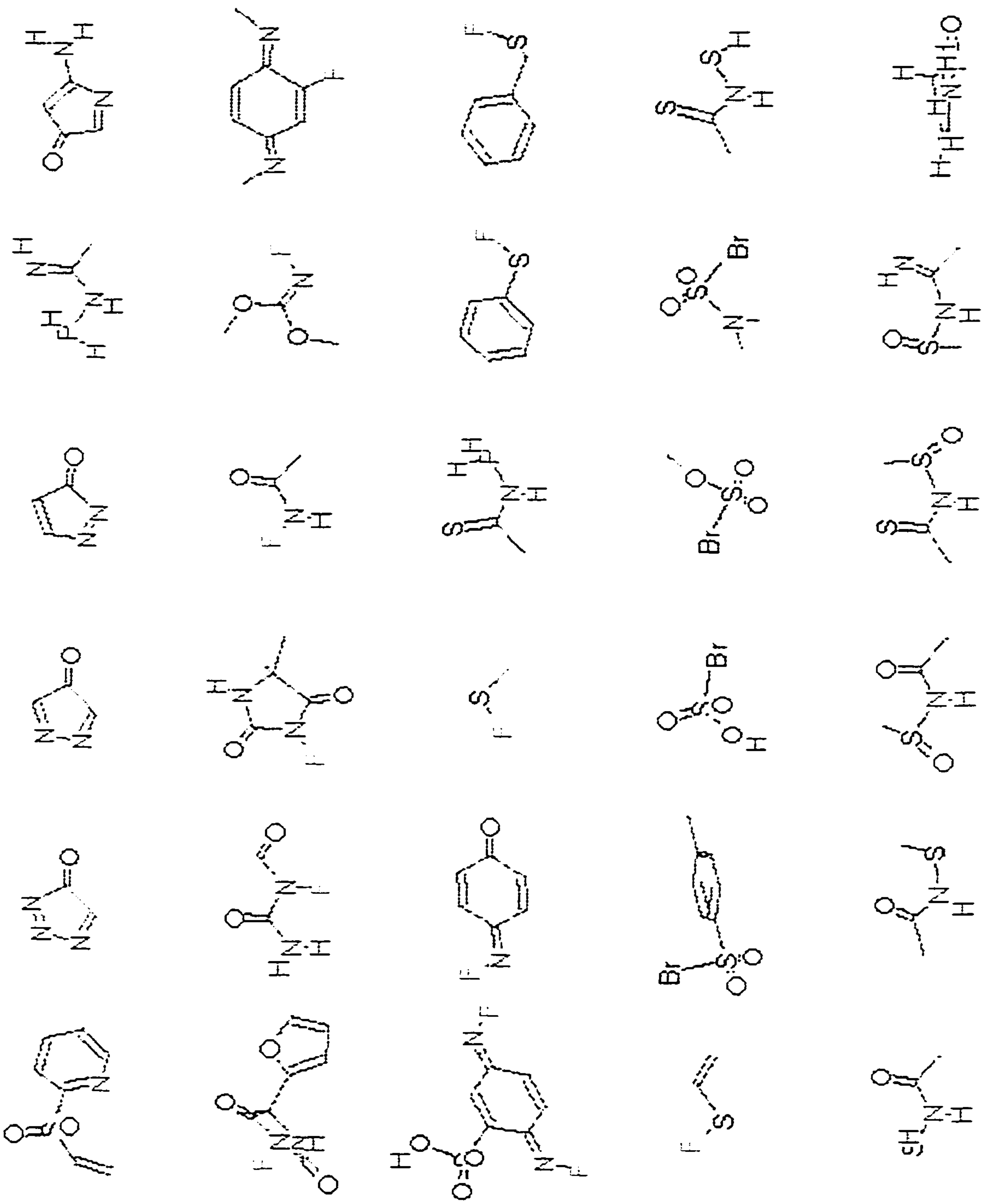


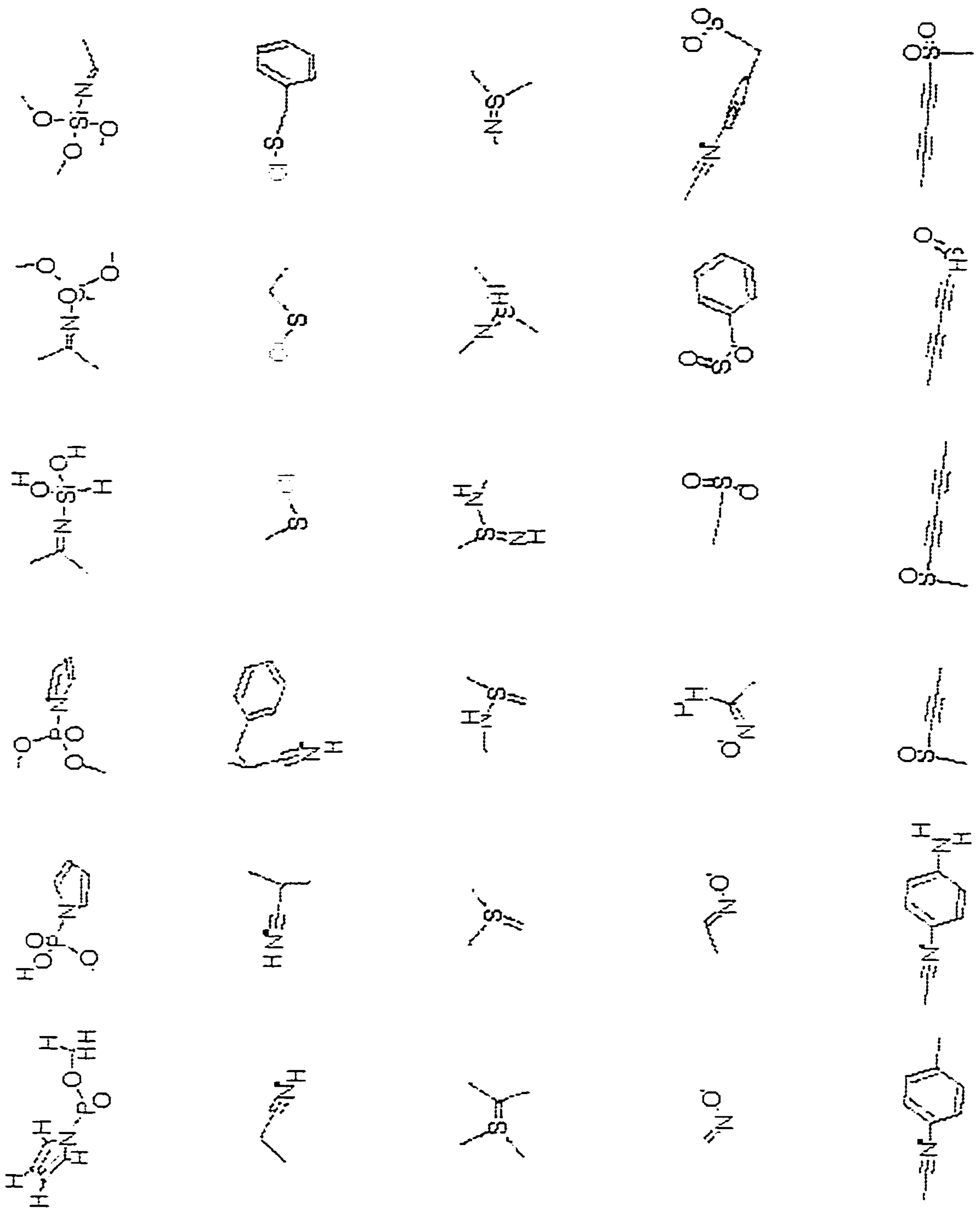


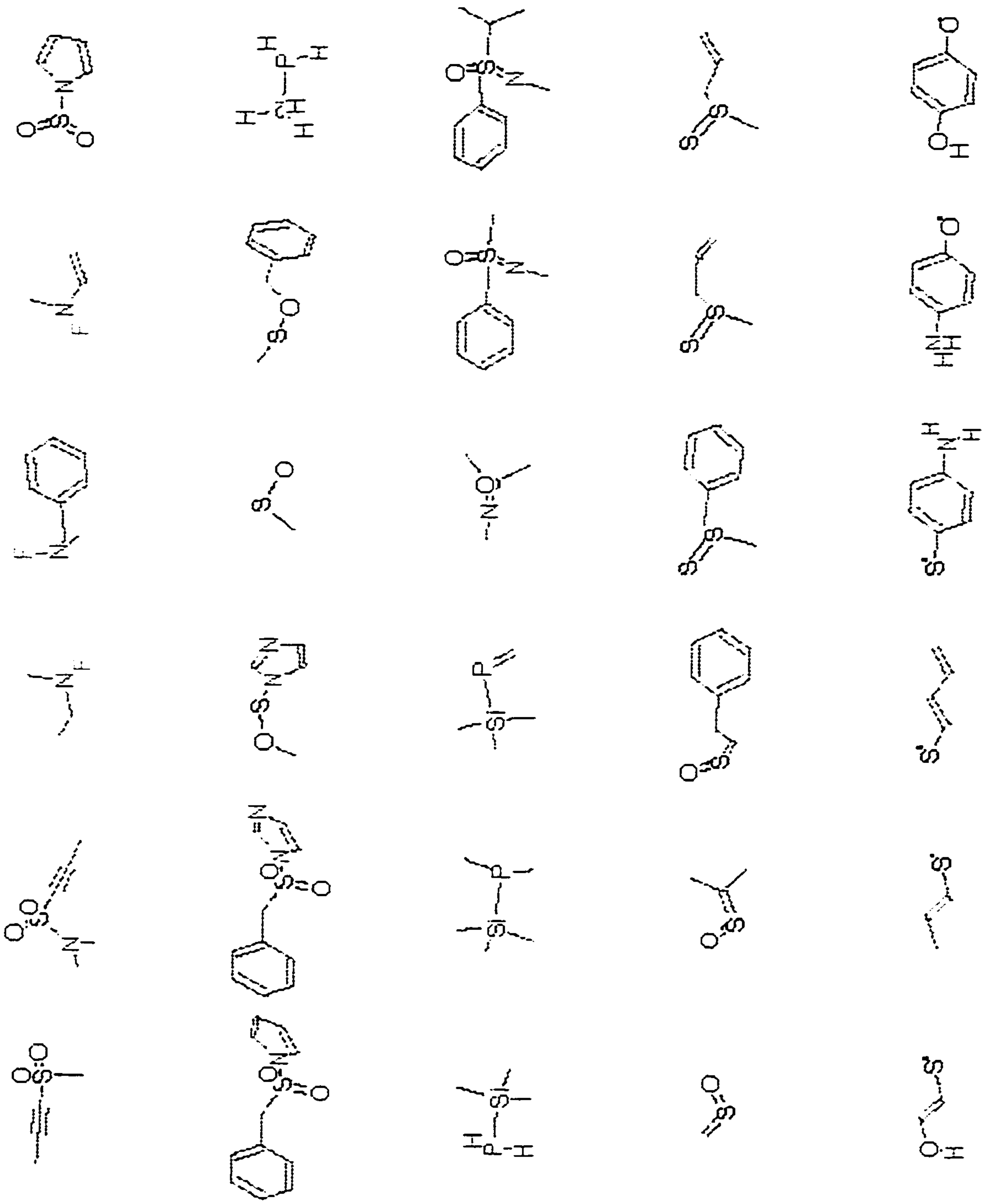


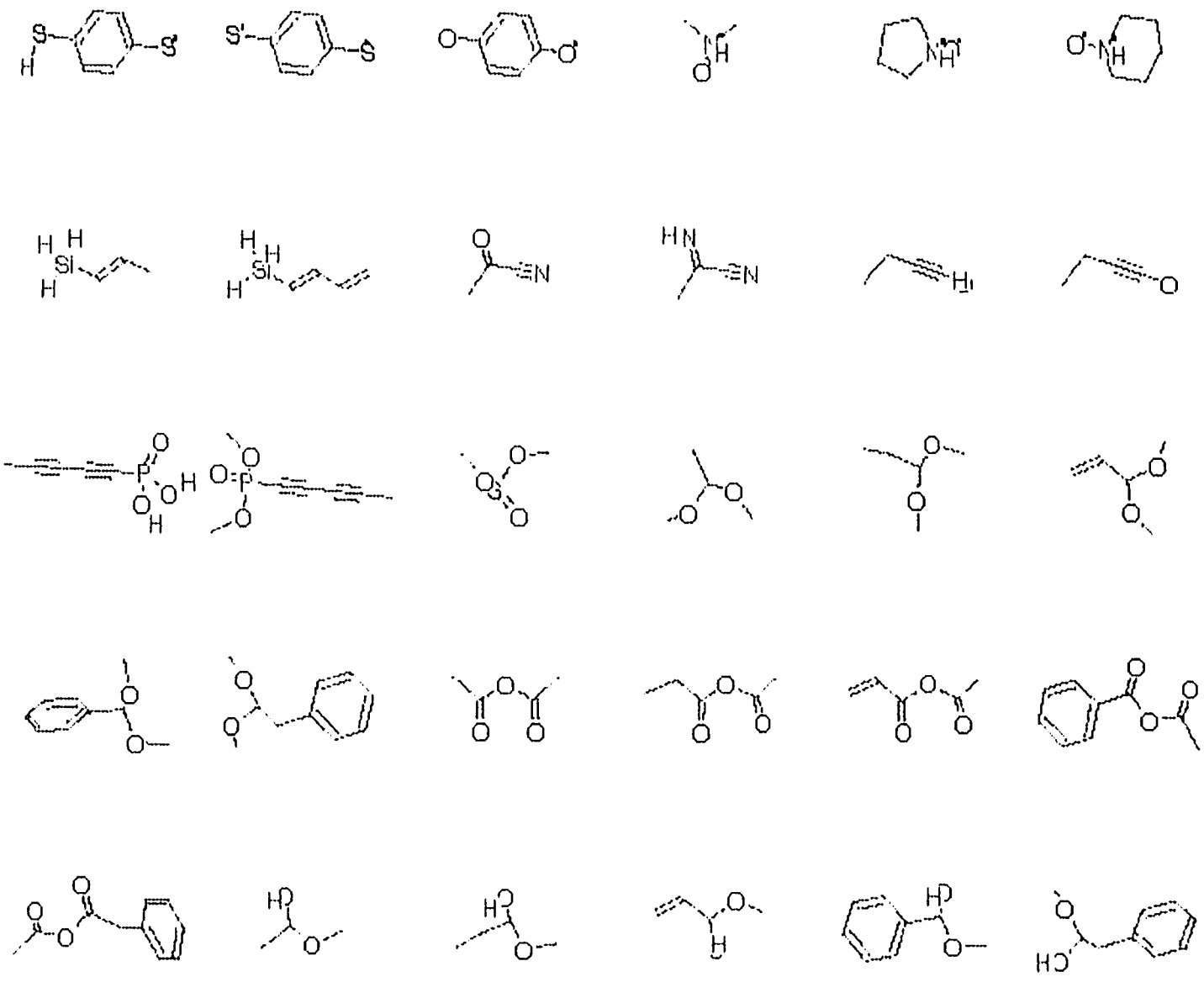


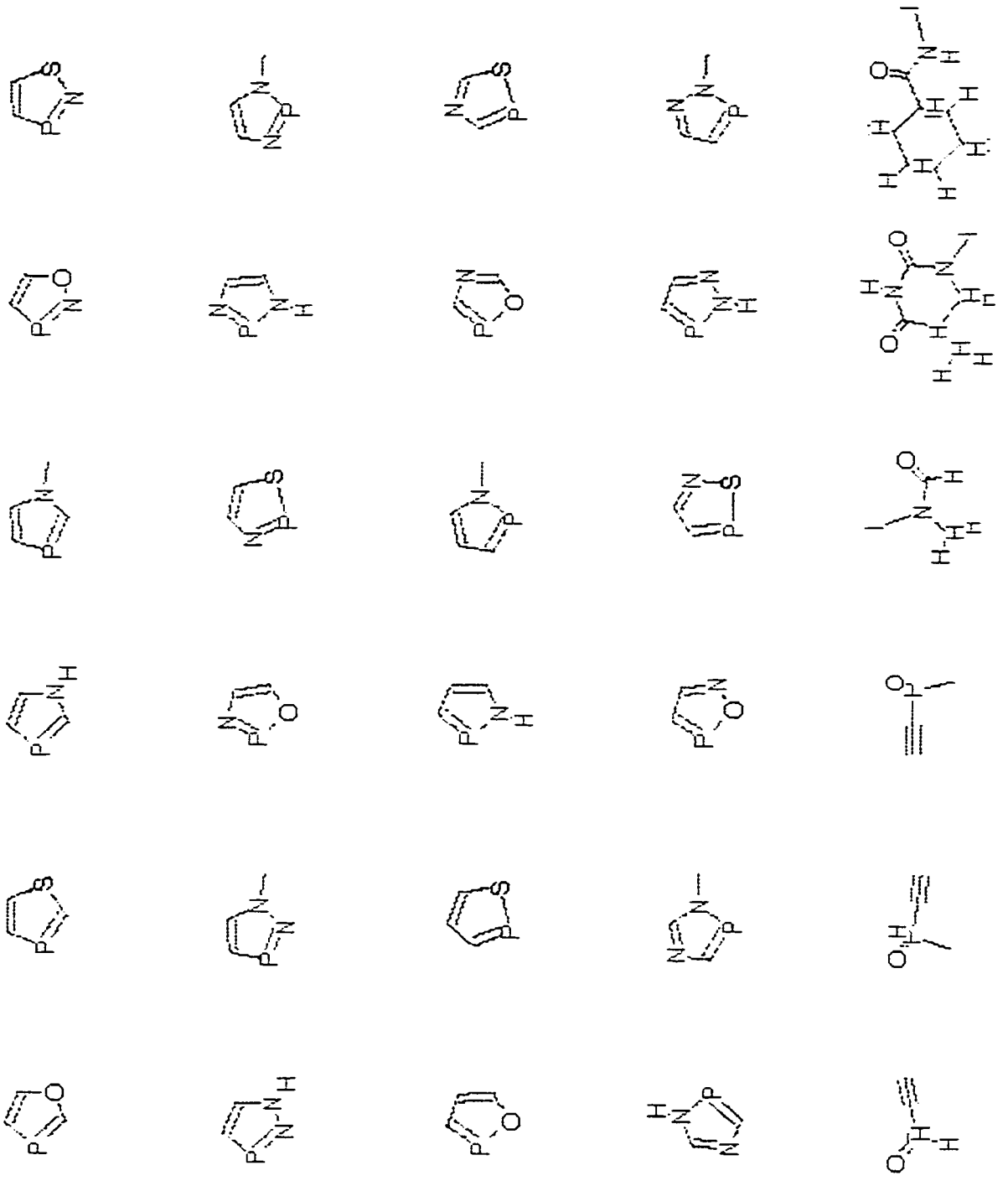


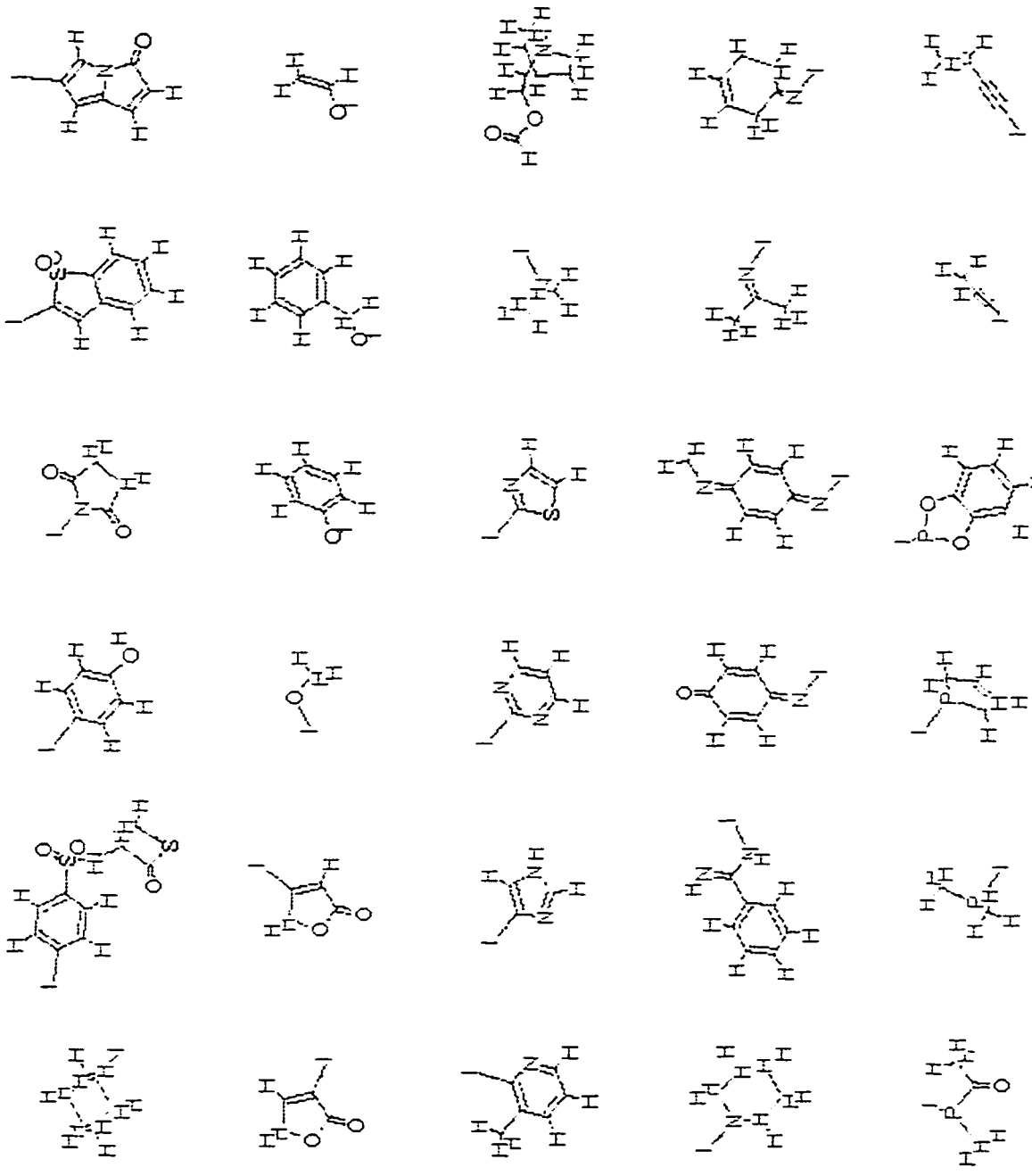


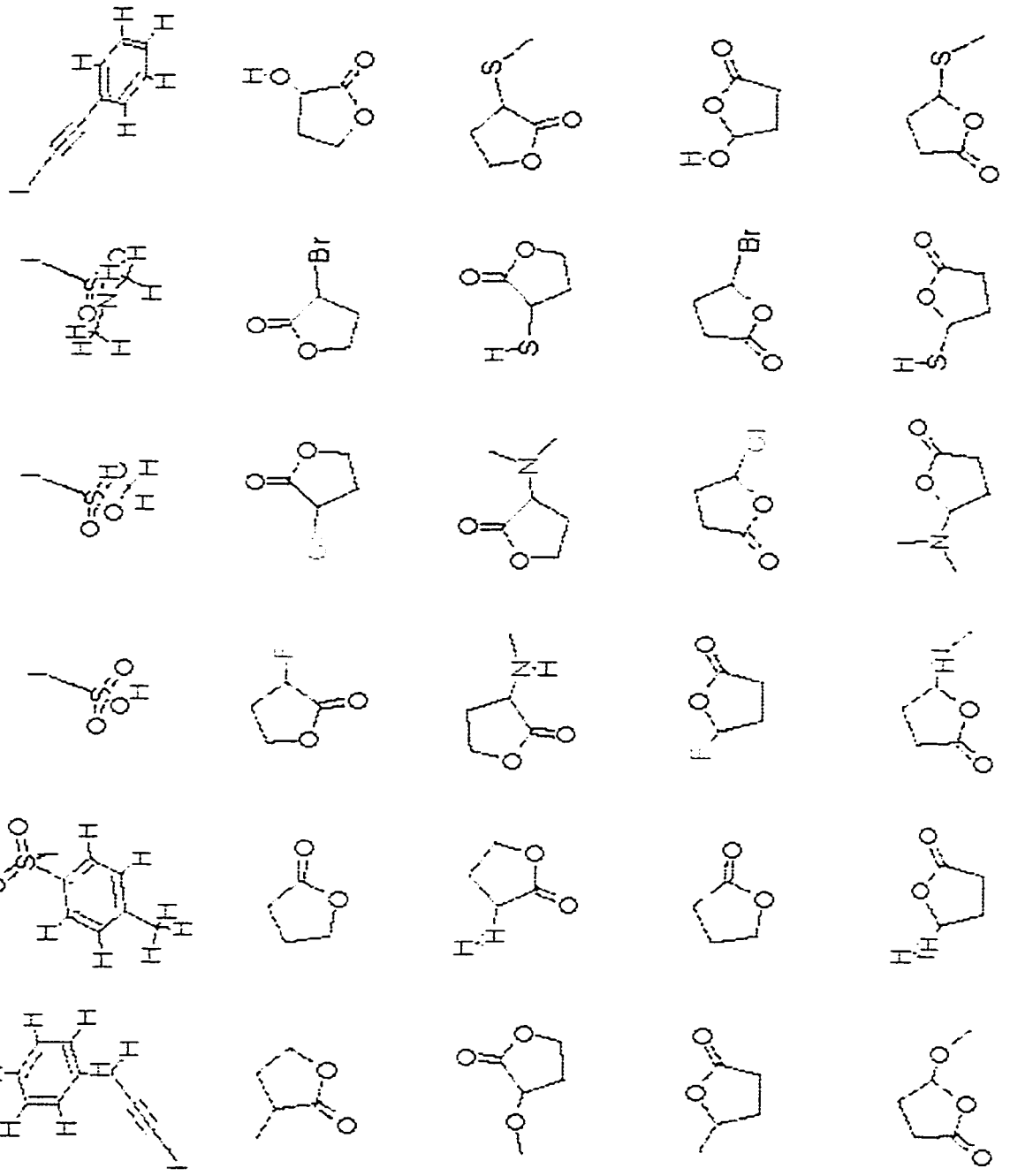


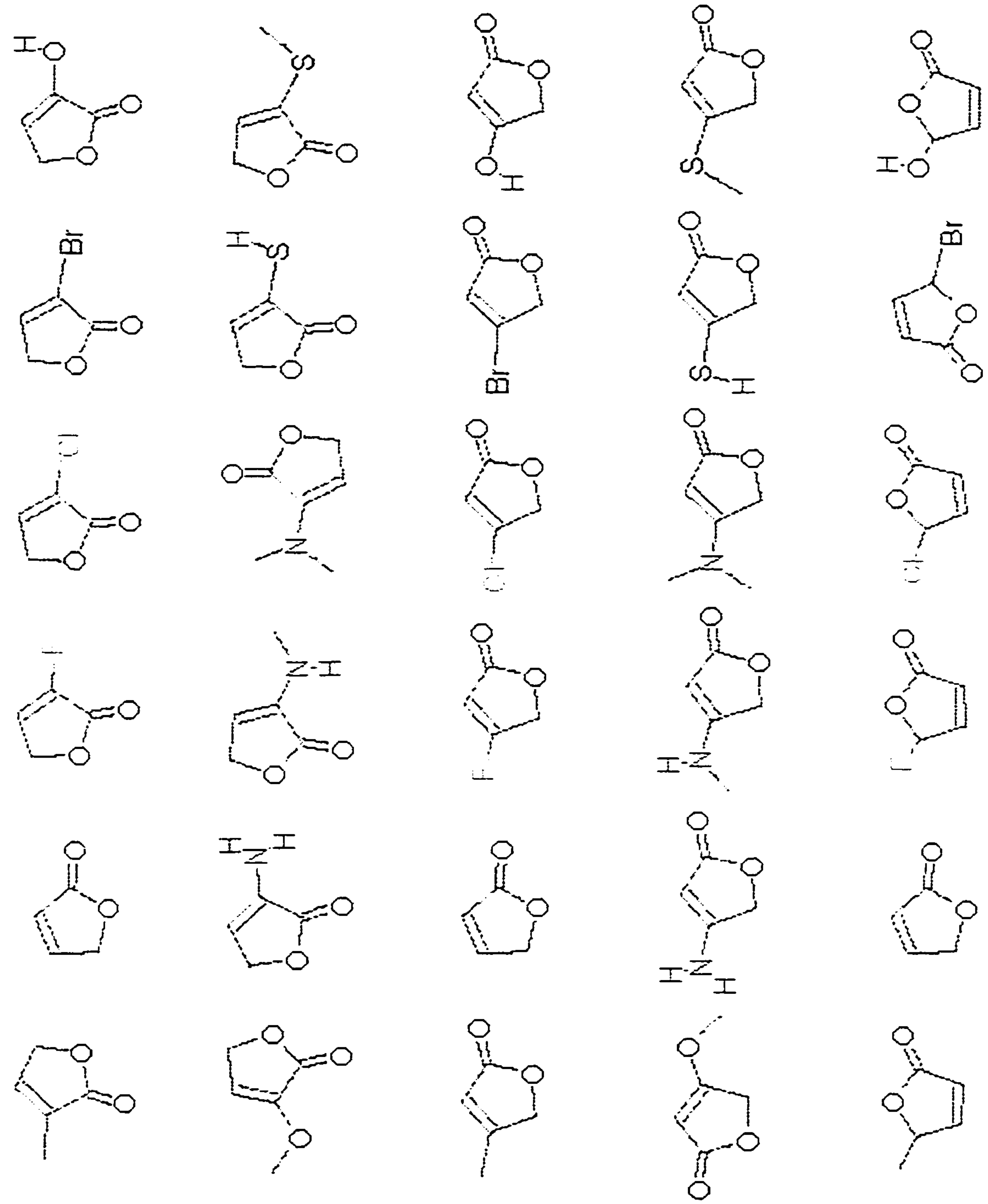


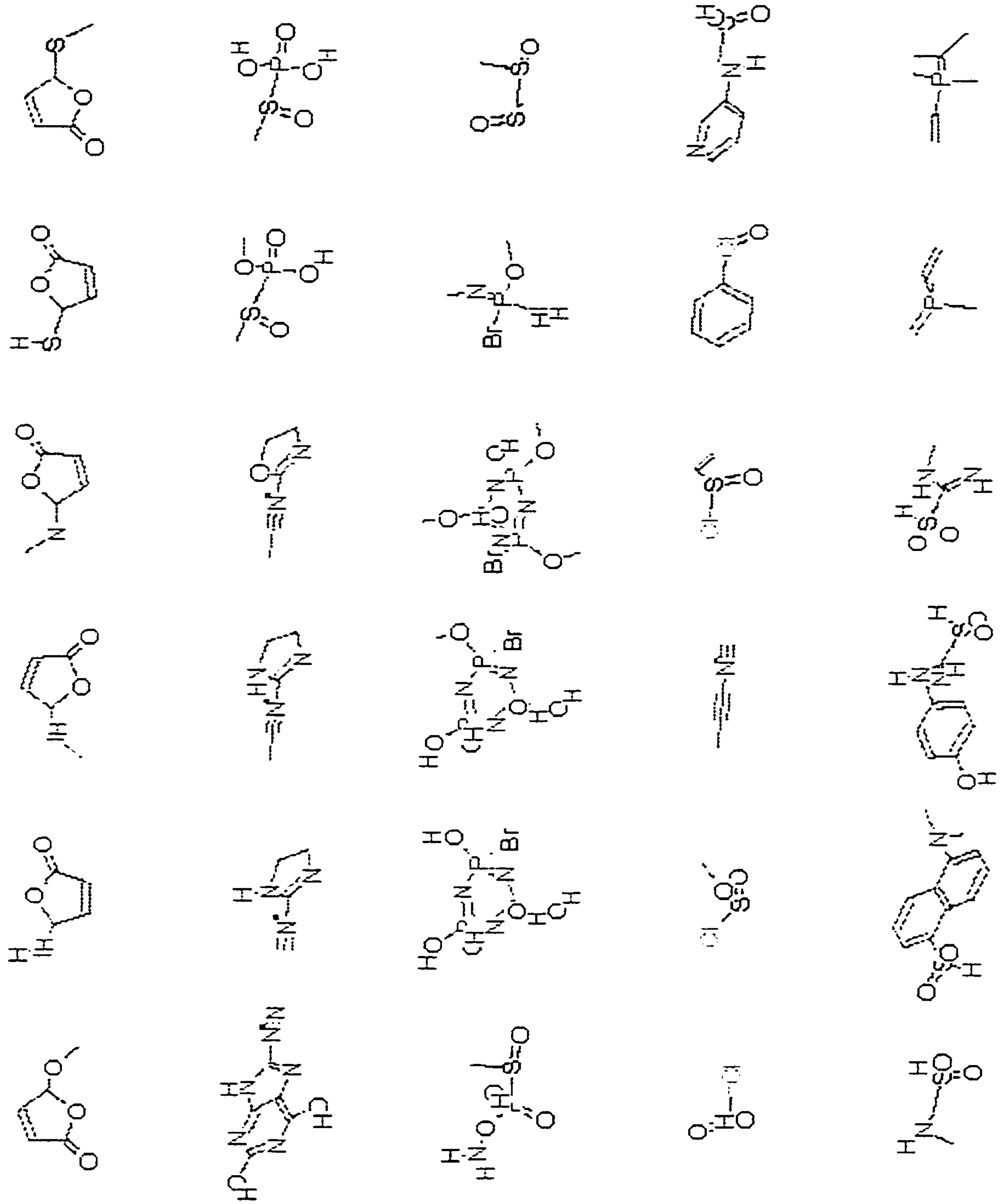


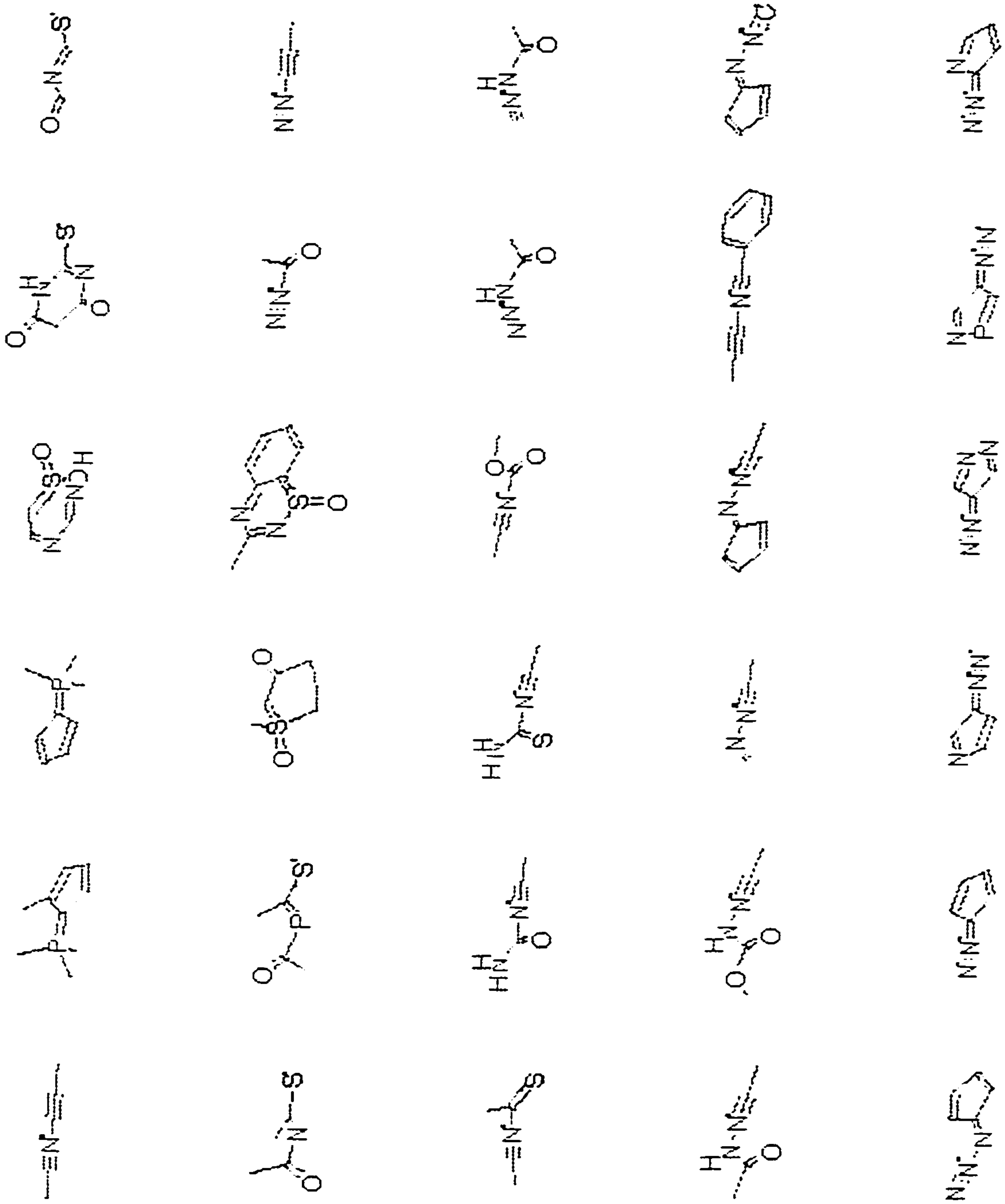


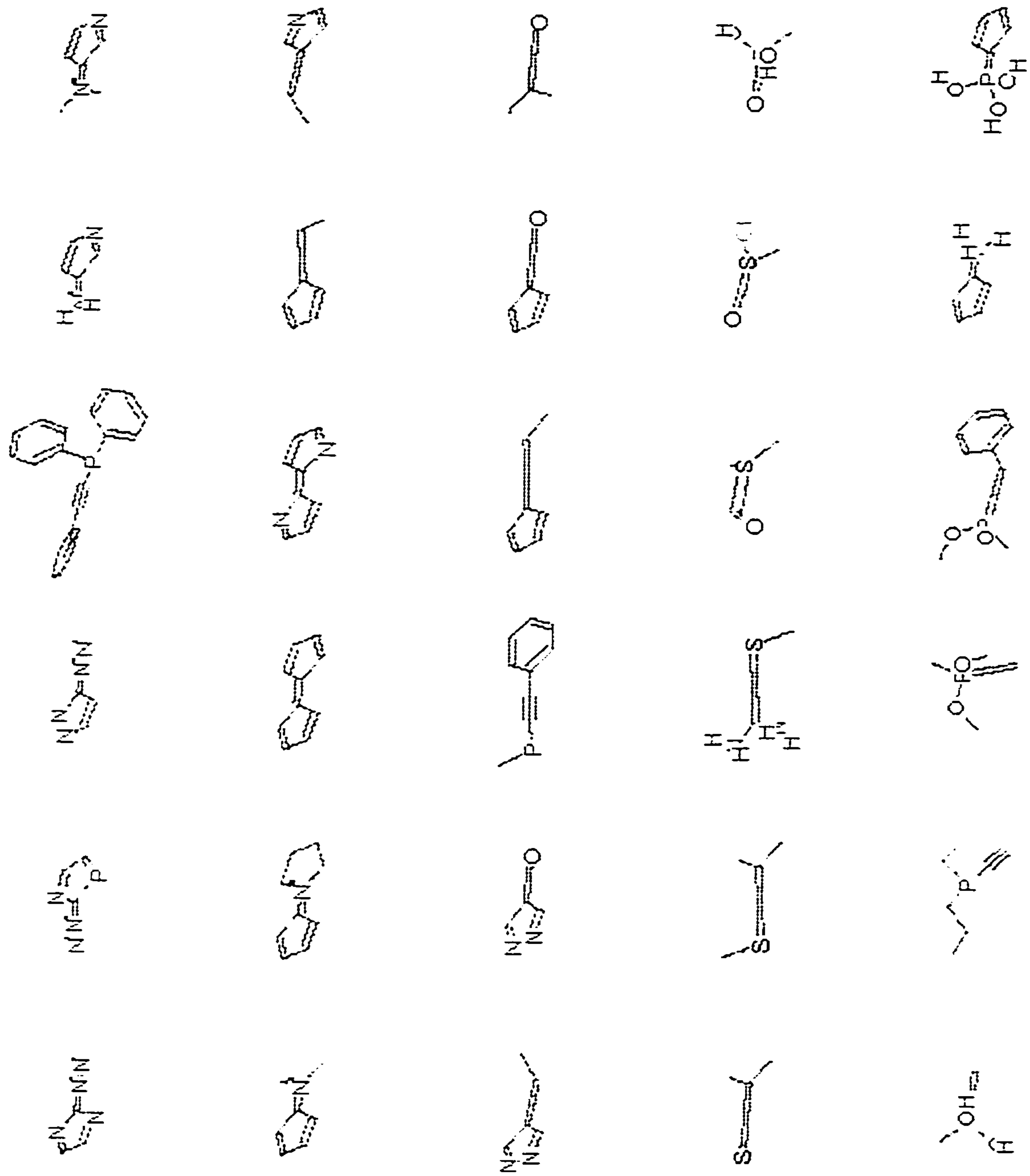


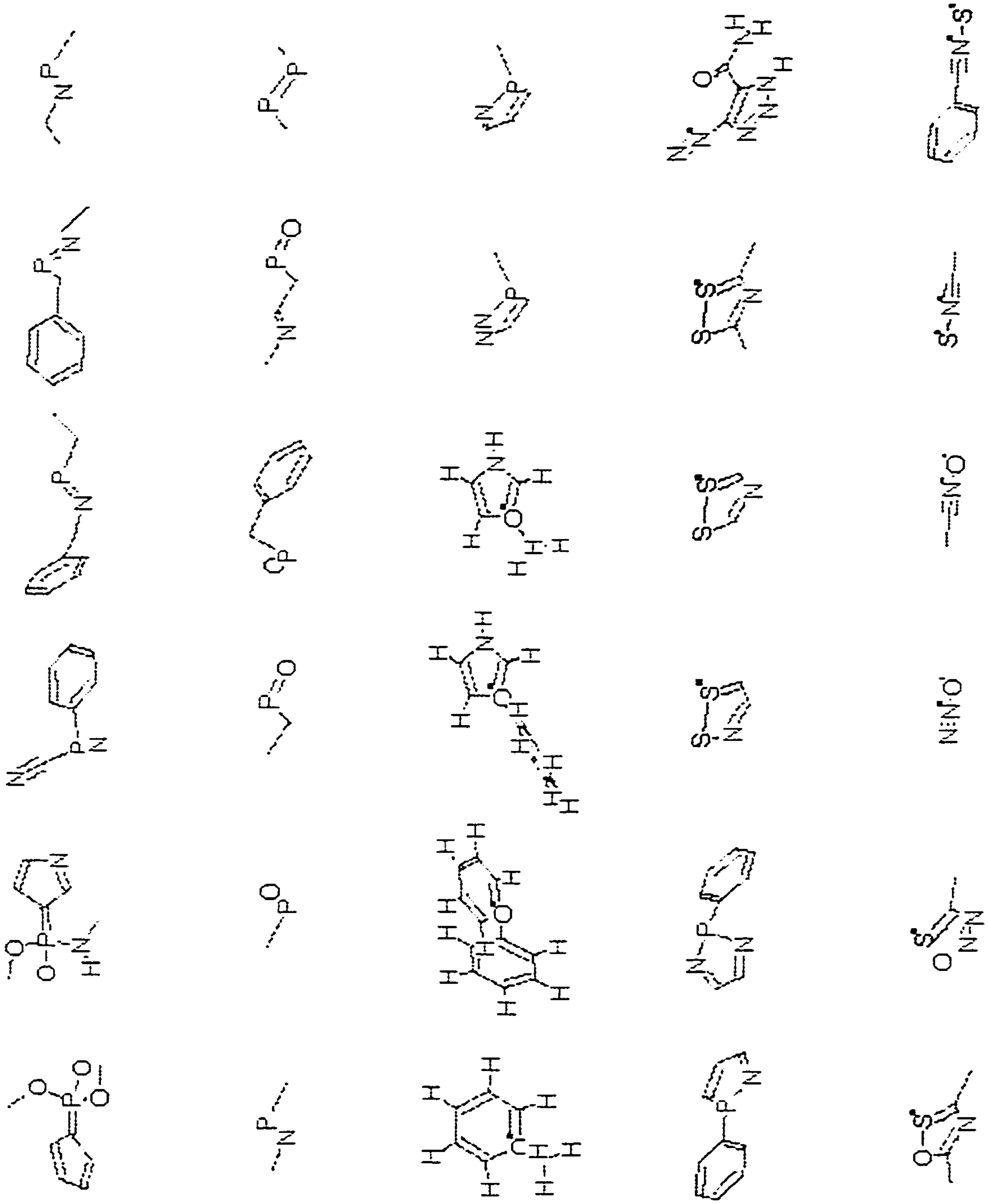


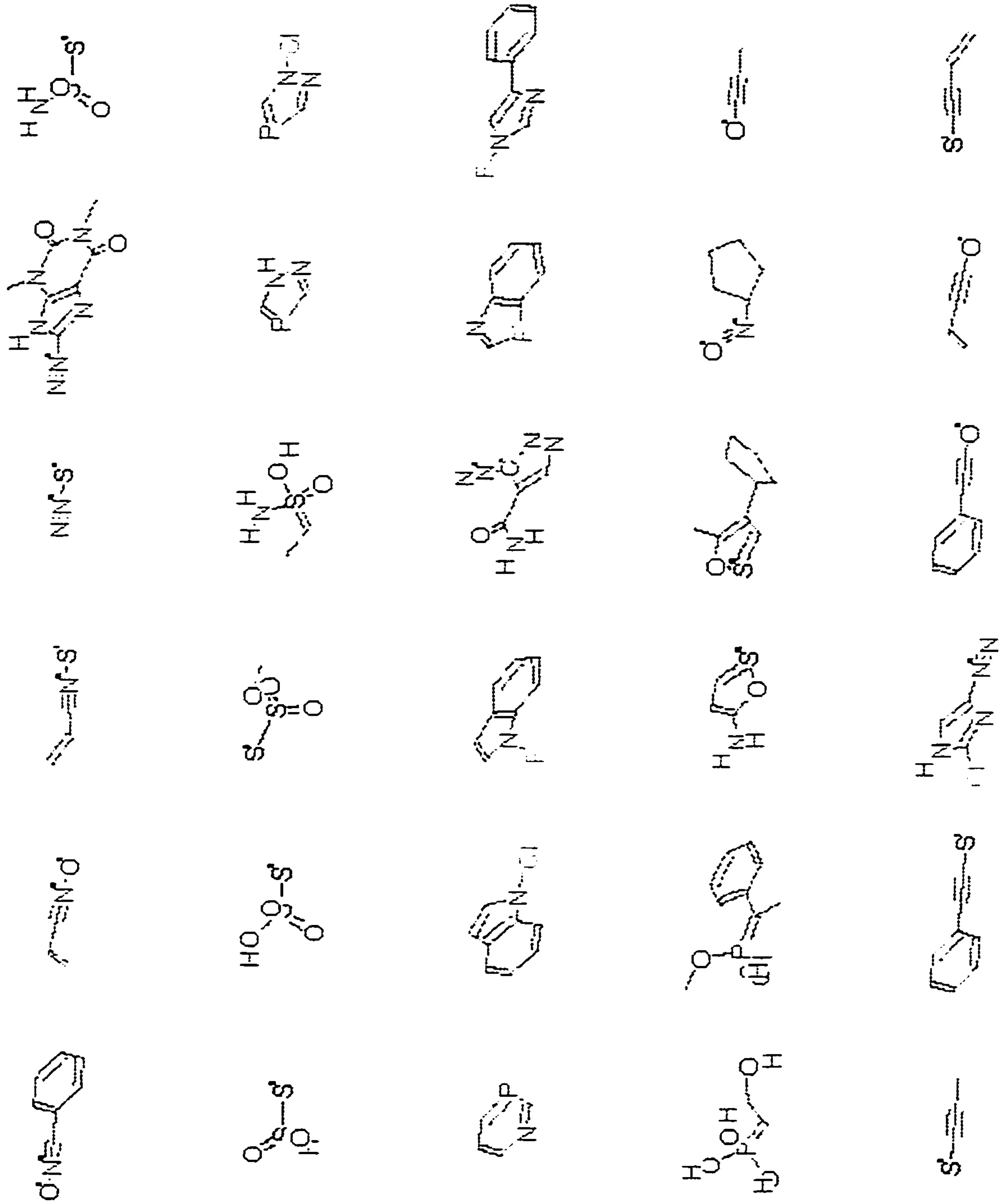


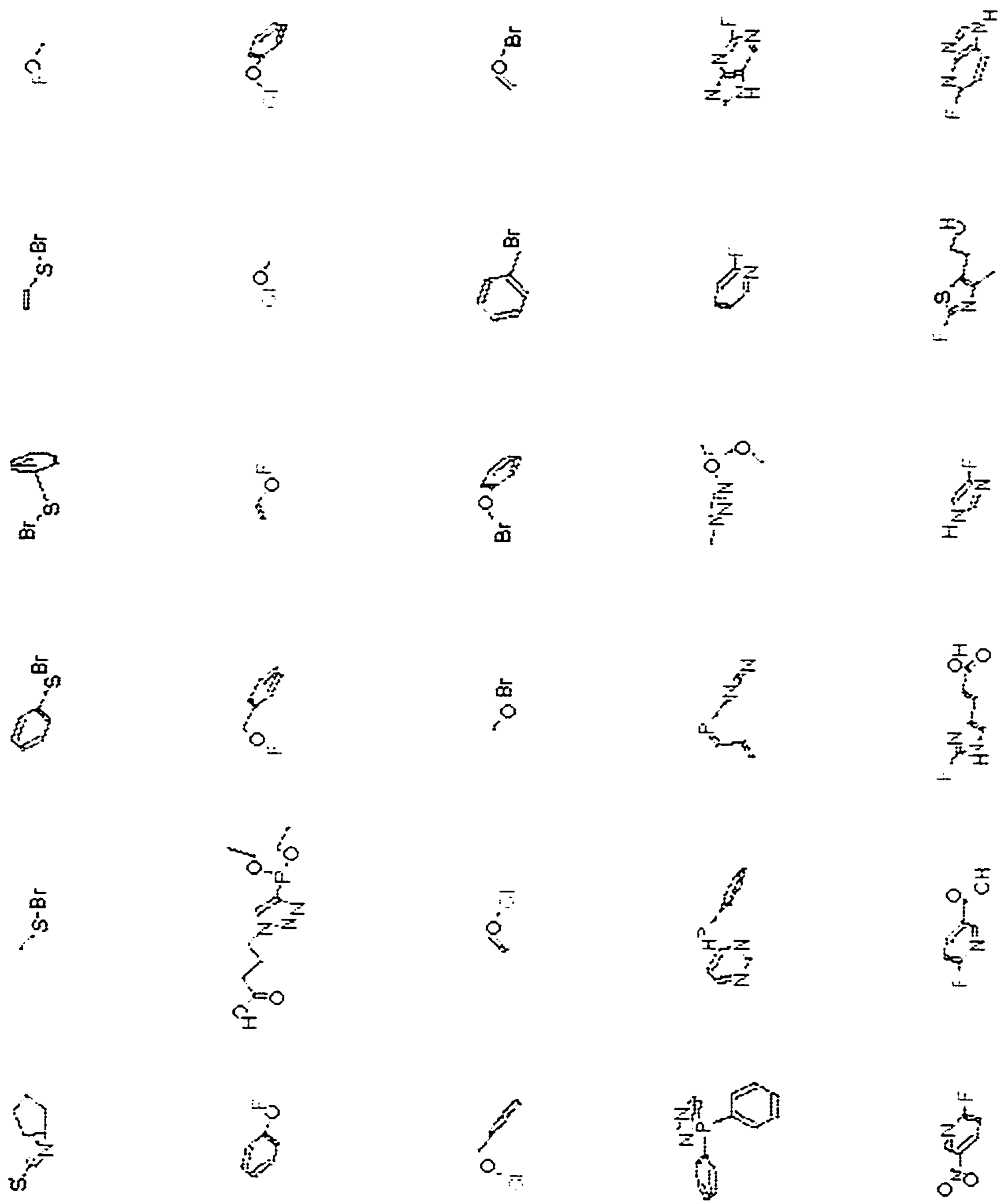


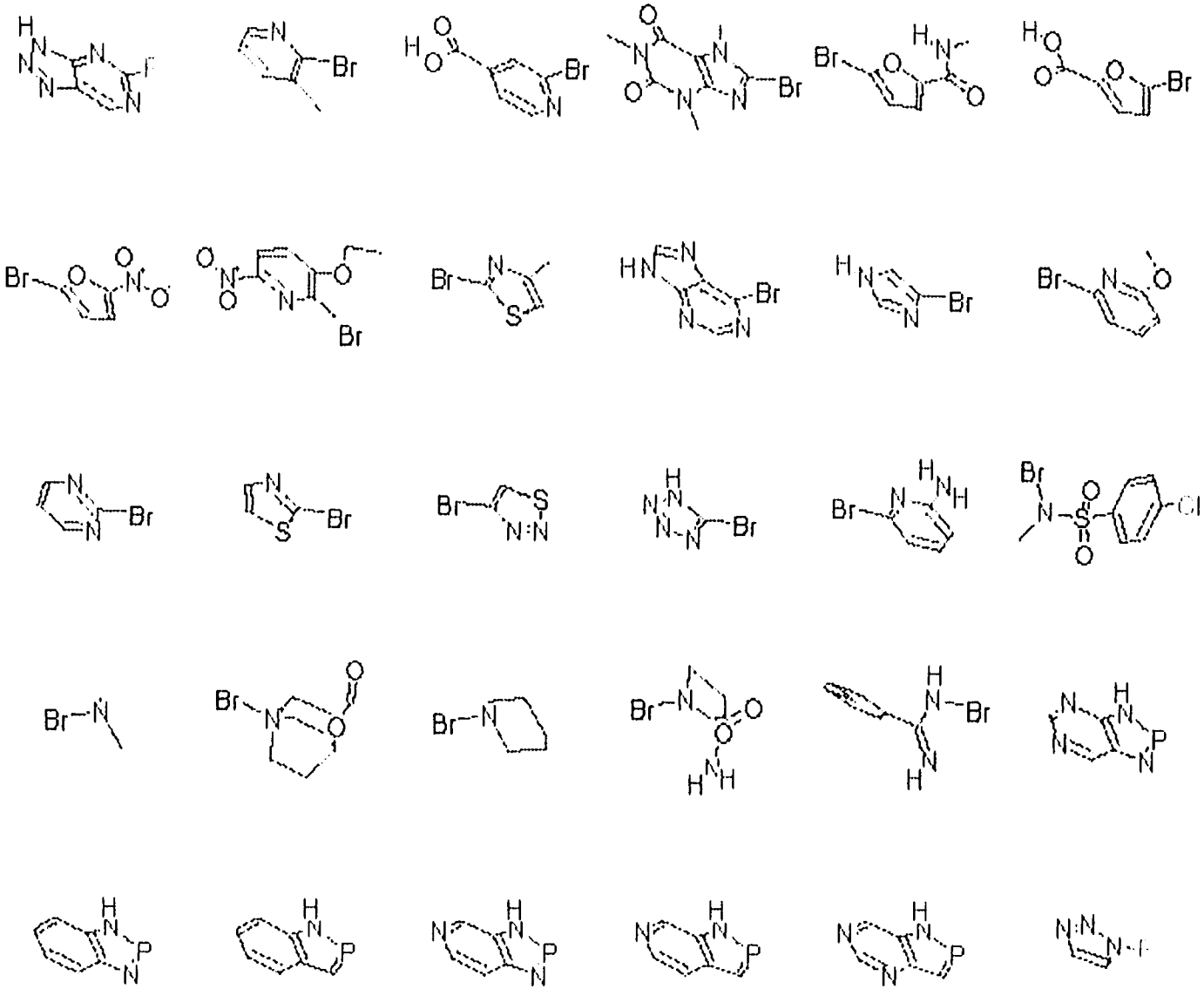


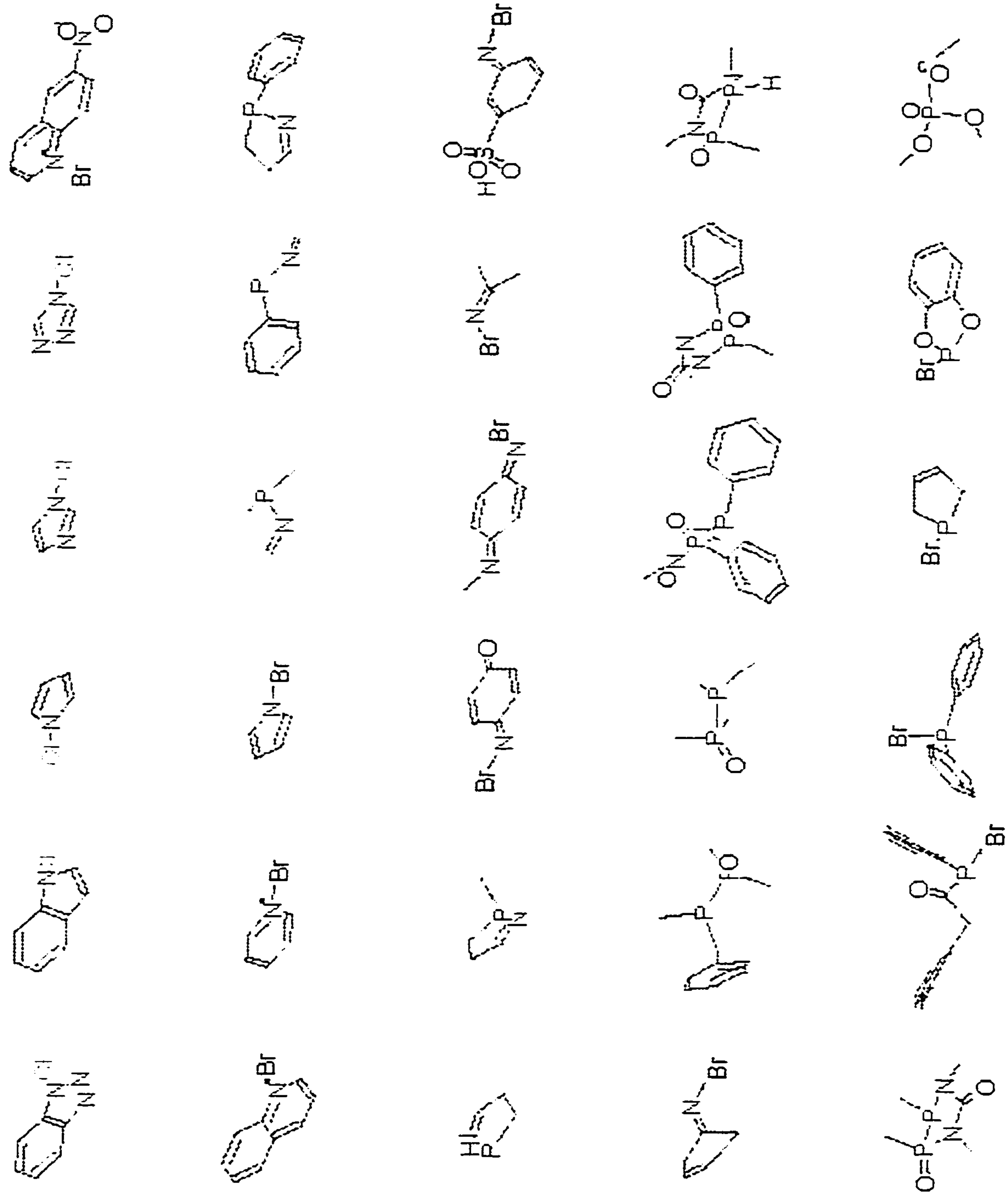


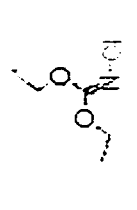
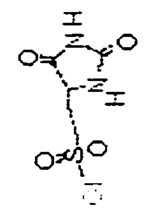
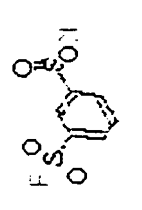
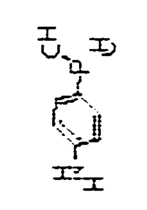
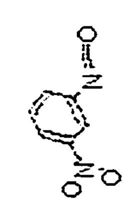
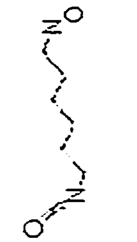
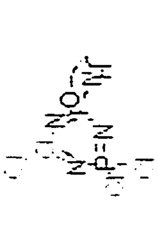
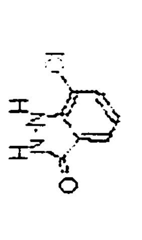
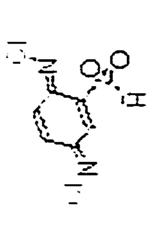
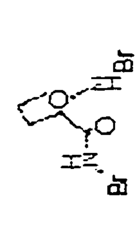
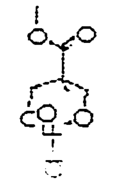
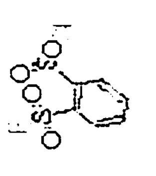
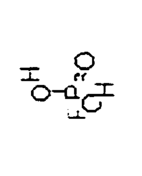
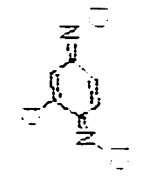
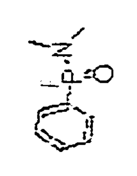
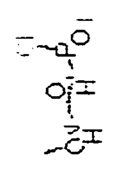
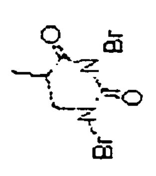
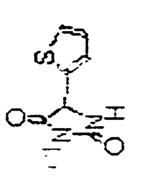
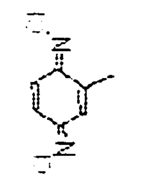
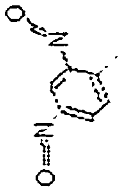
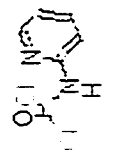
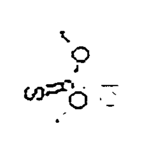
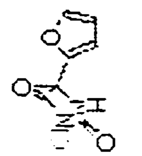
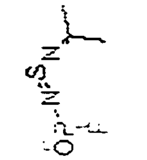
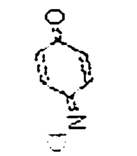
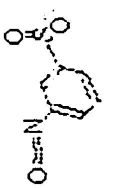
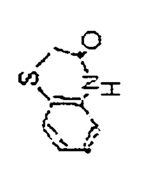
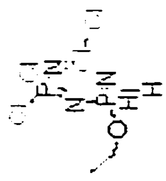
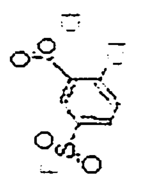
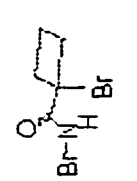


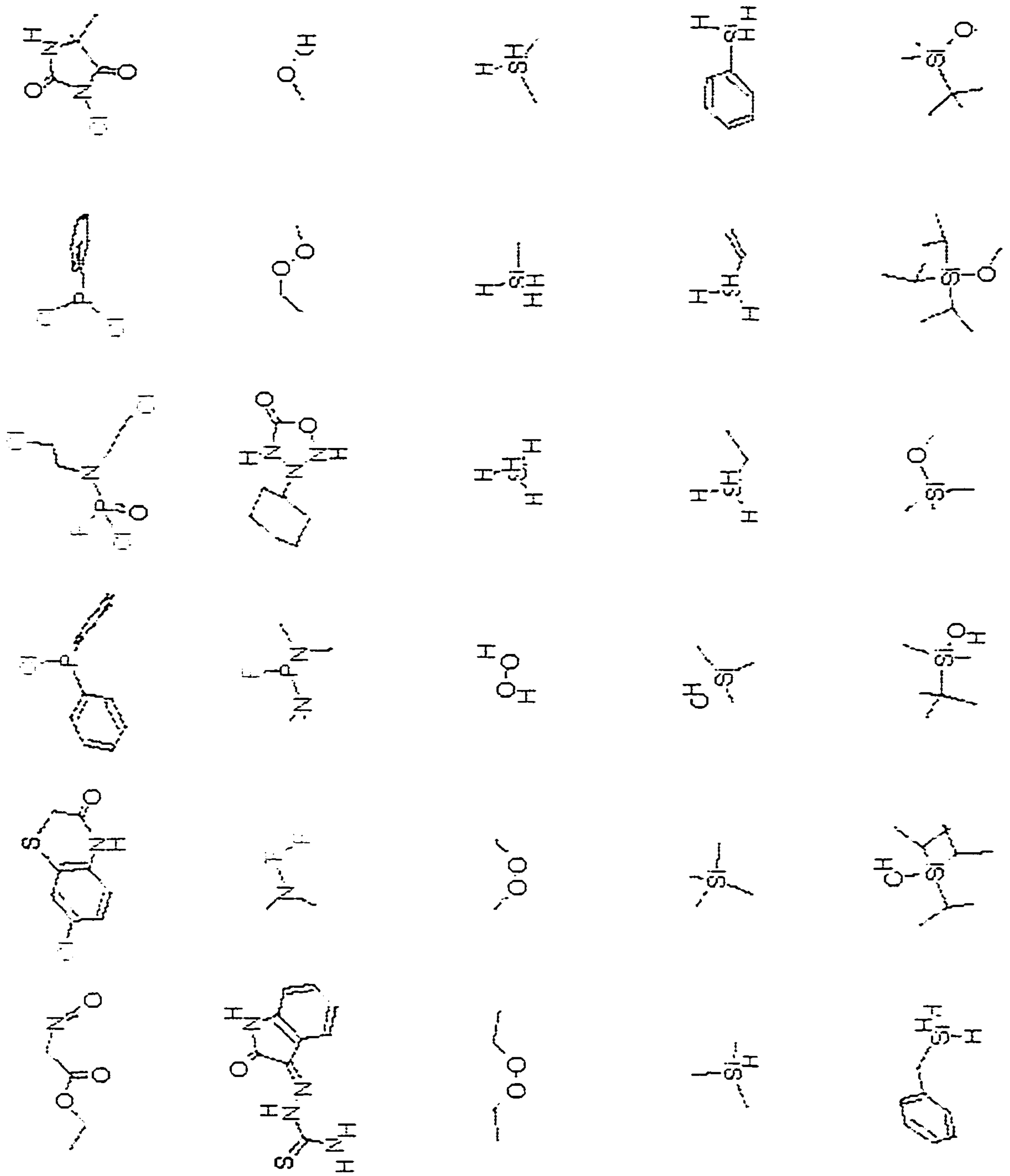










Appendix D

D.1 Charge Models

The standard two stage fitting protocol was used for RESP fit charges with the exception of the grid point density. The HF/6-31G* ESP was sampled on a face-centered-cubic grid of points with 0.5 Å spacing instead of the Connolly grid[74]. This required a re-calibration of the first stage restraint from 0.0005 to 0.001 and the second stage restraint from 0.001 to 0.01. A relativistic effective core potential with a valence space consisting of 4d, 5s, and 5p shells[75] was used on all iodine atoms in *ab initio* calculations containing this element.

AM1 atomic charges were calculated with the MOPAC-6 semi-empirical program with the following keywords in the input file: **AM1 GEO-OK MMOK EF CHARGE=*n*** where *n* is the net charge on the molecule. The same charges can be obtained from GAUSSIAN-94[76] using the **Iop(4/24=3)** keyword except for sulfur- or phosphorus-containing molecules which will have incompatible charges because GAUSSIAN-94 uses MNDO parameters for these atoms.

The AM1-BCC charges were obtained by adding bond charge corrections (BCCs) to the AM1 atomic charges using eqs. (2.2) and (2.3).

D.2 Dimer Energies

Monomers and dimers were energy minimized using the steepest descent method[77] for 10 cycles and then switched to the conjugate gradient method[77] until a DRMS of 0.1

kcal/mol Å was achieved. All dimer calculations were performed with the Cornell *et al.* force field as implemented in AMBER-5.01.[16] The MP2/6-31G*(.25) energies for the DNA dimers were taken from ref. **65**, the MP2/6-31+G** energies for the organic dimers were calculated using GAMESS[78] at the HF/6-31G* geometries.

D.3 Free Energy Calculations

All calculations were performed with the Cornell *et al.* force field as implemented in AMBER-5.01. Solute molecules were solvated in TIP3P waters with 12 Å separating the solute from the edges of the box. The solvated systems were then minimized using the steepest descent method for 10 cycles and then switched to the conjugate gradient method until a DRMS of 0.1 kcal/mol Å was achieved. The systems were then subjected to 20 ps. of molecular dynamics at 1 K in order to diffuse any buildup of potential energy that may have occurred during the placement of the waters. The next stage consisted of 20 ps. of heating to 300 K followed by 40 ps. of equilibration at the target temperature. The vdW + electrostatic perturbations were performed once for every system using the slow growth method for a total trajectory of 200 ps. Perturbations from one charge model to another were performed using the Windows approach with 400 steps of equilibration and 600 steps of data collection per window for 21 windows totaling 42 ps. The reported perturbation energies are the mean values of the forward and reverse directions. All simulations were run in the NPT ensemble at 1 atm using a constant dielectric of 1 with a 10 Å non-bonded cutoff, 2 fs. time step, and SHAKE applied to all bonds containing hydrogen atoms except those involved in the perturbations. The potential of mean force (PMF) was calculated for perturbed bonds. Separate scaling factors were used for solvent

and solute atoms with heat bath coupling constants of 0.2 ps. while in the gas phase a coupling constant of 0.4 ps. was used. The pressure relaxation time was set to 0.2 ps. The 1-4 non-bonded vdW and electrostatic interactions were scaled by ½ and 1/1.2, respectively.

Appendix E

E.1 Correlation Coefficient

Correlation coefficients are used to determine the relationship between two properties. The equation for the correlation coefficient is:

$$CC_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i and y_i are the i'th data points of properties x and y, $\bar{x}$ and $\bar{y}$ are the means of x and y, and n is the number of data points in x and y. The value of CC can lie between -1 and 1 inclusive. A value of 1 indicates a complete positive correlation while a value of zero indicates no correlation, and a value of -1 indicates a complete negative correlation.

Appendix F

F.1 Atom-type Codes

C_4	11	$N_{2^-,3,4^+}$	21	$O_{1,2}$	31	$P_{2,3}$	41
$C_3^{=C}$	12	N_3^{deloc}	22	$O_1^{ester,acid}$	32	$P_{3,4}^{=}$	42
$C_3^{=N,P}$	13	N_3^{hdeloc}	23	O_1^{lact}	33		
$C_3^{=O,S}$	14	N_2	24				
$C_{1,2}$	15	$N_{1,2^+}$	25				
C_{ar}	16						
C_{ar}^{lp}	17						
$S_{1,2}$	51	Si_4	61	F_1	71	H_1	91
S_3	52			Cl_1	72		
S_4	53			Br_1	73		
				I_1	74		

F.2 Bond Orders Codes

——	01	+—-	06	——-	09
═══	02	<u>arom</u>	07	<u>deloc</u>	09
≡≡≡	03	<u>arom</u>	08		

References

1. W. J. Hehre, L. Radom, P. v.R. Schleyer, and J. A. Pople, *Ab Initio Molecular Orbital Theory*, 1986, John Wiley & Sons, New York.
2. A. R. Leach, *Molecular Modelling, Principles and Applications*, 1996, Longman, Singapore.
3. D. Feller, and E. R. Davidson, *Chem. Rev.*, **86**, 681 (1986).
4. D. Feller, and E. R. Davidson, *Reviews in Computational Chemistry*, pp. 1-43,Editors: K. B, Lipkowitz and D. B. Boyd, 1990, VCH, New York.
5. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, **99**, 4899 (1977).
6. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, **99**, 4907 (1977).
7. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, **107**, 3902 (1985).
8. J. J. P. Stewart, *J. Comput. Chem.*, **10**, 209 (1989).
9. J. J. P. Stewart, *J. Comput. Chem.*, **10**, 221 (1989).
10. J. A. Pople and G. A. Segal, *J. Chem. Phys.*, **43**, S136 (1965).
11. D. B. Boyd and K. B. Lipkowitz, *J. Chem. Educ.*, **59**, 269 (1982).
12. N. L. Allinger and J. –H. Lii, *J. Comput. Chem.*, **8**, 1146 (1987).
13. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, **4**, 187 (1983).
14. S. L. Mayo, B. D. Olafson and W. A. Goddard III, *J. Phys. Chem.*, **94**, 8897 (1990).
15. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, **117**, 5179 (1995).

16. D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham III, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, D. M. Ferguson, R. J. Radmer, G. L. Siebel, U. C. Singh, P. K. Weiner and P. A. Kollman, AMBER 5, University of California, San Francisco, 1997.

17. U. Burkert and N. L. Allinger, *Molecular Mechanics, ACS Monograph 177*, 1982, American Chemical Society, Washington, D.C.

18. F. London, *Zeitschrift für Physik*, **63**, 245 (1930).

19. W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, **110**, 1657 (1988).

20. a) F. Momany, *J. Phys. Chem.*, **82**, 592 (1978); b) S. R. Cox and D. E. Williams, *J. Comput. Chem.*, **2**, 304 (1981); c) K. Hinsen and B. Roux, *J. Comput. Chem.*, **18**, 368 (1997); d) Z. Su, *J. Comput. Chem.*, **14**, 1036 (1993); e) D. S. Marynick, *J. Comput. Chem.*, **18**, 955 (1997); f) C. Chipot, B. Maigret, J.-L. Rivail and H. A. Scheraga, *J. Phys. Chem.*, **96**, 10276 (1992); g) R. J. Woods, M. Khalil, W. Pell, S. H. Moffat and V. H. Smith Jr., *J. Comput. Chem.*, **11**, 297 (1990).

21. a) J. Gasteiger and M. Marsili, *Tetrahedron*, **36**, 3219 (1980); b) K. Tai No, J. A. Grant and H. A. Scheraga, *J. Phys. Chem.*, **94**, 4732 (1990); c) J. A. Grant, R. L. Williams and H. A. Scheraga, *Biopolymers*, **30**, 929 (1990); d) A. K. Rappé and W. A. Goddard III, *J. Phys. Chem.*, **95**, 3358 (1991); e) W. J. Mortier, K. Van Genechten and J. Gasteiger, *J. Am. Chem. Soc.*, **107**, 829 (1985); f) U. Dinur, *J. Phys. Chem.*, **97**, 7894 (1993).

22. a) A. J. Stone, *Chem. Phys. Lett.*, **83**, 233 (1981); b) A. J. Stone and M. Alderton, *Mol. Phys.*, **56**, 1047 (1985); c) E. Sigfridsson and U. Ryde, *J. Comput. Chem.*, **19**, 377 (1998); d) U. Koch and E. Egert, *J. Comput. Chem.*, **16**, 937 (1995); e) C. Chipot,

J. G. Ángyán, G. G. Ferenczy and H. A. Scheraga, *J. Phys. Chem.*, **97**, 6628 (1993).

23. U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, **5**, 129 (1984).

24. L. Kuyper, D. Ashton, K. M. Merz Jr., and P. A. Kollman, *J. Phys. Chem.*, **95**, 6661 (1991).

25. The state of a system of N particles can be described by $6N$ values corresponding to the $3N$ coordinates for each atom and the $3N$ components of the momentum. The combination of the coordinate and momentum spaces defines the phase space.

26. The ergodic hypothesis states that a time averaged observation made on a single system has the same statistical properties as observations made at the same time from an ensemble of systems.

27. W. Jorgensen, C. Ravimohan, *J. Chem. Phys.*, **83**, 3050 (1985).

28. P. A. Kollman, *Chem. Rev.*, **93**, 2395 (1993).

29. Editors W. F. van Gunsteren and P. K. Weiner, *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*, 1989, ESCOM, Leiden.

30. K. Watanabe and M. L. Klein, *Chem. Phys.*, **131**, 157 (1989).

31. H. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren and J. Hermans, In *Intermolecular Forces*, B. Pullman, Ed., Elsevier, Dordrecht, Netherlands, 1981, p 331.

32. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, **79**, 926 (1983).

33. I. M. Svishchev, P. G. Kusalik, J. Wang, and R. J. Boyd, *J. Chem. Phys.*, **105**, 4742 (1996).

34. R. F. W. Bader, *Chem. Rev.*, **91**, 893 (1991).

35. R. F. W. Bader, *Acc. Chem. Res.*, **18**, 9 (1985).

36. K. B. Wiberg and P. R. Rablen, *J. Comput. Chem.*, **14**, 1504 (1993).

37. B. H. Besler, K. M. Merz Jr. and P. A. Kollman, *J. Comput. Chem.*, **11**, 431 (1990).

38. U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, **5**, 129 (1984).

39. C. M. Breneman and K. B. Wiberg, *J. Comput. Chem.*, **11**, 361 (1990).

40. a) D. E. Williams, *Biopolymers*, **29**, 1367 (1990); b) C. A. Reynolds, J. W. Essex and W. G. Richards, *J. Am. Chem. Soc.*, **114**, 9075 (1992).

41. C. I. Bayly, P. Cieplak, W. D. Cornell and P. A. Kollman, *J. Phys. Chem.*, **97**, 10269 (1993).

42. M. M. Francl, C. Carey, L. E. Chirlian and D. M. Gange, *J. Comput. Chem.*, **17**, 367 (1996).

43. C. Alemán, F. J. Luque and M. Orozco, *J. Comput. Chem.*, **14**, 799 (1993).

44. T. A. Halgren, *J. Comput. Chem.*, **17**, 490 (1996).

45. T. A. Halgren, *J. Comput. Chem.*, **17**, 520 (1996).

46. B. L. Bush, C. I. Bayly and T. A. Halgren, *J. Comput. Chem.*, **20**, 1495 (1999).

47. J. W. Storer, D. J. Giesen, C. J. Cramer and D. G. Truhlar, *J. Comput. Aid. Mol. Des.*, **9**, 87 (1995).

48. J. Li, T. Zhu, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. A*, **102**, 1820 (1998).

49. E. M. Duffy and W. L. Jorgensen, *J. Am. Chem. Soc.*, **122**, 2878 (2000).

50. G. A. Kaminski, W. L. Jorgensen, *J. Phys. Chem. B*, **102**, 1787 (1998).

51. The AM1 atomic charges are the standard charges produced by default by MOPAC-6 and are derived from the Coulson density matrix, see appendix A (and see page 4-8,

notes 18 and 20 of MOPAC-6 manual).

52. http://www.schrodinger.com/Mopac/html/node444.html

53. A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly, *J. Comput. Chem.*, **21**, 132 (2000).

54. G. A. Landrum YAeHMOP: Yet Another extended Hückel Molecular Orbital Package. YAeHMOP is freely available on the WWW at URL: http://overlap.chem.cornell.edu:8080/yaehmop.html

55. U. C. Singh and P. A. Kollman, *J. Comput. Chem.*, **5**, 129 (1984).

56. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran 77, 2nd edition, The Art of Scientific Computing*, Cambridge University Press, New York, 1996.

57. D. E. Williams, In *Reviews In Computational Chemistry, Vol. 2*, K. B. Lipkowitz and D. B. Boyd, VCH Publishers, New York, 1991, p.239.

58. Editor S. Budavari, *The Merck Index, An Encyclopedia of Chemicals, Drugs, and biologicals, Twelfth Edition*, 1996, Merck Research Laboratories, Whitehouse Station, NJ.

59. http://cactus.cit.nih.gov

60. http://www.ccl.net

61. The CD-ROM version of The Merck Index, 12th Edition: is available.

62. The NCI database of ca. 250000 molecules was downloaded from http://cactus.cit.nih.gov/ncidb/download.html

63. R. W. Dixon and P. A. Kollman, *J. Comput. Chem.*, **18**, 1632 (1997).

64. B. L. Bush and R. P. J. Sheridan, *Chem. Inf. Comput. Sci.*, **33**, 756 (1993).

65. Šponer, J.; Leszczynski, J.; Hobza, P. *J. Phys. Chem.* 1996, 100, 1965.

66. http://www.ccl.net

67. N-methyl-N'-nitro-N-nitrosoguanine.

68. P. R. Gerber, *J. Comput.-Aided Mol. Design*, **12**, 37 (1998).

69. C. J. Cramer and D. G. Truhlar, *J. Comput.-Aided Mol. Design*, **6**, 629 (1992).

70. R. C. Rizzo and W. L. Jorgensen, *J. Am. Chem. Soc.*, **121**, 4827 (1999).

71. E. M. Duffy and W. L. Jorgensen, *J. Am. Chem. Soc.*, **122**, 2878 (2000).

72. E. C. Meng, P. Cieplak, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, **116**, 12061 (1994).

73. C. Chipot, B. Maigret, D. A. Pearlman and P. A. Kollman, *J. Am. Chem. Soc.*, **118**, 2998 (1996).

74. M. L. Connolly, *J. Appl. Cryst.*, **16**, 548 (1983).

75. LaJohn, L. A.; Christiansen, P. A.; Ross, R. B.; Atashroo, T.; Ermler, W. C. J Chem Phys 1987, 87, 2812.

76. M. J. Frisch, G. W. Trucks, H. B. Schlegel, P. M. W. Gill, B. G. Johnson, M. A. Robb, J. R. Cheeseman, T. A. Keith, G. A. Petersson, J. A. Montgomery, K. Raghavachari, M. A. Al-Laham, V. G. Zakrzewski, J. V. Ortiz, J. B. Foresman, J. Cioslowski, B. B. Stefanov, A. Nanayakkara, M. Challacombe, C. Y. Peng, P. Y. Ayala, W. Chen, M. W. Wong, J. L. Andres, E. S. Replogle, R. Gomperts, R. L. Martin, D. J. Fox, J. S. Binkley, D. J. Defrees, J. Baker, J. P. Stewart, M. Head-Gordon, C. Gonzalez and J. A. Pople, *Gaussian94*, Gaussian, Inc., Pittsburgh PA, 1995.

77. J. A. McCammon and S. C. Harvey, *Dynamics of Proteins and Nucleic Acids*, 1989,

Cambridge University Press, Cambridge.

78. M. W. Schmidt, K .K. Baldridge, J .A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K .A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.*, **14**, 1347 (1993).