Scoring Function

The following energy terms comprise Lead-Finder scoring functions:

Van der Waals interactions are calculated with 6-12 Lennard-Jones potential

$$\Delta G_{VdW} = k_{vdW} \sum_{i \in ligand} \sum_{j \in protein} LJ_{ij}(r_{ij}) \qquad (1)$$

where k_{vdW} is corresponding scaling coefficient, summation runs over protein and ligand atoms (except acceptor of hydrogen bond and hydrogen atom itself – for these atoms hydrogen bonding energy is calculated instead), and $LJ_{ij}(r_{ij})$ is a smoothed Lennard-Jones potential depending on types of atoms *i* and *j*. Parameters for calculating $LJ_{ij}(r_{ij})$ for standard atom types (O, N, C, H, CA – aromatic carbon, NX – nitrogen atom that cannot accept hydrogen bonds, P, S) were taken from CHARMM19 force field, while some modifications were introduced for better representation of united atoms. Parameters for halogens were taken from OPLSAA and re-optimized. Standard Lennard-Jones potential was modified according to our original technique to smooth energy inside the protein interior (which demonstrates poor behavior when protein-ligand overlapping takes place) and to mimic local protein flexibility by broadening energy minimum.

Interactions with metals are also calculated with 10-12 Lennard-Jones potential in the form:

$$\Delta G_{Me} = k_{Me} \sum_{\substack{i = ligand \;\; donor \;\; atom \\ j = metal}} \alpha_{i,j} \;\; LJ_{ij}(r_{ij}) \qquad (2)$$

where k_{Me} is a scaling coefficient, a_{ij} depends on the metal coordination state and relative orientation of ligand and metal orbitals, and LJ_{ij} is a smoothed 10-12 Lennard-Jones potential accounting for radial component of interaction energy. To calculate a_{ij} coordination state of a metal ion is detected first (from protein 3D-structure), and directions along which coordination with ligand (O, N, S) atoms is possible are built up. Then a_{ij} is calculated for each vacant coordination direction as a 6-th power of cosine of corresponding angle (between ligand-metal bond and metal coordination vector). Smoothing of the Lennard-Jones potential is applied when ligand-metal overlapping takes place. Potential 10-12 was used instead of 6-12 to catch specificity of ligand-metal coordination. Constants for the Lennard-Jones potential for metal-ligand interactions were adjusted after all other parameters of the scoring function were set up. For this purpose a set of approximately 100 protein-ligand complexes (containing ligands coordinated with metal ions) was extracted from PDB and parameters of LJ potential were adjusted to fit experimentally observed geometries. In this way parameterization for Fe^{2+}, Fe^{3+}, Zn^{2+}, Mg^{2+}, Ca^{2+}, Mn^{2+}, metal ions and O, N, S ligand atoms was achieved. Energy-scaling coefficient (k_{Me}) was adjusted using the training set of protein-ligand complexes as described below.

Electrostatic interactions account for: (a) the protein-ligand interaction itself, and (b) the polar component of ligand desolvation upon binding. Protein-ligand electrostatic interactions are calculated using the screened Coulomb potential (SCP) with distance- and microenvironment-dependent dielectric permittivity. Following the original works of Mehler et al [1,2] the energy of electrostatic interactions calculated with Lead-Finder can be presented as:

$$\Delta G_{elec} = \sum_{i \in ligand} \sum_{j \in protein} k_{elec,n}(h_i, b_i) E_{elec,n}(h_i, b_i, r_{ij}, q_i, q_j) \qquad (3)$$

where h_i denotes hydrophilicity of a microenvironment of atom i, b_i – its buried fraction, q_i and q_j – are partial atomic charges, r_{ij} – interatomic distance. Hydrophilicity (h_i) of a microenvironment is a relative (compared to water) value and the way it is calculated can be found in original papers describing the SCP electrostatic model[1,2]. Ligand's partial atomic charges are calculated with Gasteiger algorithm[3]. Depending on h_i and b_i one of the three scaling constants ($k_{elec,0}$ $k_{elec,1}$ *or* $k_{elec,2}$), and one of the three functions to calculate electrostatic interaction energy ($E_{elec,0}$, $E_{elec,1}$ *or* $E_{elec,2}$) are chosen correspondingly. Energy of electrostatic interaction is calculated according to formula:

$$E_{elec,n}(r_{ij}, q_i, q_j) = \frac{q_i q_j}{D_n(h_i, b_i, r_{ij}) r_{ij}}$$

where D_n is also one of the three (D_0, D_1, D_2) screening functions describing distance dependence of dielectric permittivity, particular choice of which depends upon microenvironment properties (defined by h_i and b_i). According to original works describing the SCP electrostatic model, two parameters (h_i and b_i) describing protein-ligand microenvironment classify electrostatic interactions in a discreet fashion into buried, intermediate and surface, for which special formulas for distance-dependent dielectric permittivity (D_0, D_1 or D_2 correspondingly) are taken. Though the original paper[3] gives a rule to choose one of the three dependencies relying on h_i and b_i, these relations were reconsidered during Lead-Finder scoring function construction and slightly modified to achieve better docking success rate and energy estimations. Energy-scaling coefficients were also adjusted independently for each type of interactions.

Additionally, the electrostatic (polar) contribution of ligand desolvation upon binding to protein is evaluated using an adapted version of the Born model with the formula:

$$\Delta G_{ligand-born} = \sum_{i \in ligand} \frac{1}{2} \left(\frac{1}{D_{ES}(R_{B,i})} - \frac{1}{D_W(R_{B,i})} \right) \frac{q_i^2}{R_{B,i}} \qquad (4)$$

where $D_{ES}(R_{Bi})$ denotes the dielectric screening calculated at the distance R_{Bi} (Born radius of atom i) from the center of ligand atom i in the protein-ligand complex, and $D_W(R_{Bi})$ denotes dielectric screening calculated at the distance R_{Bi} in water. Born radii for different types of atoms were taken from publication[4] without further optimization. However, our findings suggested that term $D_{ES}(R_{Bi})$ was quite sensitive to particular microenvironment characteristics. For this reason additional parameterization of terms entering screening function was performed to achieve better docking and scoring quality.

***Hydrogen-bonding energy contribution* is** calculated as a sum of energies of individual hydrogen bonds (or, briefly, H-bonds) formed between protein and ligand (E_{hb}), and energetic penalties arising from H-bond donors and acceptors in protein and ligand, which did not form H-bonds in the complex:

$$\Delta G_{HB} = k_{hb} E_{hb} + k_{hb,lig-pen} \Delta E_{hb,lig-pen} + k_{hb,prot-pen} \Delta E_{hb,prot-pen} \qquad (5)$$

Energy of individual H-bonds is calculated with the following formula:

$$E_{hb} = \sum_{\substack{i \in ligand \\ j \in protein}} k(h_i) E_{hb,ij} \qquad (6)$$

where coefficient $k(h_i)$ depends on the hydrophilicity of particular H-bond microenvironment: for h_i<-5 one coefficient (for hydrophilic bonds) is taken; other bonds are treated as hydrophobic with another coefficient. Energy of individual H-bond ($E_{hb,ij}$) is decomposed into angular and radial contributions according to formula:

$$E_{hb,ij} = c_{AHD,ij} \cdot c_{LP,ij} \cdot LJ_{ij}$$

Where $C_{AHD,ij}$ is a squared cosine of an on angle between acceptor atom, hydrogen and donor atom; $C_{LP,ij}$ is a squared cosine of an angle between acceptor-hydrogen vector and acceptor-lone electron pair vector; LJ_{ij} is a smoothed 10-12 Lennard-Jones potential.

Energetic penalties for missing potential H-bonds in the protein-ligand complex are calculated by proprietary Lead-Finder algorithm, which accounts for accessibility of each H-bond donor and acceptor for water molecules and strength of lost H-bonds upon ligand transfer from water to protein environment. Particularly, H-bonding penalties attributed to ligand are calculated as:

$$\Delta E_{hb,lig-penalty} = \sum_{\substack{i \in ligand \\ i \in D,A}} \left(hb_p - f \cdot hb_w \right)$$

where hb_p and hb_w denote average numbers of H-bonds that ligand forms in protein-bound state and in aqueous solution correspondingly, and f – is a degree of atom exposure to solution. Standard criteria (hydrogen-acceptor distance < 2.5 Å, donor-hydrogen-acceptor angle > 120^0) are applied to count hb_p, while hb_w values were taken in a tabulated form for different types of ligand atoms (particular values were heuristically adjusted by us).

Table 1. Average number of hydrogen bonds formed by different types of ligand atoms in solution.

Donor/Acceptor type	**Number of H-bonds**
H (polar)	0.6
N (sp^2 or sp^3 hybrid)	0.6
O (sp^3 or sp^3 hybrid)	0.8

Protein loss of hydrogen bonds induced by ligand binding is calculated with Lead-Finder original formula:

$$\Delta E_{hb,prot-penalty} = \sum_{\substack{i\in ligand\\ i\notin D,A}} \sum_{j\in water} e^{-\frac{r_{ij}^2}{1.5}}$$

according to which penalties are summed over all non-polar ligand atoms overlapping with probable positions of water molecules hydrogen-bonded to protein polar atoms. Most probable positions of water molecules solvating protein in the ligand-unbound state are calculated as: placed 2.75 Å from hydrogen donor (acceptor) along hydrogen-bonding direction (direction spanned over lone electron pair). This penalty term was found to be crucial for implicit accounting of protein specific desolvation arising from tightly-bound water molecules.

Non-polar solvation favored by hydrophobic contacts in the protein-ligand complex is accounted in a classical volume-based fashion[5]:

$$\Delta G_{sol,V} = k_{sol} \sum_{\substack{i\in ligand\\ j\in protein}} S_i V_j e^{-r_{ij}^2/3.6} \qquad (7)$$

where summation runs over all protein and non-hydrogen ligand atoms and S_i and V_i denote atomic solvation parameters (energy increment and volume correspondingly), r_{ij} – interatomic distance. It should be mentioned that volume-based solvation term accounts primarily for non-specific solvation effects; while more specific contributions are accounted with additional terms calculated in a surface-based fashion according to formula:

$$\begin{aligned}\Delta G_{sol,S} = {}& \alpha_{L,polar-P} S_{L,polar-P} + \alpha_{L,polar-S} S_{L,polar-S} + \\ & \alpha_{L,polar-P} S_{L,polar-S} + \alpha_{L,nonpolar-S} S_{L,nonpolar-S}\end{aligned} \qquad (8)$$

where S denotes the area of contact (in Å^2) of polar or non-polar ligand (L) atoms with protein (P) and solvent (S), and α – are corresponding scaling constants. Inclusion of surface-based energy term reduces artifacts arising from long-range and cumulative volume-based terms, which use to overestimate contributions from loosely bound (not forming direct contacts with protein) ligand moieties. Calculating this energy term is quite computationally expensive, that is why it is included only in the most precise form of the scoring function (see 'Types of energy calculations').

Internal energy losses of the ligand upon transition from the solvent to the protein-bound state are accounted by comparing the ligand internal energies in conformations typical for the solution and protein bound state:

$$\Delta G_{internal} = k_{nb}\left(E_{nb,ES} - E_{nb,W}\right) + k_{1-4}\left(E_{1-4,ES} - E_{1-4,W}\right) \qquad (9)$$

where k_{nb} and $k_{1\text{-}4}$ are scaling constants for non-bonded (van der Waals) and 1-4 interactions (special case of non-bonded interactions between atoms separated by three chemical bonds). First term in the sum is the difference of non-bonded energies of ligand in protein-ligand complex and water, and the second term is the same difference of 1-4 interaction energies. Non-bonded interactions are calculated with standard 6-12 Lennard-Jones potential (without smoothing). 1-4 interactions are also calculated with 6-12 Lennard-Jones potential but with reduced (by 0.2 A) atomic radii. Special term for 1-4 interactions compensates the absence of fully functional description of torsion potential in current implementation of Lead-Finder. Direct inclusion of torsional penalties based on standard molecular mechanical torsional potentials is currently included only for a set of particular chemical bonds (conjugated double bonds, conjugated aromatic bonds OH-group adjacent to the aromatic or double bond, carbonyl group adjacent to double bond).

Entropic losses accounting for freezing ligand's degrees of freedom upon binding to protein are calculated in a standard linear fashion:

$$\Delta G_{entrop} = k_{tors} n_{tors}$$

where n_{tors} denotes the number of freely rotatable bonds (FRB) in the ligand (except terminal groups comprising single heavy atom and attached hydrogen atoms – internal rotation of such groups is believed to preserve upon ligand binding) and k_{tors} is a corresponding scaling factor, which was fitted with the training set.

[1] E.L. Mehler Self-Consistent, Free Energy Based Approximation To Calculate pH Dependent Electrostatic Effects in Proteins J. Phys. Chem. 1996, 100, 16006-16018.

[2] E.L. Mehler and F. Guarnieri A Self-Consistent, Microenvironment Modulated Screened Coulomb Potential Approximation to Calculate pH-Dependent Electrostatic Effects in Proteins Biophysical Journal 1999, 75, 3–22.

[3] J. Gasteiger, T. Engel, Chemoinformatics, 2003, Wile-VCH Verlag GmbH&Co. KGaA. ISBN: 3-527-30681-1.

[4] E. Gallicchio, L.Y. Zhang, R.M. Levy, The SGB/NP Hydration Free Energy Model Based on the Surface Generalized Born Solvent Reaction Field and Novel Nonpolar Hydration Free Energy Estimators J Comput Chem 2002, 23, 517–529.

[5] P.F.W. Stouten, C. Frommel, H. Nakamura, C. Sander, Mol. Simul., 1993, 10, 97.