Scientific Research Publishing

Journal of Biosciences and Medicines, 2026, 14(6), 464-484
https://www.scirp.org/journal/jbm
ISSN Online: 2327-509X
ISSN Print: 2327-5081

On Docking and Binding

Yi Fang

School of Mathematics, Jilin University, Changchun, China
Email: yi.fang3@gmail.com

How to cite this paper: Fang, Y. (2026) On Docking and Binding. *Journal of Biosciences and Medicines*, **14**, 464-484.
https://doi.org/10.4236/jbm.2026.146031

Received: May 24, 2026
Accepted: June 27, 2026
Published: June 30, 2026

Abstract

The single molecule conformational Gibbs free energy function (CGF) is rigorously derived via quantum statistics. Apply it as scoring function in docking we obtain analytic docking Gibbs free energy formula. Line by line estimates of the formula establishes the necessary and sufficient conditions to make the docking Gibbs free energy as negative as possible. These conditions form the binding pose searching strategy: the binding sites, pieces on surfaces of receptor and ligand, must be close enough, both geometrically and electronically complementary. Electronically complementary has two cases: either both binding sites are large hydrophobic pieces with smaller hydrophilic pieces or large hydrophilic pieces with smaller hydrophobic pieces. A single molecule theory of molecular folding based on CGF and single molecule thermodynamic hypothesis (SMTH) reveals that non-covalent and covalent binding will induce folding and conformational change, called post-binding reshaping. Applying this theory we resolve all three major unresolved problems in docking posted in 2017, 1. Tackling binding site flexibility; 2. Treating solvent during docking; 3. Affinity prediction in docking. Especially post-binding reshaping resolves the mystery of significant conformational change on covalent binding.

Keywords

Docking, Binding, Conformational Gibbs Free Energy Function, Post-Binding Reshaping, Binding Affinity, Hydrophobic, Hydrophilic

1. Introduction

In 1982, Kuntz *et al.* published a paper [1] entitled "A geometric approach to macromolecule-ligand interactions." That is the beginning of docking, or molecular docking, a wide spread "computational method that virtually tries to predict a complex of (usually) two binding partners. Typically, these binding partners are biological macromolecules (e.g., protein, DNA/RNA, peptide) or small molecules (e.g., endogenous ligands, drugs)." [2].

Since then, docking has become so important that each year hundred thousands papers published on docking. Indeed, Figure 1 of [3] shows that in between the years 2000 to 2020, more than 150,000 papers on "docking" and/or "molecular recognition" topics were published.

Docking becomes more and more important in structure-based drug design. Today, there are many docking web servers available, such as HADDOCK and HDOCK for protein-protein docking; GOLD, Surflex-Dock, AutoDock, and Glide, etc., for small-molecule which "are regularly utilized in structure-based drug design to predict ligand interactions with the receptor protein." [2].

In docking, a macromolecule (protein, DNA/RNA, peptide) is called receptor, another molecule (protein, DNA/RNA, peptide, or small molecule) is called ligand, [4]. In drug design, the ligand is usually a small molecule.

Binding is to put the receptor and ligand together to form either a new molecule (covalent binding) or just a functioning complex (non-covalent binding). Most docking web servers only deal with non-covalent binding. We will deal mainly with non-covalent binding and only briefly discuss covalent binding. It turns out that except chemistry considerations, covalent binding is actually simpler than non-covalent binding.

"Docking methods generally have two components: a scoring function and a sampling procedure. The scoring function is given a hypothetical binding pose for the compound and estimates the binding energy of the compound assuming that this pose is correct. The sampling procedure searches the space of potential poses to discover the pose that is assigned the most favorable score by the scoring function. The pose with the most favorable score is the predicted pose, and the score of this pose is the predicted binding energy." [5].

For non-covalent binding, the "sampling procedure" for poses is bring the ligand toward to the receptor and searching different orientations of the ligand structure to fit to the receptor structure as a pose. The process of finding poses is called the molecular recognition process.

We will use single molecule conformational Gibbs free energy function (CGF) as scoring function for docking. Since CGF has analytical formulae (4) and (8), we obtain analytic docking and binding Gibbs free energy formulae analysis of them and reveals the binding pose searching strategy: the binding sites must be close enough, both geometrically and electronically complementary. See Search 3.1 for detail.

"Molecular docking enhances the efficacy of determining the metabolic interaction between two molecules, *i.e.*, the small molecule (ligand) and the target molecule (protein), to find the best orientation of a ligand to its target molecule with minimal free energy in forming a stable complex." [6].

Applying a single molecule molecular folding theory (see [7]-[11]) based on CGF and single molecule thermodynamic hypothesis (SMTH), we resolve all three major unresolved problems in docking posted in 2017 in [12], 1. Tackling binding site flexibility; 2. Treating solvent during docking; 3. Affinity prediction in

docking.

2. Single Molecule Conformational Gibbs Free Energy Function

2.1. Conformations of Molecules

Given a molecule $\mathfrak{U}$ consists of n atoms $(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_n)$, a conformation is the listed coordinates of atomic centres $\boldsymbol{R}=(\boldsymbol{r}_1,\cdots,\boldsymbol{r}_n)\in\mathbb{R}^{3n}$.

We assume that in a conformation $\boldsymbol{R}$, all covalent bonds are correctly formed. There are standard bond lengths b_{ij} of $\mathfrak{U}$ if $\boldsymbol{a}_i\boldsymbol{a}_j$ is covalently bonded and standard bond angles $\alpha_{ij,ik}$ of $\mathfrak{U}$ if both $\boldsymbol{a}_i\boldsymbol{a}_j$ and $\boldsymbol{a}_i\boldsymbol{a}_k$ are covalently bonded. There are small positive constants δ_{ij} and $\beta_{ij,ik}$ such that all eligible bond lengths are contained in the closed interval $[b_{ij}-\delta_{ij},b_{ij}+\delta_{ij}]$ and $b_{ij}+\delta_{ij}<r_i+r_j$, where r_i is the van der Waals radius of $\boldsymbol{a}_i$; and all eligible bond angles are contained in the close interval $[\alpha_{ij,ik}-\beta_{ij,ik},\alpha_{ij,ik}+\beta_{ij,ik}]$. These should be respected in any conformation of $\mathfrak{U}$. Thus, let $r_{ij}=|\boldsymbol{r}_i-\boldsymbol{r}_j|$ and $\gamma_{ij,ik}$ be the corresponding bond lengths and angles measured in $\boldsymbol{R}$, we require that they satisfy the steric conditions (1) to (3),

$$r_{ij}\in[b_{ij}-\delta_{ij},b_{ij}+\delta_{ij}],\quad \text{if } \boldsymbol{a}_i\boldsymbol{a}_j \text{ is covalently bonded,} \tag{1}$$

$$r_{ij}-\delta_{ij}\geq r_i+r_j \quad \text{if } \boldsymbol{a}_i\boldsymbol{a}_j \text{ is not covalently bonded,} \tag{2}$$

$$\gamma_{ij,ik}\in[\alpha_{ij,ik}-\beta_{ij,ik},\alpha_{ij,ik}+\beta_{ij,ik}] \quad \text{if } \boldsymbol{a}_i\boldsymbol{a}_j \text{ and } \boldsymbol{a}_i\boldsymbol{a}_k \text{ are covalently bonded.} \tag{3}$$

2.2. The Thermodynamic System $\mathfrak{S}_R$

Let $\boldsymbol{R}=(\boldsymbol{r}_1,\cdots,\boldsymbol{r}_n)$ be a conformation of $\mathfrak{U}$. The occupied space of $\boldsymbol{R}$, $V_{\boldsymbol{R}}\subset\mathbb{R}^3$ is determined by an electron wave function Φ in quantum mechanics. It is e defined as $V_{\boldsymbol{R}}=\{\boldsymbol{x}\in\mathbb{R}^3;|\Phi(\boldsymbol{x})|^2\geq\epsilon\}$, where $\epsilon>0$ is a small number. But so far multi-electrons Schrödinger equation is unsolvable, so that the Φ is not known. Therefore, we have to use some approximation to $V_{\boldsymbol{R}}$. Figures on page 4 of [13], where the boundary $\partial V_{\boldsymbol{R}}=\{\boldsymbol{x}\in\mathbb{R}^3;|\Phi(\boldsymbol{x})|^2=\epsilon\}$, $\epsilon=0.001$ au, 1 au = a_0 = 0.529188Å, shows that $V_{\boldsymbol{R}}$ is very similar to a bunch of overlapping balls. An atom $\boldsymbol{a}_i$ occupies a space similar to a ball $B(\boldsymbol{r}_i,r_i)$, centred at the atomic centre $\boldsymbol{r}_i$ with its van der Waals radius r_i, the bunch of overlapping balls $P_{\boldsymbol{R}}=\bigcup_{i=1}^n B(\boldsymbol{r}_i,r_i)$ is a very close approximation to $V_{\boldsymbol{R}}$.

A molecule always exists in some environment, most biomolecules live in cytoplasm that is essentially aquoeous solvent or water environment, denoted as $\mathfrak{W}$. The thermodynamics system $\mathfrak{S}_{\boldsymbol{R}}\subset\mathbb{R}^3$ is tailor made for a conformation $\boldsymbol{R}$, it consists of $V_{\boldsymbol{R}}$ plus one layer of water molecules. At equilibrium, $\mathfrak{S}_{\boldsymbol{R}}$ has a Gibbs free energy $G(\mathfrak{S}_{\boldsymbol{R}})$. The single molecule conformational Gibbs free energy function (CGF) is defined as $G(\boldsymbol{R};\mathfrak{U},\mathfrak{W})=G(\mathfrak{S}_{\boldsymbol{R}})$.

2.3. Formulae of $G(\mathrm{R};\mathfrak{U},\mathfrak{W})$ and $G(\mathrm{R};\Sigma_{\mathrm{R}};\mathfrak{U},\mathfrak{W})$

A moiety or atomic group in $\mathfrak{U}$ has a hydrophobicity level, such as charged, polar, or hydrophobic, all atoms in that moiety have the same hydrophobicity level.

Suppose that there are $L>1$ hydrophobicity levels, let H_i be the set of atoms having hydrophobicity level i, thus, $\{\boldsymbol{a}_i\}_{i=1}^n = \bigcup_{i=1}^L H_i$.

A water molecule inside $\mathfrak{S}_{\boldsymbol{R}}$ has chemical potentials μ_i if the nearest atom to the water molecule belongs to H_i. We may denote $\mu_1 > \cdots \mu_l > 0 > \mu_{l+1} > \mu_L$ such that $\mu_i > 0$ is hydrophobic and $\mu_j < 0$ is hydrophilic or polar. Quantum statistics derivation via grand canonic ensemble obtained the formula.

$$G(\boldsymbol{R};\mathfrak{U},\mathfrak{W}) = U(\boldsymbol{R};\mathfrak{U}) + \mu_e\langle\hat{N}_e\rangle + \sum_{i=1}^{L}\mu_i\langle\hat{N}_i\rangle, \tag{4}$$

where μ_e is the chemical potential of electron, $\langle\hat{N}_e\rangle$ and $\langle\hat{N}_i\rangle$ are means in equilibrium of numbers of electrons and water molecules with chemical potential μ_i respectively, not necessary integers.

The intrinsic ($V_{\boldsymbol{R}}$ in vaccum) potential $U(\boldsymbol{R};\mathfrak{U})$ comes from quantum mechanics, a good approximation is given as

$$U(\boldsymbol{R};\mathfrak{U}) = \sum_{1\le\alpha<\beta\le n}\left\{\frac{q_\alpha q_\beta}{4\pi\epsilon r_{\alpha\beta}} + \epsilon_{\alpha\beta}\left[\left(\frac{r_{\alpha\beta}^0}{r_{\alpha\beta}}\right)^{12} - \left(\frac{r_{\alpha\beta}^0}{r_{\alpha\beta}}\right)^{6}\right]\right\}, \tag{5}$$

where $r_{\alpha\beta} = |\boldsymbol{r}_\alpha - \boldsymbol{r}_\beta|$ and $\epsilon_{\alpha\beta} > 0$ and $r_{\alpha\beta}^0 > 0$ are constants, q_α are approximate net charge of $\boldsymbol{a}_\alpha$ at the nuclues of the atom $\boldsymbol{a}_\alpha$, and $\sum_{\alpha=1}^N q_\alpha = 0$ if the molecule is neutral. The second term in (5) is the Lennard-Jones potential that is often used as an approximate model for the isotropic part of a total (repulsion plus attraction) van der Waals force as a function of distance.

A connected closed surfaces $S\subset\mathbb{R}^3$ are surfaces without boundary, *i.e.*, $\partial S = \varnothing$. All connected closed surfaces are topologically classified into genus, from sphere (genus 0), tyre surface (genus 1), to genus 2, 3, $\cdots$, to ∞. See [14].

By the Jordan-Brouwer Separation Theorem, see Theorem 3.44 on page 254, and Example 2.46 on page 142 of [15], any connect closed surface $S\subset\mathbb{R}^3$ divides $\boldsymbol{R}^3$ into three parts, $\mathbb{R}^3 = \Omega_S \cup \Omega_S' \cup S$, where Ω_S is a bounded domain, Ω_S' is an unbounded domain, $S = \partial\Omega_S = \partial\Omega_S'$ is the common boundary, $\overline{\Omega}_S = \Omega_{\boldsymbol{R}} \cup S$ and $\overline{\Omega'}_S = \Omega_{\boldsymbol{R}}' \cup S$.

Let $\Sigma_{\boldsymbol{R}} \subset \mathfrak{S}_{\boldsymbol{R}}$ be an interface between $V_{\boldsymbol{R}}$ and water molecules inside $\mathfrak{S}_{\boldsymbol{R}}$. In case that $V_{\boldsymbol{R}}$ has cavities (holes large enough to be able to contain a water molecule), $\Sigma_{\boldsymbol{R}}$ has more than one connected component, say S_i, $1\le i\le k$, $\Sigma_{\boldsymbol{R}} = \bigcup_{i=1}^k S_i$. Each $S_i\subset\mathbb{R}^3$ is a connected closed surface. Let $S_1 = \Sigma_{\boldsymbol{R}}^\circ \subset \Sigma_{\boldsymbol{R}}$ be the largest connected component, the cavities are $O_{\boldsymbol{R},i} = \Omega_{S_i} \subset \Omega_{\boldsymbol{R}}^\circ = \Omega_{S_1}$, $\partial O_{\boldsymbol{R}_i} = S_i$, $2\le i\le k$. Note that even there is no water molecule in some cavity $O_{\boldsymbol{R}_i}$, we still count $S_i\subset\Sigma_{\boldsymbol{R}}$ as a connected component. Thus, $V_{\boldsymbol{R}} \subset \overline{\Omega}_{\boldsymbol{R}} \subset \overline{\Omega^\circ}_{\boldsymbol{R}} \subset \mathfrak{S}_{\boldsymbol{R}}$ and $\Omega_{\boldsymbol{R}}^\circ \setminus \Omega_{\boldsymbol{R}} = \bigcup_{i=2}^k \overline{O}_{\boldsymbol{R},i}$. Let n_i be the number of water molecules inside the cavity $O_{\boldsymbol{R},i}$. Define $V_{\boldsymbol{R},i} = V_{H_i}$, $V_{\boldsymbol{R}} = \bigcup_{i=1}^L V_{\boldsymbol{R},i}$ and $\mathrm{dist}(\boldsymbol{x},C) = \inf_{\boldsymbol{y}\in C}|\boldsymbol{y}-\boldsymbol{x}|$ for any set $C\subset\mathbb{R}^3$, then $\Sigma_{\boldsymbol{R}} = \bigcup_{i=1}^L \Sigma_{\boldsymbol{R},i}$, where $\Sigma_{\boldsymbol{R},i} = \{\boldsymbol{x}\in\Sigma_{\boldsymbol{R}}; \mathrm{dist}(\boldsymbol{x},V_{\boldsymbol{R},i} \le \mathrm{dist}(\boldsymbol{x},V_{\boldsymbol{R}}\setminus V_{\boldsymbol{R},i})\}$.

We have to distinguish cavity and pocket, they both mean hole in a conformation. A cavity is contained in $\Omega_{\boldsymbol{R}}^\circ$, any path in $(\Omega^\circ)_{\boldsymbol{R}}'$ to reach the cavity must

cross through $\Sigma_{\boldsymbol{R}}^{\circ}$. A pocket is contained in $(\Omega^{\circ})'_{\boldsymbol{R}}$. Some pocket may have long narrow path in $(\Omega^{\circ})'_{\boldsymbol{R}}$ to reach the bottom of the pocket, such that on the first look, one may think it is a cavity. With a little change of conformations from $\boldsymbol{R}$ to $\boldsymbol{R}'$, a cavity inside $V_{\boldsymbol{R}}$ may become a pocket outset of $V_{\boldsymbol{R}'}$, or vice versa.

$\langle \hat{N}_i \rangle$ is proportional to the area $A(\Sigma_{\boldsymbol{R},i})$, see, for example, [16]. Therefore, there are $\nu_i > 0$, such that

$$\nu_i A(\Sigma_{\boldsymbol{R},i}) = \langle \hat{N}_i \rangle, \quad 1 \le i \le L. \tag{6}$$

By an quantum mechanics argument that taking the mean via the electron charge density function $\rho_{\boldsymbol{R}}$ ([13], p. 6), we have

$$\langle \hat{N}_e \rangle = \langle \widehat{\rho_{\boldsymbol{R}}(\boldsymbol{r}_{\boldsymbol{R}}) V(\mathfrak{S}_{\boldsymbol{R}})} \rangle = \langle \widehat{\rho_{\boldsymbol{R}}(\boldsymbol{r}_{\boldsymbol{R}})} \rangle V(\mathfrak{S}_{\boldsymbol{R}}) = \nu_e V(\mathfrak{S}_{\boldsymbol{R}}) = \nu_e [V(\Omega_{\boldsymbol{R}}^{\circ}) + V(\mathfrak{S}_{\boldsymbol{R}} \setminus \Omega_{\boldsymbol{R}}^{\circ})] \tag{7}$$

Being a mean, the the constant $\nu_e = \langle \widehat{\rho_{\boldsymbol{R}}(\boldsymbol{r}_{\boldsymbol{R}})} \rangle$ is universal for all conformations of $\mathfrak{U}$ or any complexes in the size of proteins.

Let $d_w = d + \epsilon$, where d is the diameter of a water molecule and $\epsilon > 0$ is very small. By the definitions of $\Sigma_{\boldsymbol{R}}^{\circ}$ and $\Omega_{\boldsymbol{R}}^{\circ}$, we have roughly $V(\mathfrak{S}_{\boldsymbol{R}} \setminus \Omega_{\boldsymbol{R}}^{\circ}) = d_w A(\Sigma_{\boldsymbol{R}}^{\circ})$. Substitute (6) and (7) into (4), letting $\omega_e = \nu_e \mu_e$, $\omega_i = \nu_i \mu_i$, we obtain, a geometric apporximation to (4),

$$G(\boldsymbol{R}, \Sigma_{\boldsymbol{R}}; \mathfrak{U}, \mathfrak{W}) = U(\boldsymbol{R}; \mathfrak{U}) + \omega_e V(\Omega_{\boldsymbol{R}}^{\circ}) + \omega_e d_w A(\Sigma_{\boldsymbol{R}}^{\circ}) + \sum_{i=1}^{L} \omega_i A(\Sigma_{\boldsymbol{R},i}) \tag{8}$$

Because native structures of globular proteins have smaller volume and surface area than other conformations, see [17] and [18], $\omega_e > 0$. Since also $\nu_e > 0$, $\mu_e > 0$.

There are many choices for $\Sigma_{\boldsymbol{R}}$, Van der Waals surface $\partial P_{\boldsymbol{R}}$; the most common in literature is the solvent accessible surface $\mathrm{SAS}_{\mathbf{R}}$, it is the loci of the centre of a sphere or water molecule diameter d rolling on $\partial P_{\boldsymbol{R}}$; molecular surface, or solvent excluding surface $\mathrm{SES}_{\mathbf{R}}$, also produce by a sphere of radius r rolling on $\partial P_{\boldsymbol{R}}$ but taking the loci of the boundary of the rolling sphere, see [19] and [20]. By [21], $\mathrm{SES}_{\mathbf{R}}$ is the best boundary surface of a molecule. We will take $\mathrm{SES}_{\mathbf{R}}$ as $\Sigma_{\boldsymbol{R}}$ in the remaining of the paper.

The derivation of (4) and (8) is progressively done in [7]-[10] via quantum statistics without consideration of cavities and $U(\boldsymbol{R}; \mathfrak{U})$ was set to be the Coulomb's energy of positive charges $Z_\alpha e$ at the atomic centre $\boldsymbol{r}_\alpha$, negative charges were omitted.

3. Docking

3.1. Poses

For given receptor and ligand with structures (stable conformations), for example, a macromolecule $\mathfrak{P}$ with $\boldsymbol{P} = (\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n) \in \mathbb{R}^{3n}$ and a ligand $\mathfrak{L}$ with $\boldsymbol{L} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m) \in \mathbb{R}^{3m}$, initially $V_{\boldsymbol{P}}$ and $V_{\boldsymbol{L}}$ are separated by bulk water, *i.e.*, $\mathfrak{S}_{\boldsymbol{P}} \cap \mathfrak{S}_{\boldsymbol{L}} = \varnothing$. The task of docking is to find docking poses, *i.e.*, find an orientation preserving congruence (combinations of parallel translations and rotations) $C_L : \mathbb{R}^3 \to \mathbb{R}^3$, $C_L(\boldsymbol{L}) = (C_L(\boldsymbol{x}_1), \cdots, C_L(\boldsymbol{x}_m))$,

$\boldsymbol{PC}_L(\boldsymbol{L}) = (\boldsymbol{r}_1, \cdots, \boldsymbol{r}_n, C_L(\boldsymbol{x}_1), \cdots, C_L(\boldsymbol{x}_m)) \in \mathbb{R}^{3(n+m)}$ being the conformation (*i.e.*, statisfying steric conditions (1) to (3)) of the would be complex $\mathfrak{PL}$, such that $\mathfrak{S}_{\boldsymbol{P}} \cap \mathfrak{S}_{C_L(\boldsymbol{L})} \neq \varnothing$, $\Omega^{\circ}_{\boldsymbol{R}} \cap \Omega^{\circ}_{C_L(\boldsymbol{L})} = \varnothing$. That is, $N_r \geq 0$ of water molecules in $\mathfrak{S}_{\boldsymbol{P}} \cap \mathfrak{S}_{C_L(\boldsymbol{L})}$ are removed from the thermodynamic system $\mathfrak{S}_{\boldsymbol{PC}_L(\boldsymbol{L})} = \mathfrak{S}_{\boldsymbol{P}} \cup \mathfrak{S}_{C_L(\boldsymbol{L})}$.

There will be a new cavity O_{new} in $\mathfrak{S}_{\boldsymbol{PC}_L(\boldsymbol{L})}$ bounded by $\partial O_{new} = S \cup T \cup S_{new}$, where $S = \Omega^{\circ}_{\boldsymbol{PC}_L(\boldsymbol{L})} \cap \mathrm{SES}^{\circ}_{\boldsymbol{P}}$ and $T = \Omega^{\circ}_{\boldsymbol{PC}_L(\boldsymbol{L})} \cap \mathrm{SES}^{\circ}_{C_L(\boldsymbol{L})}$ are the docking sites, and $S_{new} = \mathrm{SES}^{\circ}_{\boldsymbol{PC}_L(\boldsymbol{L})} \setminus (\mathrm{SES}^{\circ}_{\boldsymbol{P}} \cup \mathrm{SES}^{\circ}_{C_L(\boldsymbol{L})})$. Then

$$\Omega^{\circ}_{\boldsymbol{PC}_L(\boldsymbol{L})} = \Omega^{\circ}_{\boldsymbol{P}} \cup \Omega^{\circ}_{C_L(\boldsymbol{L})} \cup O_{new} \quad \text{and} \tag{9}$$

$$\mathrm{SES}^{\circ}_{\boldsymbol{PC}_L(\boldsymbol{L})} = \mathrm{SES}^{\circ}_{\boldsymbol{P}} \cup \mathrm{SES}^{\circ}_{C_L(\boldsymbol{L})} \cup S_{new}, \tag{10}$$

Let $N_{new} \geq 0$ be the number of one layer water molecules inside $\mathfrak{S}_{\boldsymbol{PC}_L(\boldsymbol{L})}$ and touching S_{new} $N_r \leq N_{ST} = N_S + N_T$, N_S the number of water molecules inside $\mathfrak{S}_{\boldsymbol{P}}$ touching S, N_T the number of water molecules inside $\mathfrak{S}_{C_L(\boldsymbol{L})}$ touching T.

Denote $[\boldsymbol{PC}_L(\boldsymbol{L}), ST]$ as a pose or docking pose with the docking sites ST. There will be infinitely many such poses.

3.2. Scoring Function

Because of "Ligand binding typically occurs in an aqueous solvent", [12], we will use $G(\mathfrak{S}_{\boldsymbol{R}})$ ($G(\boldsymbol{R}; \mathfrak{U}, \mathfrak{W})$ or $G(\boldsymbol{R}, \mathrm{SES}_{\boldsymbol{R}}; \mathfrak{U}, \mathfrak{W})$) as a scoring function. Define the docking Gibbs free energy as

$$\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}), ST]} G' = G(\mathfrak{S}_{\boldsymbol{PC}_L(\boldsymbol{L})}) - G(\mathfrak{S}_{\boldsymbol{P}}) - G(\mathfrak{S}_{C_L(\boldsymbol{L})}). \tag{11}$$

A pose $[\boldsymbol{PC}_L(\boldsymbol{L}), ST]$ becomes a binding pose if and only if $\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}), ST]} G' < 0$.

3.3. Estimate of $\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}), ST]} G'$

Denote $G(\boldsymbol{PC}_{L(\boldsymbol{L})}; \mathfrak{PL}, \mathfrak{W}) = G(\mathfrak{S}_{\boldsymbol{PC}_L(\boldsymbol{L})})$. By the formula in (4) and denote $\langle N_e \rangle$ as $N_e(\boldsymbol{R})$, etc., substituting $\boldsymbol{R}$ by $\boldsymbol{PC}_L(\boldsymbol{L})$, $\boldsymbol{P}$, and $C_L(\boldsymbol{L})$ respectively, we obtain

$$\begin{aligned}\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}), ST]} G' = {} & U(\boldsymbol{PC}_L(\boldsymbol{L}); \mathfrak{PL}) - U(\boldsymbol{P}; \mathfrak{P}) - U(C_L(\boldsymbol{L}); \mathfrak{L}) \\ & + \mu_e [N_e(\boldsymbol{PC}_L(\boldsymbol{L})) - N_e(\boldsymbol{P}) - N_e(\boldsymbol{L})] \\ & + \sum_{i=1}^{L} \mu_i [N_i(\boldsymbol{PC}_L(\boldsymbol{L})) - N_i(\boldsymbol{P}) - N_i(\boldsymbol{L})]\end{aligned}$$

The total number of water molecules in $\mathfrak{S}_{\boldsymbol{PC}_L(\boldsymbol{L})}$, $\mathfrak{S}_{\boldsymbol{P}}$, and $\mathfrak{S}_{\boldsymbol{L}}$ be

$$N^{\boldsymbol{PC}_L(\boldsymbol{L})} = \sum_{i=1}^{L} N_i(\boldsymbol{PC}_L(\boldsymbol{L})), \quad N^{\boldsymbol{P}} = \sum_{i=1}^{L} N_i(\boldsymbol{P}), \quad N^{\boldsymbol{L}} = \sum_{i=1}^{L} N_i(\boldsymbol{L}),$$

Since in $\mathfrak{S}_{\boldsymbol{PC}_L(\boldsymbol{L})}$, $N_r \geq 0$ water molecules be removed and $N_{new} \geq 0$ water molecules be added,

$$\begin{aligned}N_e(\boldsymbol{PC}_L(\boldsymbol{L})) - N_e(\boldsymbol{P}) - N_e(\boldsymbol{L}) &= 10[N^{\boldsymbol{P}} + N^{\boldsymbol{L}} + N_{new} - N_r - N^{\boldsymbol{P}} - N^{\boldsymbol{L}}] \\ &= 10(N_{new} - N_r),\end{aligned}$$

$$\begin{aligned}\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' = {} & U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L}) - U(\boldsymbol{P};\mathfrak{P}) - U(C_L(\boldsymbol{L});\mathfrak{L}) \\ & + 10\mu_e(N_{new} - N_r) \\ & + \sum_{\mu_i>0}\mu_i[N_i(\boldsymbol{P}C_L(\boldsymbol{L})) - N_i(\boldsymbol{P}) - N_i(C_L(\boldsymbol{L}))] \\ & + \sum_{\mu_i<0}\mu_i[N_i(\boldsymbol{P}C_L(\boldsymbol{L})) - N_i(\boldsymbol{P}) - N_i(C_L(\boldsymbol{L}))].\end{aligned}$$

Let N_{ST} be the number of one layer water molecules almost contacting S and T in $\mathfrak{S}_{\boldsymbol{P}}$ and $\mathfrak{S}_{C_L(\boldsymbol{L})}$, $N_r \le N_{ST}$. Let $N_{r,i}$ and $N_{new,i}$ be the numbers of corresponding water molecules with chemical potential μ_i. Then $N_r = \sum_{i=1}^{L} N_{r,i}$ and $N_{new} = \sum_{i=1}^{L} N_{new_i}$, and

$$[N_i(\boldsymbol{P}C_L(\boldsymbol{L})) - N_i(\boldsymbol{P}) - N_i(C_L(\boldsymbol{L}))] = N_{new,i} - N_{r,i}, \quad i = 1,\cdots,L.$$

$$\begin{aligned}\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' = {} & U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L}) - U(\boldsymbol{P};\mathfrak{P}) - U(C_L(\boldsymbol{L});\mathfrak{L}) \\ & + 10\mu_e(N_{new} - N_r) \\ & + \sum_{\mu_i>0}\mu_i[N_{new,i} - N_{r,i}] \\ & + \sum_{\mu_i<0}\mu_i[N_{new,i} - N_{r,i}].\end{aligned} \tag{12}$$

In terms of formula (8), (9), and (10) we have

$$\begin{aligned}\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' = {} & U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L}) - U(\boldsymbol{P};\mathfrak{P}) - U(C_L(\boldsymbol{L});\mathfrak{L}) \\ & + \omega_e[V(O_{new}) + d_w A(S_{new})] \\ & + \sum_{i=1}^{L}\omega_i[A(\mathrm{SES}_{\boldsymbol{P}C_L(\boldsymbol{L}),i}) - A(\mathrm{SES}_{\boldsymbol{P},i}) - A(\mathrm{SES}_{\boldsymbol{L},i})]\end{aligned}$$

Denote $S_i = \mathrm{SES}_{\boldsymbol{P}C_L(\boldsymbol{L}),i} \cap S$, $T_i = \mathrm{SES}_{\boldsymbol{P}C_L(\boldsymbol{L}),i} \cap T$, $S_{new,i} = \mathrm{SES}_{\boldsymbol{P}C_L(\boldsymbol{L}),i} \cap S_{new}$, then $\partial O_{new,i} = S_i \cup T_i \cup S_{new,i} \subset \mathrm{SES}_{\boldsymbol{P}C_L(\boldsymbol{L}),i}$, and $A(\partial O_{new,i}) = A(S_i) + A(T_i) + A(S_{new,i})$,

$$\mathrm{SES}_{\boldsymbol{P}C_L(\boldsymbol{L}),i} = (\mathrm{SES}_{\boldsymbol{P},i} \setminus S) \cup (\mathrm{SES}_{C_L(\boldsymbol{L}),i} \setminus T) \cup S_i \cup T_i \cup S_{new,i},$$

Since $(\mathrm{SES}_{\boldsymbol{P},i} \setminus S) \cap (\mathrm{SES}_{C_L(\boldsymbol{L}),i} \setminus T) = \varnothing$, $(\mathrm{SES}_{\boldsymbol{P},i} \setminus S) \cap \partial O_{new,i} = \varnothing$, and $(\mathrm{SES}_{C_L(\boldsymbol{L}),i} \setminus T) \cap \partial O_{new,i} = \varnothing$,

$$A(\mathrm{SES}_{\boldsymbol{P}C_L(\boldsymbol{L}),i}) = A(\mathrm{SES}_{\boldsymbol{P},i} \setminus S) + A(\mathrm{SES}_{C_L(\boldsymbol{L}),i} \setminus T) + A(S_i) + A(T_i) + A(S_{new,i}),$$

and

$$A(\mathrm{SES}_{\boldsymbol{P},i} \setminus S) - A(\mathrm{SES}_{\boldsymbol{P},i}) = -A(\mathrm{SES}_{\boldsymbol{P},i} \cap S),$$

$$A(\mathrm{SES}_{C_L(\boldsymbol{L}),i} \setminus T) - A(\mathrm{SES}_{C_L(\boldsymbol{L}),i}) = -A(\mathrm{SES}_{C_L(\boldsymbol{L}),i} \cap T),$$

we have

$$\begin{aligned}\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' = {} & U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L}) - U(\boldsymbol{P};\mathfrak{P}) - U(C_L(\boldsymbol{L});\mathfrak{L}) \\ & + \omega_e[V(O_{new}) + d_w A(S_{new})] \\ & + \sum_{\omega_i>0}\omega_i[A(S_i) + A(T_i) + A(S_{new,i}) - A(\mathrm{SES}_{\boldsymbol{P},i} \cap S) - A(\mathrm{SES}_{C_L(\boldsymbol{L}),i} \cap T)] \\ & + \sum_{\omega_i<0}\omega_i[A(S_i) + A(T_i) + A(S_{new,i}) - A(\mathrm{SES}_{\boldsymbol{P},i} \cap S) - A(\mathrm{SES}_{C_L(\boldsymbol{L}),i} \cap T)].\end{aligned} \tag{13}$$

Let A_S be the set of atoms in $\mathfrak{P}$ such that these atoms (almost) touching S

and A_T be the set of atoms in $\mathfrak{L}$ such that these atoms (almost) touching T. Let $V_S \subset V_{\boldsymbol{P}} \subset \mathbb{R}^3$ and $V_T \subset V_{C_L(\boldsymbol{L})} \subset \mathbb{R}^3$ be the subsets corresponding to A_S and A_T. The intrinsic potential $U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L})$ is

$$U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L})) = U(\boldsymbol{P};\mathfrak{P}) + U(C_L(\boldsymbol{L});\mathfrak{L}) + U(V_{\boldsymbol{P}}, V_{C_L(\boldsymbol{L})}),$$

where

$$\begin{aligned} U(V_{\boldsymbol{P}}, V_{C_L(\boldsymbol{L})}) = {} & U(V_S, V_T) + U(V_S, V_{C_L(\boldsymbol{L})} \setminus V_T) + U(V_{\boldsymbol{P}} \setminus V_S, V_T) \\ & + U(V_{\boldsymbol{P}} \setminus V_S, V_{C_L(\boldsymbol{L})} \setminus V_T). \end{aligned}$$

Thus

$$\begin{aligned} & U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L})) - U(\boldsymbol{P};\mathfrak{P}) - U(C_L(\boldsymbol{L});\mathfrak{L}) \\ & = U(V_{\boldsymbol{P}} \setminus V_S, V_{C_L(\boldsymbol{L})} \setminus V_T) + U(V_S, V_{C_L(\boldsymbol{L})} \setminus V_T) + U(V_{\boldsymbol{P}} \setminus V_S, V_{C_L(\boldsymbol{L})}) + U(V_S, V_T). \end{aligned}$$

where $U(V_{\boldsymbol{P}} \setminus V_S, V_{C_L(\boldsymbol{L})} \setminus V_T)$, $U(V_S, C_L(\boldsymbol{L}) \setminus V_T)$, $U(V_{\boldsymbol{P}} \setminus V_S, V_T)$, and $U(V_S, V_T)$, are the potential energies between V_S and $C_L(\boldsymbol{L}) \setminus V_T$, V_S and V_T, and $V_{\boldsymbol{P}} \setminus V_S$ and V_T respectively.

By formula (5), $U(\boldsymbol{R};\mathfrak{U})$ is electrostatic,

$$U(V_{\boldsymbol{P}} \setminus V_S, C_L(\boldsymbol{L}) \setminus V_T) = \sum_{\boldsymbol{a}_i \notin A_S, \boldsymbol{a}_j \notin A_T} \left\{ \frac{q_i q_j}{4\pi\epsilon r_{ij}} + \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] \right\}, \tag{14}$$

$$U(V_S, C_L(\boldsymbol{L}) \setminus V_T) = \sum_{\boldsymbol{a}_i \in A_S, \boldsymbol{a}_j \notin A_T} \left\{ \frac{q_i q_j}{4\pi\epsilon r_{ij}} + \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] \right\}, \tag{15}$$

$$U(V_{\boldsymbol{P}} \setminus V_S, V_T) = \sum_{\boldsymbol{a}_i \notin A_S, \boldsymbol{a}_j \in A_T} \left\{ \frac{q_i q_j}{4\pi\epsilon r_{ij}} + \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] \right\}, \tag{16}$$

$$U(V_S, V_T) = \sum_{\boldsymbol{a}_i \in A_S, \boldsymbol{a}_j \in A_T} \left\{ \frac{q_i q_j}{4\pi\epsilon r_{ij}} + \epsilon_{ij} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] \right\}. \tag{17}$$

In (14) to (17), the first part is electrostatic counting for phenomena such as hydrogen bonds and salt bridges, $q_i q_j > 0$ is repulsive, $q_i q_j \le 0$ is attractive or neutral, $r_{ij} = |\boldsymbol{r}_i - C_L(\boldsymbol{x}_j)|$ is the distance between $\boldsymbol{a}_i$ in A_S and and $\boldsymbol{a}_j \in A_T$. The second term is the van der Waals force. Since although V_S and V_T are close enough to remove water molecules between them but still large enough such that the distance $r_{ij} > r_{ij}^0$, and r_{ij} in (14) to (16) are even larger. Hence all van der Waals forces in (14) to (17) are negative.

Since r_{ij} are larger than distances insider a molecule, the electronic part in (14) to (17) is either negative or very small positive. Thus, $U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L}) - U(\boldsymbol{P};\mathfrak{P}) - U(C_L(\boldsymbol{L});\mathfrak{L})$ is dominated by $U(V_S, V_T)$ and the negative van der Waals forces.

Especially, if V_S and V_T are electronic complementary, *i.e.*, positive polar part of V_S is nearing the negative part of V_T, and non-polar part nearing non-polar part. Then $q_i q_j \le 0$ in (17), therefore, $U(V_S, VT) < 0$ and $U(\boldsymbol{P}C_L(\boldsymbol{L});\mathfrak{P}\mathfrak{L}) - U(\boldsymbol{P};\mathfrak{P}) - U(C_L(\boldsymbol{L});\mathfrak{L}) \le 0$.

There are two kinds of electronic complementarity, 1. V_S and V_T are both hydrophobic pieces with hydrophilic islands, *i.e.*, large pieces of hydrophobic surface with some small pieces of hydrophilic surfaces (islands) such that hydrophobic part $q_i q_j = 0$ and on hydrophilic island part $q_i q_j < 0$. 2. Similarly, hydrophilic pieces with hydrophobic islands.

In case of hydrophilic to hydrophilic case, if V_S and V_T are close enough to form hydrogen bonds between V_S and V_T (that is, the electronically complementary is strict, we cannot afford to loss even one hydrogen bond opportunity), the reduced energy is more than enough to compensate the increased energy of removing double number of intermolecular hydrogen bonds between V_S (V_T) and water molecules. That is, the sum of the first, second, fourth line in (12) and (13) will be negative.

Summary 1: In case of f hydrophilic to hydrophilic, we claim that if the sum of the second and third lines of (12) and (13) are negative or zero, $\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]}G' < 0$. In fact, hydrogen bond and salt bridge are essentially electrostatic, therefore following the Coulomb's law $\frac{q_1 q_2}{4\pi\epsilon r}$ as in (14) to (17), where ϵ is dielectric constant. An intermolecular hydrogen bond (macromolecule with water) happens at the macromolecule surface where the dielectric constant is roughly $\epsilon \cong 40$, while inside a macromolecule (or inside $\Omega^{\circ}_{\boldsymbol{PC}_L(\boldsymbol{L})}$) the dielectric is about 3, see, for example ([22], p. 72) and [23]. Thus the sum of first and fourth lines in (12) and (13) is negative. If the sum of the second and third lines of (12) and (13) is also negative or zero, $\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]}G' < 0$.

Summary 2: if the docking sites S and T are close enough and electronically complementary, the first line in (12) and (13) will be negative. In case the electronically complementary is hydrophilic pieces with hydrophobic islands, if the sum of the second and third lines of (12) and (13) are negative or zero, $\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]}G' < 0$.

The second line in (12) is negative if $N_r > N_{new}$, the best case is that $N_{new} = 0$, *i.e.*, the area $A(S_{new})$ is so small such that almost no water molecules contacting to S_{new}. The larger N_r, the more negative the second line.

The second line in (13), being the contribution from electron chemical potential $\omega_e > 0$, should be negative like in (12). But since $G(\boldsymbol{R}, \mathrm{SES}_{\boldsymbol{R}}; \mathfrak{U}, \mathfrak{W})$ in (8) is only an approximation to $G(\boldsymbol{R}; \mathfrak{U}, \mathfrak{W})$ in (4), the second line in (13) is positive. To make it as small as possible: First, $V(O_{new})$ and $A(S_{new})$ should be as small as possible, consists with the best case in (12): $N_{new} = 0$. $V(O_{new})$ very small implies that: 1) the number of water molecules in O_{new} should be very small, even zero. Then $N_r = N_{ST}$, S_i, T_i, $S_{new,i}$ are all $\varnothing$; 2) the geometry of S and T should be complementary and close to each other, *i.e.*, if a part of S is convex, the corresponding part in T should be concave, and vice versa.

Summary 3: to make the sum of the first and second lines in (12) and (13) negative as much as possible, the best scenario is $N_r = N_{ST}$ as large as possible, $N_{new} = 0$. Since S_i, T_i, $S_{new,i}$ are all $\varnothing$, the sum of the second and third lines

in (12) and (13) is always negative, the fourth line in (12) and (13) is always positive or zero. Because of Summary 1, $\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' < 0$.

3.4. Searching Strategy for Binding Poses

Thus the sure searching strategy for binding poses is

Search 3.1 (Binding Pose Searching Strategy) *To make* $\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' < 0$, *make sure that* S *and* T *are so close to each other and geometrically complementary such that* $N_r = N_{ST}$ *and* $N_{new} = 0$, *be able to form hydrogen bonds or salt bridges between* A_S *and* A_T. S *and* T *also need be electronically complementary: either* S *and* T *are both hydrophobic pieces with hydrophilic islands or both hydrophilic pieces with hydrophobic islands and make sure not even one hydrogen bond or salt bridge opportunity in* $\Omega^{\circ}_{\boldsymbol{P}C_L(\boldsymbol{L})}$ *is lost.*

Remark 3.1. *It may happen that a few cavity water molecules in* O_{new} *but still* $\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' < 0$. *Thus do not rule out new cavity water molecules without weighing pros and cons.*

According to ([24], p. 138), for small molecule or an ion ligands, many proteins' docking site are dry cavities on its surfaces, but a ligand still has to remove water molecules to approach these macromolecules.

If the ligand is another macromolecule, then there are three kinds of binding sites: 1) Surface to string, an trough on the receptor's surface meeting an extended loop of a polypeptide chain of the ligand; 2) two α helices (one each from the macromolecules) paired together to form a coiled coil; 3) the most common type of docking sites are surface to surface. The three kinds of docking sites are typical complimentary in geometry.

For macromolecule-small molecule docking, S usually is the surface of a pocket (or even a cavity such that a slight change of conformation may change the cavity to pocket), T may be the whole $\mathrm{SES}^{\circ}_{C_L(\boldsymbol{L})}$ or a subsurface of it, fitting into the pocket of $\boldsymbol{P}$.

The binding pose searching strategy **Search 3.1** can be applied to these cases by finding geometrically and electronically complementary S and T.

4. Comparing with Observations and Practical Docking Results

The binding pose searching strategy **Search 3.1** is obtained theoretically by estimating (12) and (13) line by line. For example, geometrically complementary comes from estimating line 2 of (13), $\omega_e(V(O_{new}) + d_w A(S_{new})) > 0$. To make it as small as possible, the necessary condition is that the area $A(S_{new})$ and the volume $V(O_{new})$ are as small as possible. S_{new} could be a long band, so $A(S_{new})$ small means the band is very narrow. Since O_{new} occupies the space between the binding sites $S \subset \mathrm{SES}_{\boldsymbol{P}}$ and $T \subset \mathrm{SES}_{C_L(\boldsymbol{L})}$, $V(O_{new})$ small means: 1) S and T are spacially close to each other; 2) they are geometrically complementary, *i.e.*, at a place S is convex, the correponding place of T has to be concave, and vice versa, they fit together seamlessly.

These theoretical predictions are confirmed by observations and practical docking results.

A 1999 paper entitled "Examination of shape complementarity in docking of unbound proteins." [25], checked many set of complexes and protein pairs that will bind to become complexes, these cases show that "These findings argue that simplicity in the docking process, utilizing geometrical shape criteria may capture many of the essential features in protein-protein docking. In particular, they further reinforce the long held notion of the importance of molecular surface shape complementarity in the binding, and hence in docking.".

In docking practices, the strategy of S and T are geometrically and elecronically complementary in case of both are hydrophobic pieces with hydrophilic islands works. In [26], entitled "Hydrophobic docking: a proposed enhancement to molecular recognition techniques", it was found that "In view of the higher occurrence of hydrophobic groups at contact sites, their contribution results in more intermolecular atom-atom contacts per unit area for correct matches than for false positive fits. The hydrophobic groups are also potentially less flexible at the surface. Thus, from a practical point of view, a partial representation of the molecules based on hydrophobic groups should improve the quality of the results in finding molecular recognition sites, as compared to full representation. We tested this proposal by applying the idea to an existing geometric fit procedure and compared the results obtained with full vs. hydrophobic representations of molecules in known molecular complexes. The hydrophobic docking yielded distinctly higher signal-to-noise ratio so that the correct match is discriminated better from false positive fits."

Another example is [27], entitled "A hydrophobic-Ineraction-based mechanism trigger docking between the SARS-CpV-2 Spike and Angiotensin-Converting enzyme 2" shows that when S and T are hydrophobic pieces with hydrophilic islands and both geometrically and electronically complementary, the larger the areas $A(S)$ and $A(T)$, the larger the binding affinity, meaning that the docking Gibbs free energy $\Delta_{\boldsymbol{P}C_L(\boldsymbol{L}),ST]} < 0$ is more negative. See §5.3 a single molecule definition of binding affinity that implies the ensemble defined binding affinity.

Of the hydrophilic pieces with hydrophobic islands type of both geometrically and electronically complementary binding sites in nature, [28] entitled "Hydrogen bonding and molecular surface shape complementary as a basis for protein docking" states that "A geometric docking algorithm based upon correlation analysis for quantification of geometric complementarity between protein molecular surfaces in close interfacial contact has been developed by a detailed optimization of the conformational search of the algorithm. .....The utility of the spatial and directional properties of hydrogen bonding donor and acceptor sites for the identification of candidate docking conformations is demonstrated by the reliable preliminary reduction of conformation space, the improved geometric ranking of the minimum RMS conformations of some complexes and the overall reduction of

CPU time obtained."

It is described in [29] entitled "The role of DNA shape in protein-DNA recognition" made observations "The recognition of specific DNA sequences by proteins is thought to depend on two types of mechanism: one that involves the formation of hydrogen bonds with specific bases, primarily in the major groove, and one involving sequence-dependent deformations of the DNA helix. By comprehensively analysing the three-dimensional structures of protein-DNA complexes, here we show that the binding of arginine residues to narrow minor grooves is a widely used mode for protein-DNA recognition. This readout mechanism exploits the phenomenon that narrow minor grooves strongly enhance the negative electrostatic potential of the DNA. The nucleosome core particle offers a prominent example of this effect. Minor-groove narrowing is often associated with the presence of A-tracts, AT-rich sequences that exclude the flexible TpA step. These findings indicate that the ability to detect local variations in DNA shape and electrostatic potential is a general mechanism that enables proteins to use information in the minor groove, which otherwise offers few opportunities for the formation of base-specific hydrogen bonds, to achieve DNA-binding specificity."

That is, S in DNA surface and T in protein surface are geometrically and electronically complementary, S and T form salt bridges in narrow minor groove. In major groove it "involves the formation of hydrogen bonds". The "deformation of the DNA helix" is due to post-binding reshaping, see §5.1

5. Resolving so Far Unresolved Problems in Docking and Binding

In 2017, [12] posted three major unsolved problems, 1) Tackling Binding Site Flexibility; 2) Treating Solvent During Docking; 3) Affinity Prediction In Docking. Another unresolved problem in docking is covalent docking and binding. We will resolve them one by one in this section.

5.1. Tackling Binding Site Flexibility

> "The majority of protein binding sites have some degree of flexibility, either by the motions of specific side chains or possibly by larger motions of protein backbones. Often motions are induced, or selected, by the binding of a given ligand, the well-known principle of induced fit." [12].

Suppose V_S contains atoms $(\boldsymbol{a}_{s,1},\cdots,\boldsymbol{a}_{s,l})$, $l<n$, and $V_{C_L(\boldsymbol{L})}$ constains atoms $(\boldsymbol{a}_{t,1},\cdots,\boldsymbol{a}_{t,s})$, $s<m$. Define $\boldsymbol{P}_S=(\boldsymbol{r}_{s,i},\cdots,\boldsymbol{r}_{s,l})$ and $C_L(\boldsymbol{L})_T=(C_L(\boldsymbol{x}_{t,1}),\cdots,C_L(\boldsymbol{x}_{t,s}))$. If the flexibility means conformational changing of $\boldsymbol{P}$ and $C_L(\boldsymbol{L})$, especially on $\boldsymbol{P}_S$ and $C_L(\boldsymbol{L})_T$, post-folding answers this problem.

CGF together with single molecule thermodynamic hypothesis (SMTH) is the basis of a single molecule theory of protein (or more generally, molecular) folding, ([7]-[11]). SMTH says that all stable conformations (structures) must be the local

or global minimizer conformations of $G(\boldsymbol{R};\mathfrak{U},\mathfrak{W})$ or $G(\boldsymbol{R},\mathrm{SES}_{\boldsymbol{R}};\mathfrak{U},\mathfrak{W})$. An equivalent statement of SMTH is that if $\boldsymbol{R}_0$ is not stable, then there will be a molecular folding of $\mathfrak{U}$, starting from $\boldsymbol{R}_0$, resulting in a stable conformation $\boldsymbol{R}_1$. In particular $G(\boldsymbol{R}_1;\mathfrak{U},\mathfrak{W}) < G(\boldsymbol{R}_0;\mathfrak{U},\mathfrak{W})$, [11].

Since $\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]}G' < 0$, $\mathfrak{PL}$ can bind at the pose $[\boldsymbol{PC}_L(\boldsymbol{L}_b),ST]$. Mathematically the probability of the conformation $\boldsymbol{PC}_L(\boldsymbol{L})$ to be stable is almost zero, by SMTH, folding will happen and will result in a stable conformation $\boldsymbol{P}_1\boldsymbol{L}_1$ such that $G(\boldsymbol{P}_1\boldsymbol{L}_1;\mathfrak{PL},\mathfrak{W}) < G([\boldsymbol{PC}_L(\boldsymbol{L});\mathfrak{PL},\mathfrak{W})$ and $G(\boldsymbol{P}_1\boldsymbol{L}_1,\mathrm{SES}_{\boldsymbol{P}_1\boldsymbol{L}_1};\mathfrak{PL},\mathfrak{W}) < G([\boldsymbol{PC}_L(\boldsymbol{L}),\mathrm{SES}_{\boldsymbol{PC}_L(\boldsymbol{L})};\mathfrak{PL},\mathfrak{W})$. Because $\boldsymbol{P}_1 \neq \boldsymbol{P}$ and $\boldsymbol{L}_1 \neq C_L(\boldsymbol{L})$, these difference are called post-binding reshaping (after binding conformational change in [11]).

In particular, $\boldsymbol{P}_{1,S} \neq \boldsymbol{P}_S$ and $\boldsymbol{L}_{1,T} \neq C_L(\boldsymbol{L})_T$ correspondingly.

What need do is to develop software to really calculate the flexibility.

In fact, the mechanism of allostery is just post-binding reshaping, see [11].

5.2. Treating Solvent during Docking

> "An additional degree of complexity in docking is presented by the presence or absence of water molecules in the binding site. The importance of accounting for water when docking has been recognized by many groups. For example, in a recent study, the inclusion of water molecules was shown to be able to recover 56% of observed docking failures. In another study, the authors note that it is important to include only key waters. Not all water molecules are beneficial to docking; some indeed can have an adverse effect on docking performance." [12]

CGF specifically considers one layer of environment element in the thermodynamic system $\mathfrak{S}_{\boldsymbol{R}}$, in application we do not need consider bulk water molecules. In particular, CGF avoids the heavy computation of explicit water molecules.

5.3. Affinity Prediction in Docking

> "Affinity prediction remains a largely unsolved problem in computational chemistry. The reasons for this are rooted in the theory of statistical physics, which describes the energetics of binding." [12].

Statistical physics has nothing wrong, the problem is the way to apply statistical physics. For example, statistical mechanics is not for directly calculating time to time developing of a thermodynamic system until it reaches equilibrium. The way to use statistical mechanics is to theoretically derive useful formulae for physical quantities of thermodynamic system at equilibrium such as free energy, etc..

Using CGF as scoring function, we will derive the single molecule binding affinities for non-covalent binding as follows.

By SMTH in §5.1, from the binding pose $[\boldsymbol{PC}_L(\boldsymbol{L}),ST]$, folding will produce post binding reshaping $\boldsymbol{P}_1\boldsymbol{L}_1$, such that the binding Gibbs free energy is

$$\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G = G(\boldsymbol{P}_1\boldsymbol{L}_1;\mathfrak{P}\mathfrak{L},\mathfrak{W}) - G(\boldsymbol{P};\mathfrak{P},\mathfrak{W}) - G(C_L(\boldsymbol{L});\mathfrak{L},\mathfrak{W})$$
$$\leq \Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' < 0.$$

Then the binding affinity $K_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}$ of binding at the binding pose $[\boldsymbol{P}C_L(\boldsymbol{L}),ST]$ is defined as

$$K_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]} = \exp\left(\frac{-\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G}{k_BT}\right), \tag{18}$$

where $k_B = 1.380649\times10^{-23}\,\mathrm{JK}^{-1}$ is the Boltzmann constant and T is temperature. Since folding is spontaneous, the non-covalent binding not only does not need energy input, it contributes energy $-\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G$ to the environment.

Note that it is always $K_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]} > 1$ as long as $\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' < 0$.

If $\Delta_{[\boldsymbol{P}_b C_L(\boldsymbol{L}),ST]}G' > 0$, we may say its binding affinity is less than 1, equivalent to say that stable complex cannot be formed from the pose $[\boldsymbol{P}C_L(\boldsymbol{L}),ST]$.

Also note that if $\Delta_{[\boldsymbol{P}_b C_L(\boldsymbol{L}_b),S_bT_b]}G' < \Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G' < 0$, it is not necessary that $K_{[\boldsymbol{P}_b C_L(\boldsymbol{L}_b),S_bT_b]} > K_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}$ because post-binding reshapings are involved.

Larger binding affinity means stronger binding. In reality, it is not the bigger binding affinity the better, because macromolecules functions need flexibility of working complexes, just enough binding affinity will work and is easier to break the complex down when needed.

In ensemble experiments, binding affinity is defined by equilibrium constant K via concentrations of protein (P) and ligand (X),

$$K = \frac{[PX]}{[P][X]},$$

then the binding free energy $\Delta G_{\text{binding}}$ is defined in ([30], p. 95 and 105) as,

$$\Delta G_{\text{binding}} = -RT\ln K \text{ per mol},$$

where $R = 8.3145\,\mathrm{mol}^{-1}\mathrm{K}^{-1}$ is the universal gas constant. It is roughly the single molecule ΔG times the Avogadro constant $A = 6.02214076\times10^{23}$. Therefore, the ensemble binding affinity can be expressed as ([24], p. 62), it is also called equilibrium constant K,

$$K = \exp\left(-\frac{\Delta G^\circ}{RT}\right).$$

K is usually obtained by experiments. If we replace ΔG° by $A\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G$ as the Gibbs free energy difference per mol, then by $R = k_BA$, single molecule affinity $K_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}$ is exactly the ensemble affinity K.

The difference is, ensemble binding affinity can only be obtained from experiments, single molecule binding affinity can be calculated via CGF formulae (4) and (8).

5.4. Covalent Docking and Binding

Another problem is covalent binding. “We note that there are yet more areas

where docking could have impact but where the solutions offered as yet are somewhat limited. Some software features the ability to perform covalent docking, for example, but the methods used could be significantly enhanced. Further, we are currently unaware of solutions that try to tackle significant conformational change on covalent binding to the protein in a structure agnostic way." [12].

Let $\mathfrak{P}$ and $\mathfrak{L}$ be the receptor and ligand with stable conformations $\boldsymbol{P}=(\boldsymbol{r}_1,\cdots,\boldsymbol{r}_n)$ and $\boldsymbol{L}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m)$ respectively. In case of covalent binding, in both $\mathfrak{P}$ and $\mathfrak{L}$ some covalent bonds will be broken and a new covalent bond will be formed between $\mathfrak{P}$ and $\mathfrak{L}$, therefore, the number of atoms of the new molecule will be not be $n+m$, say it is $M=s+t$, $0\le s\le n$, $0\le t\le m$, *i.e.*, some atoms are removed due to bonds breaking. Relabelling if necessary, suppose that $\boldsymbol{a}_{s+1},\cdots,\boldsymbol{a}_n$ were removed in $\mathfrak{P}$ and $\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{m-t}$ were removed in $\mathfrak{L}$, denote $\boldsymbol{P}'=(\boldsymbol{r}_1,\cdots,\boldsymbol{r}_s)$ and $\boldsymbol{L}'=(\boldsymbol{x}_{m-t+1},\cdots,\boldsymbol{x}_m)$. Relabelling if necessary, the new covalent bond is between $\boldsymbol{a}_s$ in $\mathfrak{P}$ and $\boldsymbol{a}_{m-t+1}\in\mathfrak{L}$.

Bring $\boldsymbol{L}'$ to $\boldsymbol{P}'$ such that the distance between $\boldsymbol{a}_s$ in $\mathfrak{P}$ and $\boldsymbol{a}_{m-t+1}$ in $\mathfrak{L}'$ is the standard bond length r, *i.e.*, find an orientation preserving congruence C_L such that $\boldsymbol{P}'C_L(\boldsymbol{L}')=(\boldsymbol{r}_1,\cdots,\boldsymbol{r}_s,C_L(\boldsymbol{x}_{m-t+1}),\cdots,C_L(\boldsymbol{x}_m))\in\mathbb{R}^{3M}$ is a conformation, *i.e.*, satisfying steric conditions (1) to (3), and $r=|\boldsymbol{r}_s-C_L(\boldsymbol{x}_{m-t+1})|$. Call such C_L eligible.

Relabelling $\boldsymbol{a}_i$ as $\boldsymbol{a}_{s+i-m+t}$ in $\mathfrak{L}$, the new molecule $\mathfrak{PL}$ will have atoms $(\boldsymbol{a}_1,\cdots,\boldsymbol{a}_s,\boldsymbol{a}_{s+1},\cdots,\boldsymbol{a}_{s+t})$.

For every eligible C_L, let $\boldsymbol{b}=\dfrac{C_L(\boldsymbol{x}_{m-t+1})-\boldsymbol{r}_s}{r}$, R_ϕ the rotation around $\boldsymbol{b}$ of angle ϕ. Let $\Phi\subset[0,2\pi]$ such that $\boldsymbol{P}'R_\phi\circ C_L(\boldsymbol{L}')$ is still a conformation, *i.e.*, satisfying steric conditions (1) to (3). All docking poses will be conformations $\{\boldsymbol{P}'R_\phi\circ C_L(\boldsymbol{L}')=\boldsymbol{R}_\phi\}_{\phi\in\Phi}$ for all eligible C_L s.

By SMTH in §5.1, folding will push $\boldsymbol{R}_\phi$ to a minimizer conformation $\boldsymbol{R}_N$ of $G(\boldsymbol{R};\mathfrak{PL},\mathfrak{E})$ such that $G(\boldsymbol{R}_N;\mathfrak{PL},\mathfrak{W})<G(\boldsymbol{R}_\phi;\mathfrak{PL},\mathfrak{W})$. Define $\Delta_{C_L,\phi}G=G(\boldsymbol{R}_N;\mathfrak{P}'\mathfrak{L},\mathfrak{W})-G(\boldsymbol{R}_\phi;\mathfrak{PL},\mathfrak{W})<0$. This post-binding reshaping $\boldsymbol{R}_N\ne\boldsymbol{R}_\phi$ tackles "significant conformational change on covalent binding to the protein in a structure agnostic way." Although there are multiple minimizier conformations for the new molecule $\mathfrak{PL}$, $\boldsymbol{R}_N$ is the native structure of the new molecule $\mathfrak{PL}$ in the environment $\mathfrak{W}$, regardless which binding pose is chosen to start the folding. On the other hand, non-covalent binding may result in different conformations $\boldsymbol{P}_1\boldsymbol{L}_1$ of the complex $\mathfrak{PL}$ for different binding poses.

The binding affinity is defined by

$$K_{\mathfrak{PL},C_L,\phi}=\exp\left(\frac{-\Delta_{C_L,\phi}G}{k_BT}\right),\tag{19}$$

While the non-covalent binding is spontaneous, not only does not need energy input, but also release energy $-\Delta_{[\boldsymbol{P}C_L(\boldsymbol{L}),ST]}G$ to the environment. The covalent binding may release energy, or need energy input, depending on the amounts of energies released by breaking old covalent bonds and energy input for the new covalent bond, and the $\Delta_{C_L,\phi}G$ above.

6. Ensemble of Ligands

We consider non-covalent binding affinity in ensembles of receptors and ligands. Covalent binding can be similarly treated, but because of energy changes are complicatedly involved with chemical reactions, binding affinity is not that important as in non-covalent case.

6.1. Boltzmann Distribution

Let $\mathfrak{S}$ be a thermodynamic system surrounded by heat bath of constant temperature and pressure. Suppose that all microscopic states of $\mathfrak{S}$ may have energies E_i, $i=1,2,\cdots$. Suppose that $\mathcal{Z}=\sum_i \exp\left(-\frac{E_i}{k_B T}\right)<\infty$, the probability ρ_i of a microstate has energy E_i is

$$\rho_i=\frac{\exp\left(-\frac{E_i}{k_B T}\right)}{\mathcal{Z}}>0. \tag{20}$$

The ρ_i is called the Boltzmann distribution, $\mathcal{Z}$ is the partition function. The rigorous proof of Boltzmann distribution is non-trivial, see ([31], pp. 200-207).

6.2. Probability of Binding in an Ensemble of Ligands

Suppose there are M ligands and one receptor in a region Ω. Let $\mathfrak{P}$ be the receptor with Gibbs free energy $\epsilon_R=G(\boldsymbol{P};\mathfrak{P},\mathfrak{E})$, each ligand $\mathfrak{L}$ in the ensemble has Gibbs free energy $\epsilon_L=G(\boldsymbol{L};\mathfrak{L},\mathfrak{E})$. Let $\epsilon_B=\epsilon_{\mathrm{RC}}=G(\boldsymbol{P}_1\boldsymbol{L}_1;\mathfrak{P}\mathfrak{L},\mathfrak{E})$ as in §5.3. Divide Ω into lattices of N equal boxes, $N\gg M$, the receptor occupies one of the box. Suppose that the binding happens purely by chance, then there are $\frac{(N-1)!}{(M-1)!(N-M)!}$ ways (microstates) that one of the ligand binds $\mathfrak{P}$, each of them with Gibbs free energy $\epsilon_B+(M-1)\epsilon_L$; and $\frac{(N-1)!}{M!(N-M-1)!}$ ways that no binding happens, each of them with Gibbs free energy $\epsilon_R+M\epsilon_L$. By Boltzmann distribution (20), the partition function is

$$\mathcal{Z}=\exp\left(\frac{-[\epsilon_B+(M-1)\epsilon_L]}{k_B T}\right)\frac{(N-1)!}{(M-1)!(N-M)!}+\exp\left(\frac{-(\epsilon_R+M\epsilon_L)}{k_B T}\right)\frac{(N-1)!}{M!(N-M-1)!}.$$

Denote $\Delta\epsilon=\epsilon_B-\epsilon_R-\epsilon_L$. Because that

$$\frac{\frac{(N-1)!}{(M-1)!(N-M)!}}{\frac{(N-1)!}{M!(N-M-1)!}}=\frac{M!(N-M-1)!}{(M-1)!(N-M)!}=\frac{M}{N-M},$$

the probability of binding happens is

$$p_{\text{bound}} = \frac{\exp\left(\frac{-[\epsilon_B + (M-1)\epsilon_L]}{k_B T}\right)\frac{(N-1)!}{(M-1)!(N-M)!}}{\exp\left(\frac{-[\epsilon_{\text{RL}} + (M-1)\epsilon_L]}{k_B T}\right)\frac{(N-1)!}{(M-1)!(N-M)!} + \exp\left(\frac{-(\epsilon_{\text{R}} + M\epsilon_L)}{k_B T}\right)\frac{(N-1)!}{M!(N-M-1)!}}$$

$$= \frac{\exp\left(\frac{-\Delta\epsilon}{k_B T}\right)\frac{(N-1)!}{(M-1)!(N-M)!}}{\exp\left(\frac{-\Delta\epsilon}{k_B T}\right)\frac{(N-1)!}{(M-1)!(N-M)!} + \frac{(N-1)!}{M!(N-M-1)!}}$$

$$= \frac{\frac{M}{N-M}\exp\left(\frac{-\Delta\epsilon}{k_B T}\right)}{1 + \frac{M}{N-M}\exp\left(\frac{-\Delta\epsilon}{k_B T}\right)} \cong \frac{\frac{M}{N}\exp\left(\frac{-\Delta\epsilon}{k_B T}\right)}{1 + \frac{M}{N}\exp\left(\frac{-\Delta\epsilon}{k_B T}\right)} \quad \text{since } N \gg M. \tag{21}$$

Adjusting M such that $p_{\text{bound}} \approx 1/2$, *i.e.*, the two items in the denomenator of (21) is roughly equal. Equality of these two terms roughly amounts to $\exp(-\Delta\epsilon / k_B T) \cong (N-M)/M \gg 1$. the statement that the entropy lost in stealing one of the ligands from solution to bind it to the receptor is just made up for by the energetic gain ($\Delta\epsilon$) associated with binding the ligand to the receptor. Thus, whenever we consider ensemble of solute in solvent, mixed entropy appears.

Let $V(box)$ be the volume of a lattice box, the concentration of ligand is $c = M / NV(box)$. Assume the ground concentration is $c_0 = 1/V(box)$, then $c / c_0 = M / N$, denote $\beta = 1/k_B T$, we get that

$$p_{\text{bound}} = \frac{\frac{c}{c_0}e^{-\beta\Delta\epsilon}}{1 + \frac{c}{c_0}e^{-\beta\Delta\epsilon}}. \tag{22}$$

Thus, the probabilty p_{bound} is a function of the ligand concentration c. As described in ([32], p. 244), "This classic result goes under many different names depending upon the field (such as the Langmuir adsorption isotherm or a Hill function with Hill coefficient n = 1).".

6.3. Entropy of Mixing

$S_L = -k_B \ln\frac{c}{c_0} > 0$ is the entropy of mixing coming from arrangements of the M ligand in the solvent. Such that the chemical potential of the ligand is

$$\mu_L = \epsilon_L - TS_L = \epsilon_L + k_B T \ln\frac{c}{c_0} = \mu_L^\circ + k_B T \ln\frac{c}{c_0}. \tag{23}$$

Remember that $\Delta\epsilon = \epsilon_B - \epsilon_R - \epsilon_L$. If there are $N_R > 1$ receptors and N_L ligands, then the probability of each receptor binds with a ligand is given by p_{bound} in (22). Let

$$R + L \rightleftharpoons RL \tag{24}$$

represent the binding and $\mu_R^\circ = \epsilon_R$, $\mu_{\text{RL}}^\circ = \epsilon_{\text{RL}} = \epsilon_B$,

$$\mu_R = \mu_R^\circ + k_B T \ln \frac{c_R}{c_0}, \quad \mu_L = \mu_L^\circ + k_B T \ln \frac{c_L}{c_0}, \quad \mu_B = \mu_{RL} = \mu_{RL}^\circ + k_B T \ln \frac{c_{RL}}{c_0}.$$

be the chemical potentials of receptor, ligand, and bound complex RL.

6.4. Ensemble of Receptors and Ligands

In the ensemble of constant temperature and pressure, in any time the Gibbs free energy of the system $\mathfrak{S}$ is given by $G(T, P, N_R, N_L, N_{\mathrm{RL}})$ and

$$\frac{\partial G}{\partial N_R} = \mu_R, \quad \frac{\partial G}{\partial N_L} = \mu_L, \quad \frac{\partial G}{\partial N_{\mathrm{RL}}} = \mu_{\mathrm{RL}} = \mu_B, \tag{25}$$

see, for example, [31]. In equilibrium, $\mathrm{d}G(T, P, N_R, N_L, N_{RL}) = 0$, since T and P are constants, we have that

$$\begin{aligned} 0 = \mathrm{d}G(T, P, N_R, N_L, N_{RL}) &= \frac{\partial G}{\partial N_R}\mathrm{d}N_R + \frac{\partial G}{\partial N_L}\mathrm{d}N_L + \frac{\partial G}{\partial N_{RL}}\mathrm{d}N_{RL} \\ &= \mu_R \mathrm{d}N_R + \mu_L \mathrm{d}N_L + \mu_{RL}\mathrm{d}N_{RL}. \end{aligned}$$

In the beginning, $N_{RL} = 0$, whenever one receptor and one ligand bind, N_R and N_L are reduced by one, and N_{RL} is increased by one, thus $N = N_R + N_L + N_{RL}$ is not a constant, stoichiometric reasoning shows that for $N = N_R + N_L + N_{RL}$,

$$\mathrm{d}N_R = -\mathrm{d}N, \mathrm{d}N_L = -\mathrm{d}N, \mathrm{d}N_{RL} = \mathrm{d}N,$$

let $\nu_R = \nu_L = -1, \nu_{RL} = 1$, we have

$$0 = \mathrm{d}G(T, P, N_R, N_L, N_{RL}) = (\mu_R \nu_R + \mu_L \nu_L + \mu_{RL}\nu_{RL})\mathrm{d}N,$$

therefore,

$$\mu_{RL} - \mu_R - \mu_L = 0, \quad \mu_{RL}^\circ - \mu_R^\circ - \mu_L^\circ = k_B T \left(\ln \frac{c_R}{c_{R0}} + \ln \frac{c_L}{c_{L0}} - \ln \frac{c_{RL}}{c_{RL0}} \right).$$

$$\Delta\epsilon = \Delta\mu^\circ = \mu_{RL}^\circ - \mu_R^\circ - \mu_L^\circ = -k_B T \ln \left[\frac{c_{RL}}{c_{RL0}} \left(\frac{c_R}{c_{R0}} \right)^{-1} \left(\frac{c_L}{c_{L0}} \right)^{-1} \right],$$

therefore

$$e^{-\beta\Delta\epsilon} = \exp\left(\frac{-\Delta\mu^\circ}{k_B T} \right) = \exp\left\{ \ln \left[\frac{c_{RL}}{c_{RL0}} \left(\frac{c_R}{c_{R0}} \right)^{-1} \left(\frac{c_L}{c_{L0}} \right)^{-1} \right] \right\} = \frac{c_{R0}}{c_R} \frac{c_{L0}}{c_L} \frac{c_{RL}}{c_{RL0}},$$

and writing the concentration c_R as $[R]$, etc.,

$$\frac{c_{RL0}}{c_{R0} c_{L0}} \exp\left(\frac{-\Delta\epsilon}{k_B T} \right) = \frac{c_{RL}}{c_R c_L} = \frac{[RL]}{[R][L]} = K_{\mathrm{eq}} = \frac{1}{K_d}, \tag{26}$$

c_{R0}, c_{L0} and c_{RL0} are usually set as 1 M. In equilibrium ensemble theory of binding, K_{eq} is the equilibrium constant, K_d is the so-called disassociation constant of the binding (24), ([32], p. 269). These quantities can be obtained only by experiments, no ensemble theory of binding to calculate it. Because

$$\Delta\epsilon = \epsilon_B - \epsilon_R - \epsilon_L = \Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]}G,$$

by (18) we have

$$K_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]} = \exp\left(\frac{-\Delta_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]}G}{k_BT}\right) = \exp\left(\frac{-\Delta\epsilon}{k_BT}\right). \tag{27}$$

By (26), (27), and (18), we have

$$K_{[\boldsymbol{PC}_L(\boldsymbol{L}),ST]} = \exp\left(\frac{-\Delta\epsilon}{k_BT}\right) = \frac{c_{R0}c_{L0}}{c_{RL0}}K_{\text{eq}} = \frac{c_{R0}c_{L0}}{K_d c_{RL0}}. \tag{28}$$

Equation (28) shows in ensemble case that the binding affinity is proportional to the equilibrium constant K_{eq} and inverse proportional to disassociation constant K_d. In equilibrium ensembele setting, the equilibrium constant K_{eq} and the binding affinity have to be obtained through different experiments separately, see ([32], chapter 6).

7. Conclusions

As said before, protein functioning depends on docking and binding, ([30], chapter 4). If there is no docking, there will be no binding. If there is no binding, there will be no protein function. Therefor, studying of docking and binding is indispensable in biological science.

Using CGF as scoring function enables an analytic docking Gibbs free energy formulae (12) and (13), analysis them line by line reveals the binding searching strategy **Search 3.1**.

The single molecule thermodynamic hypothesis (SMTH) shows the mechanism of non-covalent docking and binding is just molecular folding.

The CGF is rigorously derived via quantum statistics, making the single molecule binding affinity calculable, unlike the ensemble binding affinity that can only be obtained by experiments.

Together with the resolution of the three so far unresolved problems in docking posted in [12], (1. Tackling binding site flexibility; 2. Treating solvent during docking; 3. Affinity prediction in docking. and tackles "significant conformational change on covalent binding to the protein in a structure agnostic way.") these results demonstrate the power of the single molecule theory of molecular folding based on CGF and SMTH.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E. (1982) A Geometric Approach to Macromolecule-Ligand Interactions. *Journal of Molecular Biology*, **161**, 269-288. https://doi.org/10.1016/0022-2836(82)90153-x

[2] Pantsar, T. and Poso, A. (2018) Binding Affinity via Docking: Fact and Fiction. *Molecules*, **23**, Article No. 1899. https://doi.org/10.3390/molecules23081899

[3] Allen, P., Nguyen, N., Humphrey, N.D., Mao, J., Chavez-Bonilla, D. and Sorin, E.J. (2024) A Hands-On Collaboration-Ready Single- or Interdisciplinary Computational Exercise in Molecular Recognition and Drug Design. *Education Sciences*, **14**, Article No. 139. https://doi.org/10.3390/educsci14020139

[4] Nivatya, H.K., Singh, A., Kumar, N., Sonam, Sharma, L., Singh, V., *et al.* (2025) Assessing Molecular Docking Tools: Understanding Drug Discovery and Design. *Future Journal of Pharmaceutical Sciences*, **11**, Article No. 111. https://doi.org/10.1186/s43094-025-00862-y

[5] Paggi, J.M., Pandit, A. and Dror, R.O. (2024) The Art and Science of Molecular Docking. *Annual Review of Biochemistry*, **93**, 389-410. https://doi.org/10.1146/annurev-biochem-030222-120000

[6] Uppathi, P., Rajakumari, S. and Saritha, K.V. (2025) Molecular Docking: An Emerging Tool for Target-Based Cancer Therapy. *Critical Reviews™ in Oncogenesis*, **30**, 1-13. https://doi.org/10.1615/critrevoncog.2024056533

[7] Fang, Y. (2012) Gibbs Free Energy Formula for Protein Folding. In: Morales-Rodriguez, R., Ed., *Thermodynamics—Fundamentals and Its Application in Science*, InTech, 47-82. https://doi.org/10.5772/52410

[8] Fang, Y. (2014) A Gibbs Free Energy Formula for Protein Folding Derived from Quantum Statistics. *Science China Physics, Mechanics & Astronomy*, **57**, 1547-1551. https://doi.org/10.1007/s11433-013-5288-x

[9] Fang, Y. (2015) Thermodynamic Principle Revisited: Theory of Protein Folding. *Advances in Bioscience and Biotechnology*, **6**, 37-48. https://doi.org/10.4236/abb.2015.61005

[10] Yi, F. (2019) Single Molecule Thermodynamics Hypothesis of Protein Folding and Drug Design. *Journal of Biosciences and Medicines*, **7**, 164-172. https://doi.org/10.4236/jbm.2019.711015

[11] Fang, Y. (2024) Applications of a Single Molecule Theory of Protein Dynamics. *Journal of Biosciences and Medicines*, **12**, 311-335. https://doi.org/10.4236/jbm.2024.126025

[12] Cole, J., Davis, E., Jones, G. and Sage, C.R. (2017) Chapter 12. Molecular Docking—A Solved Problem? In: Davis, A.M. and Edge, C., Eds., *Silico Drug Discovery Tools*, Elsevier, 297-318. https://doi.org/10.1016/b978-0-12-409547-2.12352-2

[13] Bader, R.F.W. (1994) Atoms in Molecules: A Quantum Theory. Clarendon Press.

[14] Seifert, H. and Threlfall, W. (1980) A Textbook of Topology. Academic Press.

[15] Hatcher, A. (2002) Algebraic Topology. Cambridge University Press.

[16] Pegram, L.M. and Record, M.T. (2006) Partitioning of Atmospherically Relevant Ions between Bulk Water and the Water/Vapor Interface. *Proceedings of the National Academy of Sciences*, **103**, 14278-14281. https://doi.org/10.1073/pnas.0606256103

[17] Novotný, J., Bruccoleri, R. and Karplus, M. (1984) An Analysis of Incorrectly Folded Protein Models. Implications for Structure Predictions. *Journal of Molecular Biology*, **177**, 787-818. https://doi.org/10.1016/0022-2836(84)90049-4

[18] Novotný, J., Rashin, A.A. and Bruccoleri, R.E. (1988) Criteria That Discriminate between Native Proteins and Incorrectly Folded Models. *Proteins: Structure, Function, and Bioinformatics*, **4**, 19-30. https://doi.org/10.1002/prot.340040105

[19] Connolly, M.L. (1983) Analytical Molecular Surface Calculation. *Journal of Applied Crystallography*, **16**, 548-558. https://doi.org/10.1107/s0021889883010985

[20] Connolly, M.L. (year) Molecular Surfaces. http://www.biohedron.com/

[21] Jackson, R.M. and Sternberg, M.J.E. (1993) Protein Surface Area Defined. *Nature*, **366**, 638. https://doi.org/10.1038/366638b0

[22] Finkelstein, A.V. and Ptitsyn, O.B. (2016) Protein Physics: A Course of Lectures. 2nd, Updated and Extended Edition, Academic Press.

[23] Sheu, S.-Y., Yang, D.-Y., Selzle, H.L. and Schlag, E.W. (2003) Energetics of Hydrogen Bonds in Peptides. *Proceedings of the National Academy of Sciences*, **100**, 12683-12687. https://doi.org/10.1073/pnas.2133366100

[24] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K. and Walter, P. (2015) Molecular Biology of the Cell. 6th Edition, Garland Science.

[25] Norel, R., Petrey, D., Wolfson, H.J. and Nussinov, R. (1999) Examination of Shape Complementarity in Docking of Unbound Proteins. *Proteins*: *Structure*, *Function*, *and Genetics*, **36**, 307-317. https://doi.org/10.1002/(sici)1097-0134(19990815)36:3<307::aid-prot5>3.3.co;2-i

[26] Vakser, I.A. and Aflalo, C. (1994) Hydrophobic Docking: A Proposed Enhancement to Molecular Recognition Techniques. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **20**, 320-329. https://doi.org/10.1002/prot.340200405

[27] Li, J.C., Ma, X.L., Guo, S., *et al*. (2020) A Hydrophobic-Ineraction-Based Mechanism Trigger Docking between the SARS-CpV-2 Spike and Angiotensin-Converting Enzyme 2. *Global Challenges*, **4**, Article ID: 2000067.

[28] Meyer, M., Wilson, P. and Schomburg, D. (1996) Hydrogen Bonding and Molecular Surface Shape Complementarity as a Basis for Protein Docking. *Journal of Molecular Biology*, **264**, 199-210. https://doi.org/10.1006/jmbi.1996.0634

[29] Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The Role of DNA Shape in Protein-DNA Recognition. *Nature*, **461**, 1248-1253. https://doi.org/10.1038/nature08473

[30] Bahar, I., Jernigan, R.L. and Dill, K.A. (2017) Protein Actions: Principle and Modelling. Garland Scinece. Taylor & Francis Group, LLC.

[31] Greiner, W., Neise, L. and Stöker, H. (1994) Thermodynamics and Statistical Mechanics. Springer-Verlag.

[32] Phillips, R., Condev, J., Theriot, J. and Garcia, H.G. (2012) Illustrated by Orme, Nigel, Physical Biology of the Cell. 2nd Edition, Garland Science.