# Chapter 2: Computational Methodology

In this chapter the fundamentals of the methodologies used in this thesis are described.

## Computational Carbohydrate Grafting

### Generation of 3D glycan libraries

A 3D virtual glycan library (GLibrary-3D) was built based on structures available from the online database GlycomeDB ([www.glycome-db.org](http://www.glycome-db.org)) [1]. In 2011 GlycomeDB contained 3,570 N- and O-linked glycan sequences found in humans, of which 3,086 contained sufficient information to be converted into 3D structures. Because most of the structures contained in GlycomeDB were determined using mass spectrometric techniques, not all reported sequences included sufficient information to uniquely define the glycan. For example, many sequences do not include information regarding inter-residue linkage positions, and these were generally excluded from the virtual library. However, in the case of certain human glycan sequences, which display only a limited number of linkage possibilities, such as the disaccharides Neu5Acα(2-3/6)Gal or Galβ(1-3/4)GlcNAc, each linkage permutation was constructed. Additionally, on the basis of known glycan structures, all ring types were assumed to be pyranose. These assumptions resulted in a total of 7,127 unique putative human glycan structures. For glycans containing 1-6 or 2-6 linkages, each stable rotamer of the ω-angle (+/-60°, 180°) was generated. Additional rotamers were built for the ϕ-angles in 2-3 linkages (-60° and 180°) [2], leading to a library of 207,693 glycan 3D structures (GLibrary-3D). Glycan sequences were converted to 3D structures using an automated version of the Carbohydrate Model Building Tool of GLYCAM-Web ([www.glycam.org](http://www.glycam.org)) [3].

A second, high-definition version of the 3D library is also available, where only glycan sequences with fully defined inter-glycosidic linkages, but that may have an undefined reducing-terminal anomeric configuration, are included. There are 2,799 of these more fully defined glycan sequences, whose various rotamer and anomeric configuration permutations are modelled using 60,925 3D structures (GLibrary-3DHD).

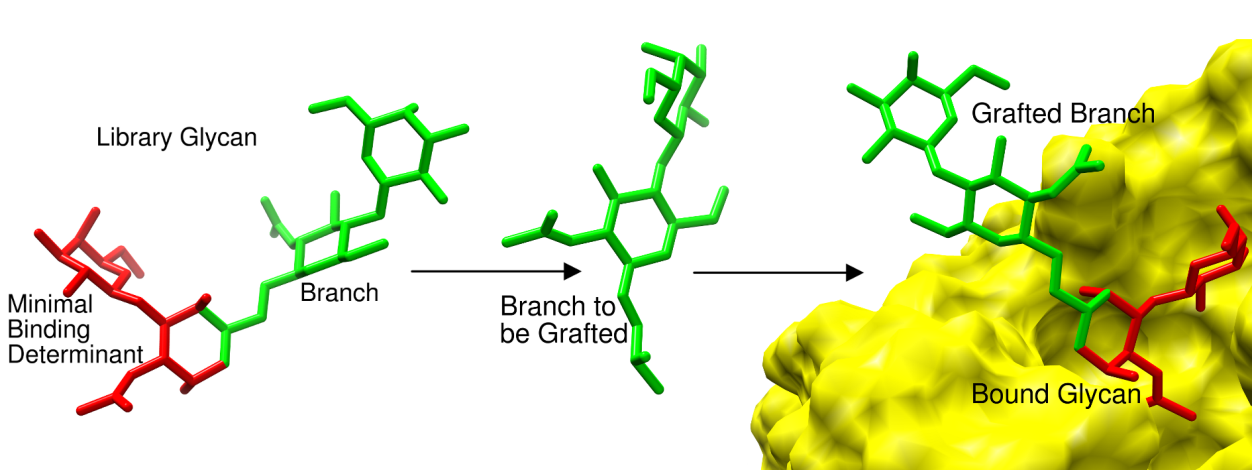
Following the same procedures separate libraries for each of the currently buildable glycans present on the most recent (since version 4.0) Consortium for Functional Glycomics glycan arrays.

### Library searching

During grafting, glycans which contain the minimal binding determinant (MBD) of the GBP are searched for within the library. The MBD is the portion of a glycan that directly interacts with the GBP, and is primarily responsible for the affinity of the interaction. Additional branches from this MBD core may inhibit binding but are not essential for a detectable high affinity interaction.

### Superimposing and Wiggling

Unlike traditional molecular docking, which attempts to generate poses for the entire glycan in the binding site, CCG splices the glycan branches onto the appropriate positions in the bound MBD of the glycan present in a protein-carbohydrate 3D complex. During grafting, the coordinates for any branches attached to the MBD are superimposed onto the appropriate linkages of the co-complexed MBD (Figure 2.1).



**Figure 2.1:** Illustration of the grafting process.The branches from the minimal binding determinant containing glycans in the 3D library are excised and spliced onto the bound minimal determinant. The grafted branches are then assessed for steric clashes with the GBP surface. This process is illustrated for the grafting of the glycan branch Galβ1-4GlcNAcβ1-6 (green, stick) onto the TF-antigen (red, stick) in the JAA-F11 binding site (yellow solvent-accessible surface).

In order to relieve minor clashes with the protein surface, the glycosidic linkages of a grafted branch are permitted an amount (≤ ± 20 °) of rotational motion to reflect their inherent flexibility [4].

### Predicting Binders and Non-binders

Quantification of any remaining steric overlaps between the grafted branches and the protein surface enables discrimination between putative binding partners and putative non-binders. Any vdW overlaps were quantified using Equation 2.1, where *A* is area of overlap (Å2), *r* is the assigned radius (Å), and *d* is the actual distance between the atoms (Å). Bondi radii with the Rowland and Taylor [5] modifications were assigned. If the sum of any two radii was greater than the distance between them, an overlap *A*, was computed (Equation 2.1, [6]) and the individual overlaps between each atom in the glycan and the receptor were summed.

[2.1]

The overlap, , was then normalized relative to the surface area of a buried carbon atom (36.32 Å2) giving an intuitive number called relative overlap (RO) that relates to the number of atoms buried within the protein surface. A glycan was predicted to be a non-binder if its total RO was greater than one.

Currently, no attempt is made to rank the structures in terms of theoretical affinities; the power of the CCG method is the rapid identification of a subset of putative binders, which can subsequently be examined in more detail either experimentally or theoretically.

### CFG Glycan Array Linkers

In addition to checking if the branches are physically tolerated in the binding site, CCG also checks if the linker used to conjugate the glycan to the array surface is physically tolerated in the binding site and whether or not it is able to present the glycan relative to the surface to enable the GBP to recognize it. If both the linker and the branches of a glycan from the 3D library permit binding then the glycan is predicted to be a binder and can be compared to the available experimental glycan array screening data.

By grafting onto a bound carbohydrate motif, a level of speed and accuracy in the prediction of the 3D structures is achieved that would otherwise be impossible using either traditional virtual screening or experimental techniques alone. In addition, CCG facilitates the screening of vast libraries of glycans that can encompass the entire known human glycome, as well as synthetic or hypothetical structures, extending the CCG screening capability far beyond the scope of current experimental glycan microarrays.

## Molecular Mechanical Methods

In molecular mechanics (MM) or “force field” type methods, atoms are described using a “classical” representation. This approximation allows the study of large biomolecular systems or phenomena that operate on relatively long time scales. In the classical representation, bonded interactions such as bond stretching and angle bending, are modelled using simple harmonic functions such as Hooke's (spring) law. This precludes the study of any phenomena that are based upon electron movement, such as bond making or breaking. Torsion angles are modelled using Fourier expansion functions. Non-bonded interactions (electrostatic and van der Waals) are modelled using Coulombic and Lennard-Jones potentials, respectively. The contributions from each of the bonded and non-bonded interactions are summed (Equation 2.2, [7] ) to calculate the Potential Energy *V(rN­)*.

[2.2]

Both non-bonded and bonded terms are developed from small molecular fragments and then combined together additively when applied to larger molecules. The force fields are designed to be transferable within a particular set of related molecules, such as within different proteins, lipids, or carbohydrates. It is now common to develop the individual force fields in a consistent manner so that simulations can be performed using a mixture of molecular classes. Despite the seemingly crude approximations, molecular mechanics force fields can be just as accurate as, and are much faster than, high level quantum mechanics calculations when reproducing certain experimental observables, such as structures and relative energies. However this approach is heavily dependent on the care and philosophy adopted during force field parameterization.

## Molecular Dynamics

Physicist Richard Feynman famously said: "life is nothing but the wiggling and jiggling of atoms" [8]. This is a profoundly important observation; knowledge of the microscopic states of a system allows direct extrapolation to macroscopic properties observed in the laboratory. Molecular dynamics is essentially an attempt to model the time-dependant motion and interactions of atoms and molecules. A classical force field is used to calculate the potential energy for any particular conformation and configuration of a molecular system via Equation 2.2. The forces acting on each individual atom are calculated from the gradient of the potential energy using Equation 2.3 [7].

[2.3]

Atomic accelerations and therefore motion can be calculated using Newton's second law, which states that a body with mass *m*, being acted upon by a force *F*, experiences acceleration *a* in the same direction as the force and according to equation 2.3.

The force *F* acting on each particle is calculated from the derivative of the potential energy returned from the force field. The acceleration *a*, calculated from the force, is used along with the current velocities and atomic positions of the system to calculate how the positions of each particle will change over time using the formula in Equation 2.4 [7].

[2.4]

Given the atomic accelerations, it is then possible to determine the position of each atom after a short timestep, *Δ*t. The accelerations of the particles are calculated at a particular time point *t* and when combined with the current positions and velocities predict new positions and velocities at a time *t + Δt*. During this time step the force acting on each particle is assumed to be constant. There are many algorithms used but all assume that the positions and dynamical properties can be approximated as Taylor series expansions.

The positions at a new step, *r(t + Δt),* or a previous step, *r(t + Δt),* can be calculated using equations 2.5 and 2.6, respectively.

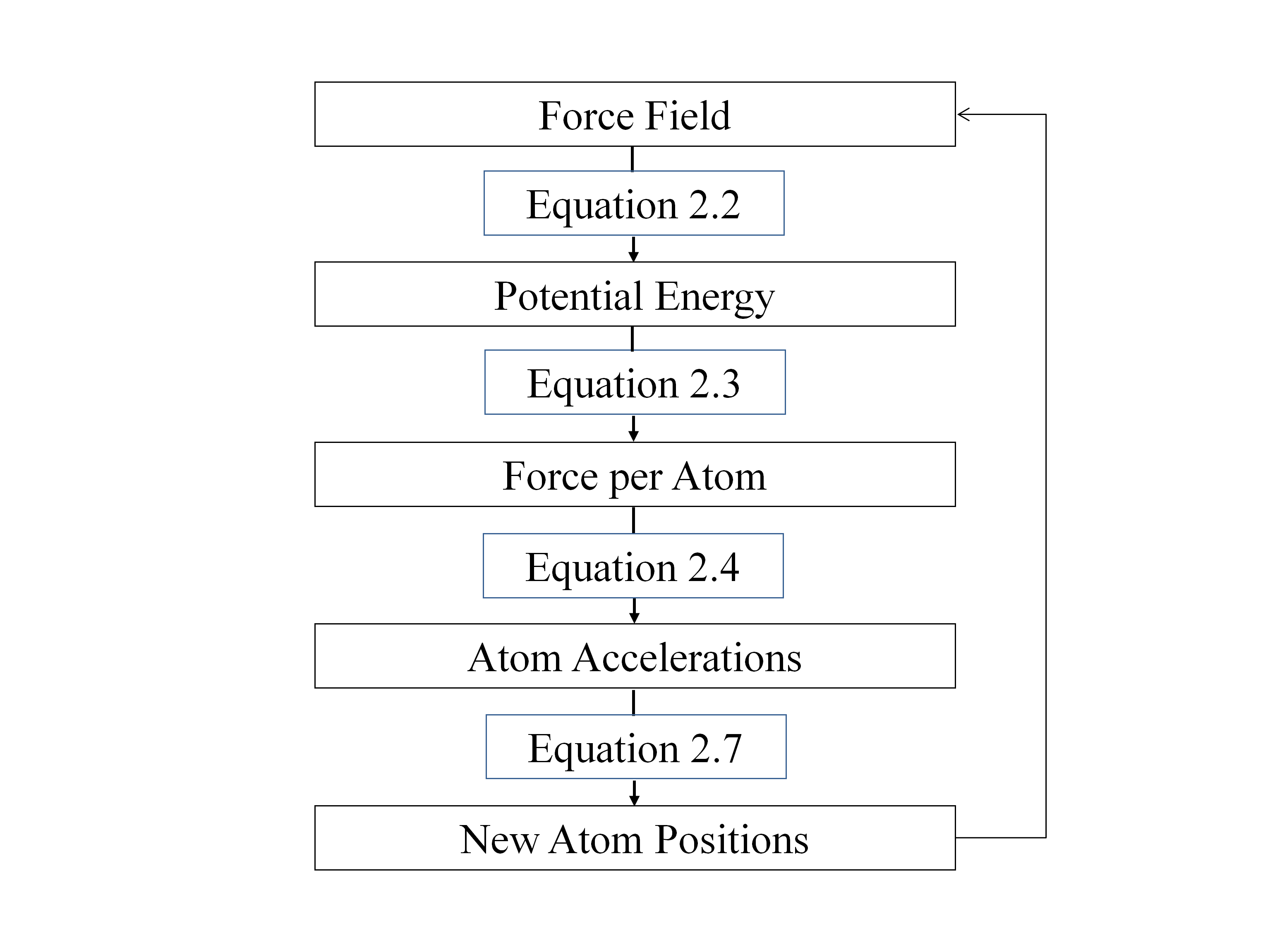
[2.5]

[2.6]

The Verlet algorithm [9] is a combination of these two equations, where the velocity terms are cancelled. This allows just the positions *r,* and accelerations *a,* at time *t*, along with the positions from the previous step *r(t - Δt),* to calculate the positions at the next step, *r(t + Δt)*, using Equation 2.7.

[2.7]

For the first step in a molecular dynamics calculation, velocities are assigned to the particles according to a Maxwell-Boltzmann distribution at the temperature of interest. The time span or “step size” used is a crucial determinant in how long the simulation will take to complete, but the maximum size of the time step depends on the time order of the fastest motion within the system. In biological systems this is the vibration of a covalent bond between hydrogen and a heavy atom and is on the order of 0.5 fs. However, most MD simulations will constrain this motion allowing much larger time steps of up to 2 fs. Choosing a time step that is too small is time inefficient but choosing one that is too large leads to instabilities as particles move too much in between the force calculation step, resulting in large vdW overlaps or electronic repulsions.



**Schema 2.1:** Outline of the calculations involved in a single step of a MD simulation.

An initial configuration of the system is needed in order to start a molecular dynamics calculation. This is commonly taken from experimental data such as 3D structures, generated by NMR or X-ray crystallography. Theoretical models or a combination of theoretical and experimental models can also be used. For simulations involving biomolecules the system is placed in a box of water and appropriately charged counter-ions are added to ensure the system is neutral. The positions of the solvent molecules are then energy minimized to allow it to adjust and reorient relative to the artificially introduced solute. Molecular dynamics begins with an equilibration phase, where velocities are assigned and the system is brought to the correct temperature and pressure. Velocities, energies and positions are saved at regular intervals allowing calculation of average properties or visualization of the resulting trajectory. The measure for sufficient equilibration is when properties such as temperature and pressure fluctuate around the desired value and the total energy of the system remains constant. Structural equilibration can also be important depending on what is being studied. Once the system is equilibrated a “production” phase begins during which the properties of interest can be monitored.

## Developing Partial Charges

The distribution of electrons within a molecule generates an electrostatic potential which affects the molecules interaction with its surrounding electronic environment. This molecular electron density is approximated in MM using point charges centred on each atom [10]. In order to model molecules not already covered by the AMBER set of force fields, molecular mechanics type point charges for atoms may be calculated using *ab initio* quantum mechanical calculations. The QM electrostatic potential (*Vi*) for a point (*i*) at a position (*ri*) is given by Equation 2.8:

[2.8]

If *qj* is the net atomic charge on atom *j* then the MM electrostatic potentialat a point *i* can be expressed using Equation 2.9.

[2.9]

The net atomic charges are derived from a least-squares fitting of the MM electrostatic potentials to QM electrostatic potentials, by minimization of the function *Y* in Equation 2.10.

[2.10]

Using QM and MM potentials, a restrained electrostatic potential (RESP) fit was used to calculate the point charges in accordance with regular AMBER development protocols. An attempt is made to derive a set of partial charges that both reproduce the overall charge of the molecule and also reproduces the quantum mechanical electrostatic potential at each point. In developing charges for the AMBER force field an algorithm known as CHELPG was developed that places points on a grid surrounding the van der Waals surface [11]. This method was further modified to include RESP [12].

## References

1. Ranzinger R, Herget S, Wetter T, von der Lieth CW (2008) GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinformatics* 9: 384.

2. Poppe L, Stuike-Prill R, Meyer B, van Halbeek H (1992) The solution conformation of sialyl-alpha (2----6)-lactose studied by modern NMR techniques and Monte Carlo simulations. *Journal of Biomolecular NMR* 2: 109-136.

3. WoodsGroup (2005 - 2012) GLYCAM Web. In: Woods RJ, editor. Athens, GA: *Complex Carbohydrate Research Center, The University of Georgia*.

4. DeMarco ML, Woods RJ (2008) Structural glycobiology: a game of snakes and ladders. *Glycobiology* 18: 426-440.

5. Rowland RS, Taylor R (1996) Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der Waals Radii. *The Journal of Physical Chemistry* 100: 7384-7391.

6. Rychkov G, Petukhov M (2007) Joint neighbors approximation of macromolecular solvent accessible surface area. *Journal of Computational Chemistry* 28: 1974-1989.

7. Leach AR (2001) Molecular Modelling: Principles and Applications second edition: *Prentice Hall*.

8. Feynman R (1963) Lectures on Physics

9. Verlet L (1967) Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* 159: 98-103.

10. Woods RJ, Khalil M, Pell W, Moffat SH, Smith VH (1990) Derivation of net atomic charges from molecular electrostatic potentials. *Journal of Computational Chemistry* 11: 297-310.

11. Breneman CM, Wiberg KB (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry* 11: 361-373.

12. Bayly CI, Cieplak P, Cornell W, Kollman PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* 97: 10269-10280.