EXERCISE I.12

HYDROPHOBICITY IN DRUG DESIGN

Pietro Cozzini[1] and Francesca Spyrakis[2]

[1]Laboratory of Molecular Modelling, Department of General and Inorganic Chemistry, Chemical-Physics and Analytical Chemistry, University of Parma, 43100 Parma, Italy; [2]Department of Biochemistry and Molecular Biology, University of Parma, 43100 Parma, Italy

Hydrophobicity represents the tendency of a substance to repel water and to avoid the complete dissolution in water. The term "hydrophobic" means "water fearing", from the Greek words hydro, water, and phobo, fear. Being that hydrophobicity is one of the most important physicochemical parameters associated with chemical compounds, several studies have been carried out to understand, evaluate, and predict this parameter [1–8]. In fact, hydrophobicity governs numerous and different biological processes, such as, for example, transport, distribution, and metabolism of biological molecules; molecular recognition; and protein folding. Therefore, the knowledge of a parameter that describes the behavior of solutes into polar and nonpolar phases is essential to predict the transport and activity of drugs, pesticides, and xenobiotics.

The hydrophobic effect can be defined as "the tendency of nonpolar groups to cluster, shielding themselves from contact with an aqueous environment". The hydrophobic effect in proteins can also be described as the tendency of polar species to congregate in such a manner to maximize electrostatic interactions. Proteins, in fact, organize themselves to expose polar side-chains toward the solvent, and retain hydrophobic amino acid in a central hydrophobic core. The hydrophobic effect constitutes one of the main determinants of globular protein molecules structure and folding: The hydrophilic regions tend to surround hydrophobic areas, which gather into the central hydrophobic core, generating a protein characterized by a specific and function-related three-dimensional structure. This driving force not only guides protein folding processes, but also any kind of biological interaction. Biological molecules interact, mainly, via electrostatic forces, including hydrogen bonds or hydrogen-bonding networks, often formed through water molecules. During a protein-ligand association, water molecules not able to properly locate themselves at the complex interface, are displaced and pushed into the bulk solvent, increasing entropy. Thus, it is possible to define the hydrophobic effect as a free energy phenomenon, constituted by both

enthalpic and entropic phenomena [9].

The hydrophobic character of different amino acids was deeply studied, and the possibility of creating amino acid hydrophobicity scales was pursued by several biochemical researchers with different methods and approaches [10,11]. A complete understanding of the forces that guide amino acid interactions within proteins could lead to the prediction of protein structure and processes that drive a protein to fold into its native form.

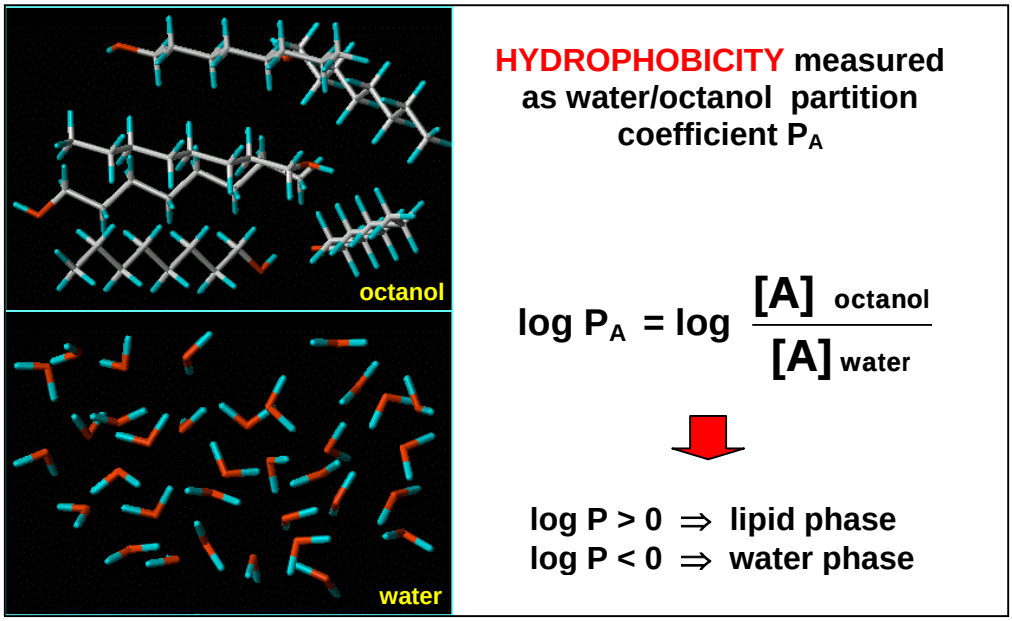
The octanol/water partition coefficient (log $P_{O/W}$) constitutes a quantitative, and easily accessible, hydrophobicity measurement. *P* is defined as the ratio of the equilibrium concentration of a substance dissolved in a two-phase system, formed by two immiscible solvents:

$$P_{O/W} = \frac{C_{octanol}}{C_{water}}$$

As a result, the partition coefficient *P* is the quotient of two concentrations and is normally calculated in the form of its logarithm to base 10 (log *P*), because *P* ranges from 10^{-4} to 10^{8}.

Log *P* values are widely used in bio-accumulation studies, in drug absorption and toxicity predictions and, recently, even in biological interactions modeling [12,13]. Several endeavours have been carried out to develop rapid and reliable log *P* estimation methodologies, capable of predicting the partition coefficient values for compounds not experimentally tested.

The common and standard procedure adopted for experimental log *P* estimation is the shake-flask method, used to determine the hydrophobicity of compounds ranging from –2 to 4 log *P* values. Log $P > 0$ characterize hydrophobic substances soluble in the lipid phase, while log $P < 0$ typifies polar compounds soluble in the water phase (Panel 1).



Panel 1

Fig. 1

As an experimental alternative, high-performance liquid chromatography (HPLC) is used for more hydrophobic compounds ranging from 0 to 6 log *P* values. Log *P* can be experimentally measured, or predicted from structural data. Experimental measurements are often time-consuming and difficult to make, thus, the need to properly and rapidly estimate hydrophobic parameters is more and more pressing. This need was also triggered by the advent of molecular modeling and the screening of large molecular libraries in the perspective of virtual screening and drug design.

Simultaneously, with new computational applications and molecular modeling progress and achievements, several methods, capable of predicting log *P* values for thousand of compounds, have been developed, and can now be classified into five major classes [14]: substituent methods, fragments methods, methods based on atomic contribution and/or surface areas, methods based on molecular properties, and, finally, methods based on solvatochromic parameters.

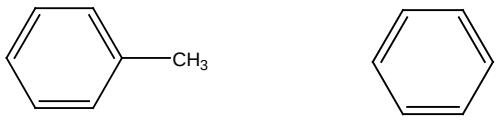
The first "by substituent" approach was proposed by Fujita and coworkers in 1964 [15]. Their technique is based on the following equation:

$$\pi = \log P_X - \log P_H$$

where P_X represents the partition coefficient of a derivative between 1-octanol and water and P_H that of the parent compound. Being that π typically is derived from equilibrium processes, it is possible to directly consider it as a free energy constant. As a consequence, log *P* represents an additive-constitutive, free energy-related property, numerically equivalent to the sum of the parent log *P* compound, plus a π term, representing the log *P* difference between a determinate substituent and the hydrogen atom which has been replaced [16]. As an example, the log *P* determination for

the methyl group is reported.

$$\log P_{CH_3} = \log P - \log P$$



The following "by fragments" methods was supported by Rekker and Mannhold, who stated that log *P* can be calculated as the sum of the fragment values plus certain correction factors. They determined the averaged contributions of simple fragments, using a large database of experimentally measured log *P* values [17,18]. Rekker did not indicate which fragment could be considered a valid fragment. The log *P* of molecules can be calculated using the formula

$$\log P = \sum a_n f_n + \sum b_m F_m$$

where *a* is the number of occurrences of fragment *f* of type *n* while *b* is the number of occurrences of correction factor *F* of type *m*.

The well-known CLOGP method clearly represents an improvement of the Rekker approach and, in fact, can be expressed by the same equation. CLOGP program breaks molecules into fragments and sums these constant fragment values and structure-dependent correction values taken from Hansch and Leo's database, to predict log *P* of several organic molecules. The program divides the target molecule into different fragments following a set of simple rules not alterable by users. CLOGP represents the first stand-alone program developed by Pomona MedChem, following Rekker general formulation. The program is now available on the Web (http://www.daylight.com/daycgi/clogp).

Different from chemical group fragments, the methods based on atomic contribution and/or surface area use atomic fragments and surface area data to predict hydrophobicity. The contribution of each atom to a molecule, in terms of hydrophobicity, can be evaluated by multiplying the corresponding atomic parameter by the degree of exposure to the surrounding solvent. The exposure degree is typically represented by the solvent-accessible surface area (SASA). The first promoters of this method were Broto and his colleagues, who developed a 222 descriptors set, made by combinations of up to four atoms with specific bonding pathways up to four in length, reaching a precision of about 0.4 log units [19]. Later, the concept of SASA was used by Iwase [20] and Dunn [21] in principal component analysis, to improve their log *P* estimations. Dunn computed the isotropic surface area, calculating the number of water molecules able to hydrate the polar portions of the solute molecules. As an example, one water molecule was allowed for groups as nitro, aniline, ketones, and tertiary amines, while two waters are allowed for other amines, three for carboxyls, and five for amide groups. The use of SASA parameters has been extended and introduced in several log *P* calculation algorithms, like the program HINT created by Abraham and Kellogg in 1991, which will be subsequently discussed and used for a practical session.

Various researchers did not agree with previously reported fragmental methods, claiming that a

molecule is rarely a simple sum of its parts and prediction of any molecular property on empirical or calculated fragments has no scientific basis [22]. The Bodor's method computes log *P* as a function of different calculated molecular properties, like conformations, ionization, hydration, ion-pair formation, keto-enol tautomerism, intramolecular and intermolecular H-bond formation, folding, and so forth.

The fifth log *P* determination method, based on solvatochromic comparisons, was proposed by Kamlet and coworkers [23] and constitutes, once more, a molecular properties methodology. Log *P* can be calculated through the following equation:

$$\log P_{oct} = a\,V + b\,\pi^* + c\,\beta_H + d\,\alpha_H + e$$

V is a solute volume term, π^* is a polarity/polarizability solute term, β_H is an independent measure of solute hydrogen-bond acceptor strength, α_H the corresponding hydrogen-bond donor strength, while *e* is the intercept. π^*, β_H, and α_H represent solvatochromic parameters obtained averaging multiple normalized solvent effects on a variety of properties, involving many different types of indicators.

Several research groups have tried to extend to amino acids the log *P* calculations, in order to better understand and investigate events like protein folding and biological interactions. However, experimental methods, like chromatography or site-directed mutagenesis, give ambiguous and different results [11]. Generally, each amino acid is characterized by a wide range of hydrophobicity values, thus, deciding and stating which value should correspond to a true measure becomes very difficult and time-consuming.

In order to obtain rapid and proper estimation of biological molecule hydrophobicity, in 1987 Abraham and Leo extended to common amino acids the fragment method of calculating partition coefficients [10]. Fundamental hydrophobic fragments, obtained from partitioning experiments performed on thousands of compounds, were subsequently reduced to atomic values with inherent bond, ring, chain, branching, and proximity factors. The derived hydrophobic atomic constants and the corresponding SASAs constituted the key parameter of the software **HINT** (**H**ydropathic **INT**eractions), able to directly calculate them for small molecules like ligands, or to obtain them from a residue-based dictionary. The program was thus created with the purpose of rapidly and properly estimating biological interactions such as protein–protein, protein–DNA, and protein–ligand and folding phenomena.

Why should we use log *P* to study and predict recognition and interactions between biological molecules? At least three reasonable answers could be given: (i) log *P* is essentially an experimental reproducible measurement; (ii) partition experiments are low cost and perform relatively rapidly; and (iii) log *P* is directly related to the free energy of binding. In fact, being that hydrophobicity is defined in terms of solubility, log $P_{o/w}$, and consequently also the hydrophobic atomic constants,

implicitly enclose hydrophobic and solvation/desolvation effects, directly related to the entropic contribution involved in molecular associations. The formation of a complex between a protein and a ligand in aqueous solution can be represented by the following equilibrium:

$$P_{aq.} + L_{aq.} \underset{k_{-1}}{\overset{k_{+1}}{\rightleftarrows}} P'L'_{aq.'}$$

where P is the protein, L the ligand, P'L' the new complex, and k_{+1} and k_{-1} are, respectively, the association and dissociation constants.

$$K_a = K_d^{-1} = \frac{[PL]}{[P][L]}$$

Both K_a (association) and K_d (dissociation) are related to the activity of the reacting species n, but, if extremely dilute solutions are considered, activities can be substituted by concentrations. Starting from the constant values it is possible to calculate the free energy of binding associated to the binding event, using the following relation:

$$\Delta G° = -RT \ln K_d$$

T is the absolute temperature, R the gas constant and $\Delta G°$ the binding free energy variation measured in standard condition (298 °K, 1 atm, and 1 M concentration for both reagents and products).

$P_{o/w}$ is also an equilibrium constant for solute transfer between octanol and water:

$$\log P_{o/w} = -\Delta G°/2.303\ RT$$

where R and T are constants. It derives that

$$\log P_{o/w} = k\ \Delta G°$$

where $k \approx -0.733$ kcal mol^{-1} at 298 K. Because

$$\Sigma a_i = \log P_{o/w}$$

it is obvious the relationship between hydrophobic atomic constants a_i and $\Delta G°$, thus, including both enthalpic and entropic contribution [9].

HINT can be defined as a natural and intuitive force field, able to estimate, using experimentally determined log P values, not only the enthalpic but also the entropic effects included in noncovalent interactions, like hydrogen bonding, Coulombic forces, acid-base and hydrophobic contacts. Hydrophobic and polar contacts, both identified as hydropathic interactions, are strictly related to solvent partitioning phenomena. In fact, the solubilization of a ligand in a mixed solvent system,

like water and octanol, involves the same processes and atom–atom interactions as biomolecular interactions within or between proteins and ligands [24]. The program was designed to consider and investigate hydrophobicity and hydropathic interactions in several biological areas. HINT is able to (i) calculate hydrophobic atomic constant for each atom in small molecule or even in macromolecule and quantitatively score molecular interactions, (ii) create hydrophobic maps or fields for small molecules in protein environments, (iii) map the hydrophobic and polar nature of the surrounding receptor from the structure of small interacting molecules, providing a hydrophobic interaction template for the definition of secondary and tertiary protein structure, and (iv) suggest modes of inter-helix interactions in trans-membrane ion channel [25]. All these features and capabilities make HINT a suitable tool, not only for the study of single and simple interactions, but also for the virtual screening of organic libraries and for structure-based drug design.

Interactions between atom–atom couples are calculated using the following equation:

$$b_{ij} = a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij}$$

where b_{ij} represents the interaction score between atoms i and j, a is the hydrophobic atomic constant, S is the SASA, T_{ij} is a logic function assuming –1 or +1 value, depending on the character of the interacting polar atoms, while R_{ij} and r_{ij} are a function of the distance between atoms i and j. The whole interaction between two molecules, like protein and ligand, or protein and DNA, can be represented as

$$\Sigma\Sigma\, b_{ij} = \Sigma\Sigma\, a_i S_i a_j S_j T_{ij} R_{ij} + r_{ij}$$

$b_{ij} > 0$ identifies favorable interactions, while $b_{ij} < 0$ the unfavorable ones. Interactions can be divided into: polar–polar, hydrophobic–hydrophobic, and hydrophobic–polar. While hydrophobic–hydrophobic contacts are always positively scored, polar interactions, depending on the charge of interacting groups can be favorable (acid–base), or unfavorable (acid–acid and base–base). Hydrophobic–polar contacts are constantly negatively scored by HINT, so they negatively contribute to the global binding energy. The HINT hydrophobic–polar interaction score term represents an empirical free-energy evaluation for the energy cost to desolvate the polar regions of proteins or ligands, placing them in a hydrophobic environment.

The HINT software allows us to reduce the information from bulk molecule solvent partitioning, to discrete interactions between biological molecules, i.e., ligand–protein, protein–protein, protein–DNA, and protein–ligand–water.

Small differences have been revealed in hydrophobicity estimations between HINT and CLOGP. Some examples are reported in Table 1 [25].

Table 1

Compound	HINT	CLOG-P
anthracene	4.45	4.49
1,3-butadiene	1.76	1.90
n-butylamine	0.97	0.92
cyclopentane	2.94	2.80
hexachlorobenzene	5.79	6.42
N-nitrosomorpholine	–0.41	–0.64
aldosterone	0.55	–0.14
cortisone	0.49	0.20
testosterone	3.35	3.35

HINT PRACTICAL APPLICATIONS

1. PROTEIN–LIGAND INTERACTIONS

Within an homogeneous biological set, HINT can be easily used to score and predict the free energy associated to protein-ligand complex formation. Starting from good crystallographic data and well experimentally determined Ki or IC_{50} values (Table 2), it is possible to obtain linear relationships between experimental $\Delta G°$ and computationally calculated HINT score values.

Table 2 reports the HINT score protein-ligand values calculated for two different homogenous set, formed, respectively, by eight bovine trypsin-ligand complexes and by nine tryptophan synthase-ligand complexes, for which experimental inhibition constants are reported in literature.

Table 2

PDB code	protein	$\Delta G°_{binding}$ (kcal/mol)	Hint score
1TNJ	bovine trypsin	–2.66	677
1TNK	bovine trypsin	–2.02	720
1TNI	bovine trypsin	–2.30	834
1TNL	bovine trypsin	–2.54	1360
1TNG	bovine trypsin	–3.98	923
1TNH	bovine trypsin	–4.57	972
3PTB	bovine trypsin	–6.43	1634
1PPH	bovine trypsin	–8.04	2663
1CX9	tryptophan syntethase	–9.58	2595
1C29	tryptophan syntethase	–9.00	2793
1C9D	tryptophan syntethase	–8.97	3094
1CW2	tryptophan syntethase	–8.76	3094
1C8V	tryptophan syntethase	–8.92	2571
2TRS	tryptophan syntethase	–7.20	2646
1QOP	tryptophan syntethase	–7.20	2721
1A50	tryptophan syntethase	–8.56	2914
2TSY	tryptophan syntethase	–4.65	905

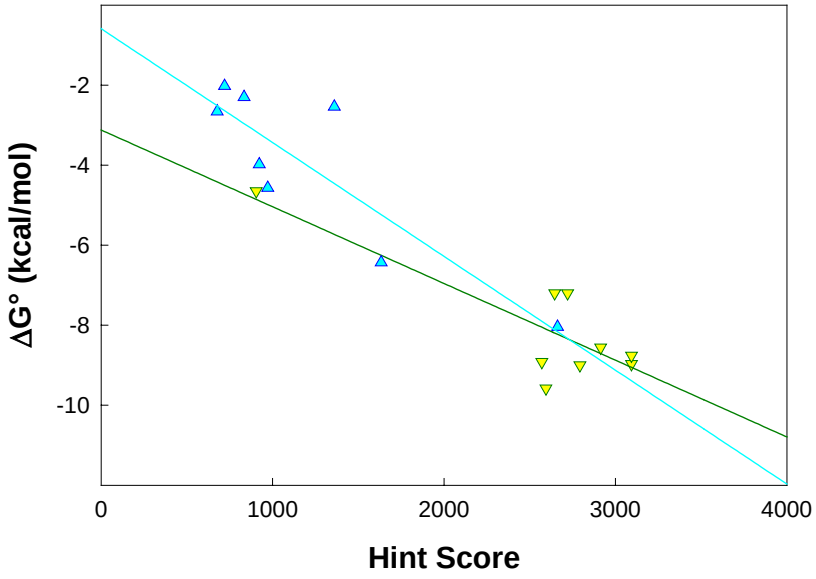


Fig. 2 Plots of experimental Δ*G*° vs. HINT score units for bovine trypsin (cyan triangle) and tryptophane synthase (green triangle).

The regression analyses of bovine trypsin and tryptophan synthase data series are shown in Fig. 2 and, respectively, represented by the following equations:

$$\Delta G° = -0.0019\ HS_{P-L} - 3.1210$$

$$\Delta G° = -0.0028\ HS_{P-L} - 0.5880$$

with R = 0.83, (standard error) SE = 0.90 kcal/mol for trypsin-ligand complexes and R = 0.87 and SE = 1.16 kcal/mol for tryptophan synthase-ligand complexes.

Thus, it is possible to predict the binding free energy of new hypothetical trypsin or tryptophan synthase ligands, for which the experimental inhibition constant value has not been yet determined, just calculating the HINT score value for the new potential complex, as shown in Fig. 3.

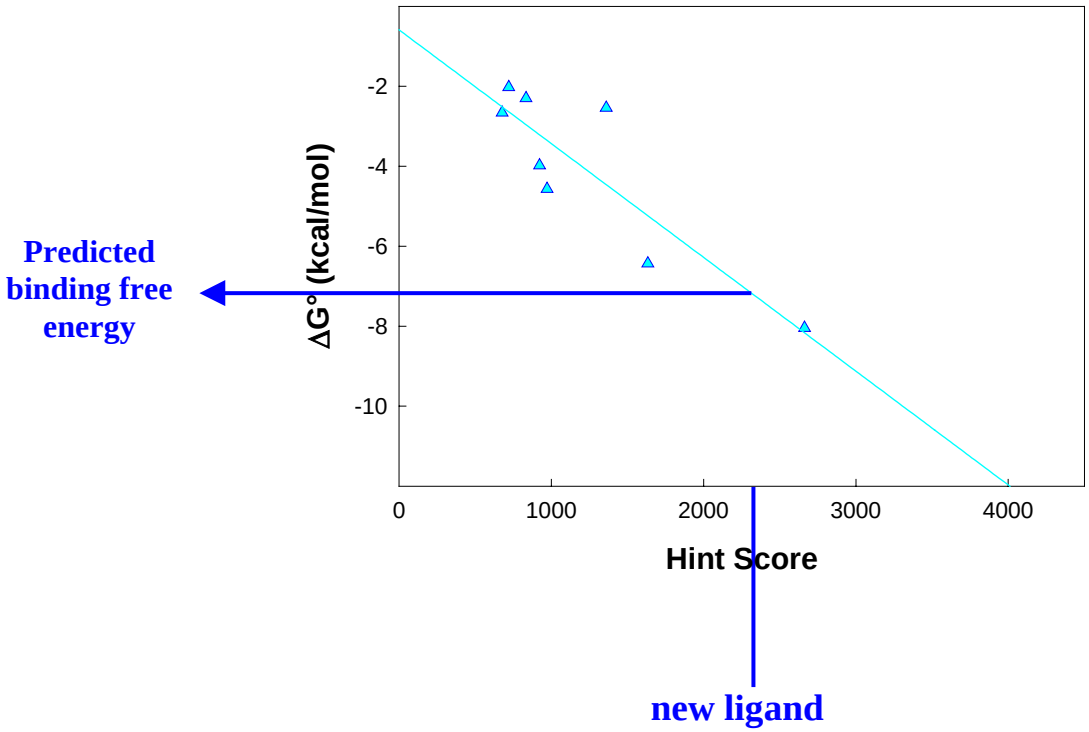


Fig. 3 Prediction of the binding free energy of a new potential bovine typsin ligand, from the protein–ligand HINT score

It is more difficult to find a good relationship between experimental and computational data, for a heterogeneous set of protein–ligand complexes, characterized by different active site polarity, ligands with diverse chemical nature, and inhibition constants varying among 10 or more orderd of magnitude [13]. In the following analysis, 93 different crystallographic protein–ligand complexes were examined and scored, in order to define a general relationship between $\Delta G°$ and HINT score. Experimental and calculated data, with both the protein nature and the crystallographic resolution values, are reported in Table 3, while the general relation is shown in Fig. 4.

Table 3

PDB code	Protein	Crystal resolution (Å)	$\Delta G°_{binding}$ (kcal/mol)	Hint score
1ETS	bovine thrombin	2.30	–11.17	3623
1ETT	bovine thrombin	2.50	–8.00	2131
1ETR	bovine thrombin	2.20	–10.49	2848
1UVT	bovine thrombin	2.50	–10.38	1834
1A2C	human thrombin	2.10	–8.97	3019
1A4W	human thrombin	1.80	–8.05	3110
1BHX	human thrombin	2.30	–9.30	2283
1D6W	human thrombin	2.00	–8.10	4005
1FPC	human thrombin	2.30	–9.52	2299
1C4U	human thrombin	2.10	–14.09	3882
1C4V	human thrombin	2.10	–14.67	4390
1C5N	human thrombin	1.50	–6.39	2334
1C50	human thrombin	1.90	–4.75	2498
1D4P	human thrombin	2.07	–8.57	3363
1KTT	human thrombin	2.10	–8.33	3586
1OYT	human thrombin	1.67	–9.85	3660

1AE8	human thrombin	2.00	–8.91	3616
1AFE	human thrombin	2.00	–6.25	3487
1BCU	human thrombin	2.00	–6.85	1755
1D9I	human thrombin	2.30	–12.38	4010
1TOM	human thrombin	1.80	–11.28	5271
1D3T	human thrombin	3.00	–8.73	1660
1D3Q	human thrombin	2.90	–8.88	1734
1D3P	human thrombin	2.10	–10.87	2094
1D3D	human thrombin	2.04	–12.35	2339
1DWB	human thrombin	3.16	–3.95	909
2YAS	hydroxynitrile lyase	1.72	–7.14	1988
5YAS	hydroxynitrile lyase	2.20	–4.43	1012
3YAS	hydroxynitrile lyase	1.85	–2.07	1206
1ADL	adipocyte lipid-binding protein	1.60	–9.16	3107
1LIE	adipocyte lipid-binding protein	1.60	–9.62	3001
1LID	adipocyte lipid-binding protein	1.60	–9.83	3486
1LIF	adipocyte lipid-binding protein	1.60	–9.64	3445
1HBP	retinol binding protein	1.90	–9.72	932
1ERB	retinol binding protein	1.90	–9.57	816
1FEL	retinol binding protein	1.80	–9.19	488
1TNJ	bovine trypsin	1.80	–2.66	677
1TNK	bovine trypsin	1.80	–2.02	720
1TNI	bovine trypsin	1.90	–2.30	834
1TNL	bovine trypsin	1.90	–2.54	1360
1TNG	bovine trypsin	1.80	–3.98	923
1TNH	bovine trypsin	1.80	–4.57	972
3PTB	bovine trypsin	1.70	–6.43	1634
1PPH	bovine trypsin	1.90	–8.04	2663
1CX9	tryptophan syntethase	2.30	–9.58	2595
1C29	tryptophan syntethase	2.30	–9.00	2793
1C9D	tryptophan syntethase	2.30	–8.97	3094
1CW2	tryptophan syntethase	2.00	–8.76	3094
1C8V	tryptophan syntethase	2.20	–8.92	2571
2TRS	tryptophan syntethase	2.04	–7.20	2646
1QOP	tryptophan syntethase	1.40	–7.20	2721
1A50	tryptophan syntethase	2.30	–8.56	2914
2TSY	tryptophan syntethase	2.50	–4.65	905
1BXQ	penicillopepsin	1.40	–10.02	4294
1BXO	penicillopepsin	0.95	–13.59	4435
1PPL	penicillopepsin	1.70	–11.62	3995
1PPM	penicillopepsin	1.70	–9.13	3908
1PPK	penicillopepsin	1.80	–10.40	3859
1APV	penicillopepsin	1.80	–12.23	4629
1APW	penicillopepsin	1.80	–10.87	4206
1FQ4	saccharopepsin	2.70	–8.70	4214
1FQ6	saccharopepsin	2.70	–10.70	3323
1FQ7	saccharopepsin	2.80	–7.37	3556
1LGR	glutamine synthetase	2.80	–4.17	2268
1ADF	alcohol dehydrogenase	2.90	–6.24	3467
2YPI	triosephosphate isomerase	2.50	–6.55	1978
1ULB	purine nucleoside phosphorylase	2.75	–7.23	2391
1DIH	dihydrodipicolinate reductase	2.20	–7.83	5517
1LYB	cathepsin	2.50	–15.5	5498
4HMG	hemagglutinin	3.00	–3.48	3459
1HXW	Hiv-1 protease	1.80	–14.71	3607
1HVJ	Hiv-1 protease	2.00	–14.25	3460
1HXB	Hiv-1 protease	2.30	–13.49	3135
1HTG	Hiv-1 protease	2.00	–13.20	4226
7HVP	Hiv-1 protease	2.40	–13.11	4311
1HPV	Hiv-1 protease	1.90	–12.57	3080
1HPS	Hiv-1 protease	2.30	–12.57	3124
4PHV	Hiv-1 protease	2.10	–12.51	3932
1AAQ	Hiv-1 protease	2.50	–11.45	3416
1HTF	Hiv-1 protease	2.20	–11.04	2641
1HIH	Hiv-1 protease	2.20	–10.97	3210
1SBG	Hiv-1 protease	2.30	–10.56	3037
1HVK	Hiv-1 protease	1.80	–13.80	3935
1HVI	Hiv-1 protease	1.80	–13.74	3734
1HVL	Hiv-1 protease	1.80	–12.27	3415
1HIV	Hiv-1 protease	2.00	–12.27	3660
1HBV	Hiv-1 protease	2.30	–8.68	2042
1QBT	Hiv-1 protease	2.10	–14.44	5170
1DMP	Hiv-1 protease	2.00	–12.99	4988
1AJX	Hiv-1 protease	2.00	–10.79	3357
1G35	Hiv-1 protease	1.80	–11.06	4198
1G2K	Hiv-1 protease	1.95	–10.82	3525
1AJV	Hiv-1 protease	2.00	–10.52	3916

The protein-ligand HINT score for the 93 complexes provides the following equation:

$$\Delta G° = -0.0018\ HS_{P\text{-}L} - 3.9041$$

with R = 0.68 and SE = 2.33 kcal/mol.

This general relationship could be used to predict the behavior of any potential ligand, not belonging to a specific class of inhibitors.

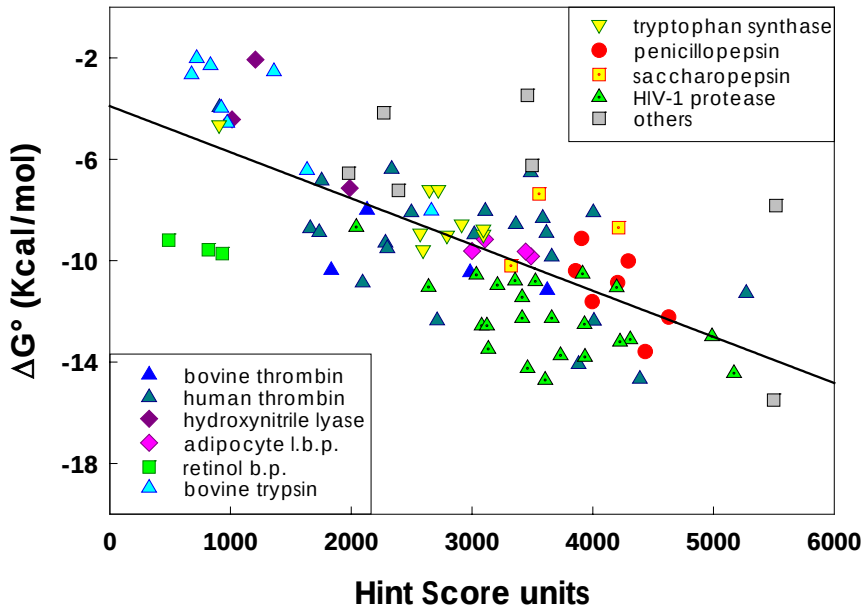


Fig. 4 Plot of experimental $\Delta G°$ vs. HINT score units, for a set of 93 different crystallographic protein–ligand complexes

2. CONSERVED WATER MOLECULES IN PROTEIN–LIGAND INTERACTIONS

It is well known that water molecules play a very significant role in biological recognition and interactions, exploiting its bridging properties, between proteins and ligands, proteins and proteins and proteins and nucleic acids [37]. Water can act directly, in water-mediated hydrogen bonds and, indirectly, in ligand, protein desolvation and hydrophobic interactions [37].

The HINT force field has been used to evaluate the role of conserved water molecules, through the same rapid protocol used to estimate protein–ligand interactions. Since water is integral in log $P_{o/w}$, salvation/desolvation and hydrophobic effects, the solvent bulk effects are implicitly encoded in the HINT parameters, but the constrained individual solvent molecules, bridging protein and ligand associations, must be explicitly considered and evaluated. Thus, the global HINT score for a complex interaction mediated by water molecules is given by two different contributions:

$$HS_{TOTAL} = HS_{protein\text{-}ligand} + HS_{ligand\text{-}water}\ [+ HS_{protein\text{-}water}]$$

If we assume that all the bridging interface-placed water molecules are pre-existing, and contribute to define the geometry and the chemical nature of the binding pocket, the latter protein-

water term can be ignored.

The specific contribution of a well-known conserved water molecule has been studied in a set of 23 crystallographic HIV-1 protease-inhibitor complexes, retrieved from Protein Data Bank (Fig. 5, Table 4, [26]). Water 301 is located on the HIV-1 protease symmetry axis, hydrogen-bonded to the protein monomers through Ile 50 and Ile 150 and specific peptidic inhibitors (Fig. 6). This conserved water has been crystallographically detected in the free form of the enzyme [27] and in all complexed structures [28–32], except when the ligands specifically designed to displace it were present in the binding pocket [33,34].

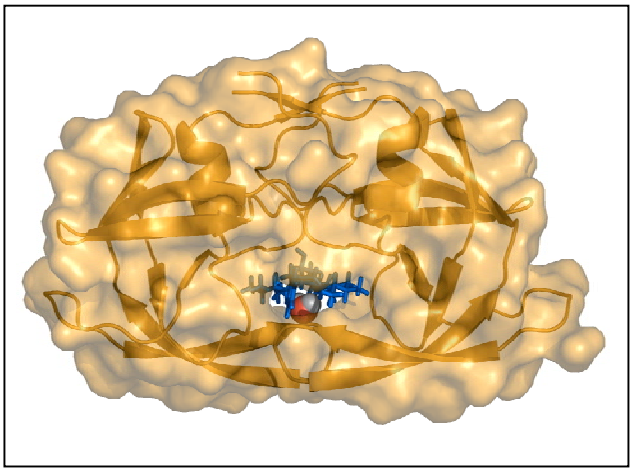


Fig. 5 HIV-1 protease complexed with CGP 53820 inhibitor. The protein is represented in ribbon tube cartoons, the ligand in sticks, and water 301 in spacefill.

PDB code	res (Å)	$\Delta G_{binding}$	HS_{P-L}	HS_{L-W301}
1HXW	1.80	-14.71	3607	1454
1HVJ	2.00	-14.25	3460	1203
1HXB	2.30	-13.49	3135	1049
1HTG	2.00	-13.20	4226	1272
7HVP	2.40	-13.11	4311	1229
1HPV	1.90	-12.57	3080	1058
1HPS	2.30	-12.57	3124	829
4PHV	2.10	-12.51	3932	789
1AAQ	2.50	-11.45	3416	633
1HTF	2.20	-11.04	2641	726
1HIH	2.20	-10.97	3210	1080
1SBG	2.30	-10.56	3037	1126
1HVK	1.80	-13.80	3935	1064
1HVI	1.80	-13.74	3734	1211
1HVL	1.80	-12.27	3415	1253
1HIV	2.00	-12.27	3660	1326
1HBV	2.30	-8.68	2042	777
1QBT	2.10	-14.44	5170	-
1DMP	2.00	-12.99	4988	-
1AJX	2.00	-10.79	3357	-
1G35	1.80	-11.06	4198	-
1G2K	1.95	-10.82	3525	-
1AJV	2.00	-10.52	3916	-

Table 4 Protein–ligand and ligand–water HINT scores determined for the 23 complexes.

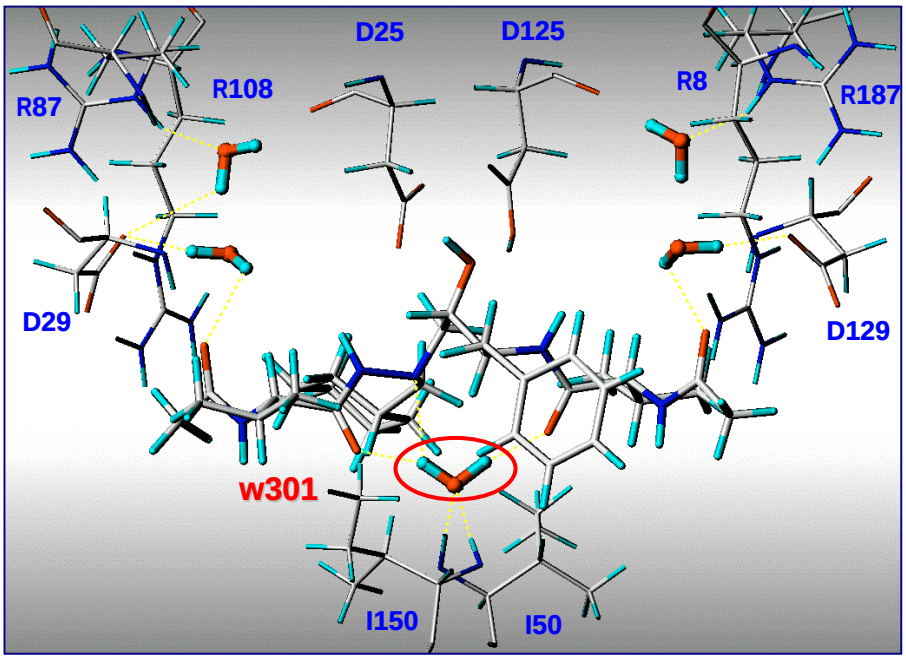


Fig. 6 HIV-1 protease active site. Water 301 placed at the protein–ligand interface is highlighted. Yellow dotted lines represent hydrogen bonds between ligand and water and protein and water.

The correlation between $HS_{protein–ligand}$ and experimental binding free energy is shown in Fig. 7. The linear regression is provided by the equation

$$\Delta G° = –0.0012\ HS_{P–L} –7.903$$

with a relatively poor R = 0.55 and an SE of 1.30 kcal/mol. The inclusion of water 301 contribution significantly improved the correlation between computational and experimental data, leading to an R = 0.80 and an SE of 1.0 kcal/mol. The new relationship, represented by the subsequent equation, is represented in Fig. 7.

$$\Delta G° = –0.0017\ HS_{TOTAL} –4.789$$

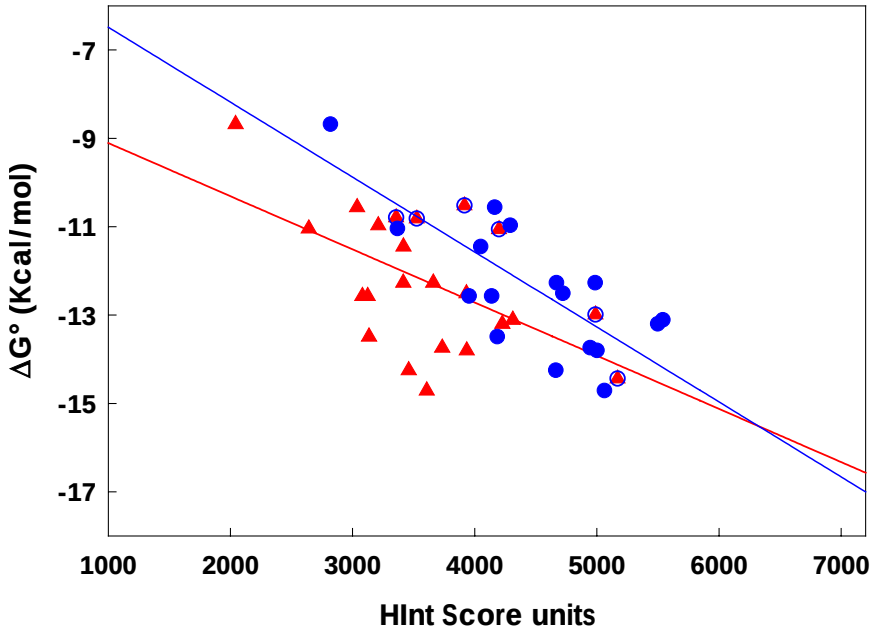


Fig. 7 Correlation between HINT score and experimental free energy of binding for the 23 analyzed HIV-1 protease complexes. Red line and triangles: correlation and scores without water 301 contribution. Blue line and circles: correlation and scores with water 301 contribution.

These results clearly testify for the value of explicitly modeling water behavior and contribution at molecular interfaces, and also the reliability of the HINT force field, in estimating the energetics of biological molecules through water mediation.

3. MODELING pH AND IONIZATION STATE IN BIOLOGICAL INTERACTIONS

In the process of developing a new potential lead compound, the exact attribution of the protonation state of ionizable groups, placed on both ligand and protein active site, is fundamental for obtaining reliable results.

Again, hydropathic analysis can be used to perform an automatic "computational titration", to define the precise number and location of hydrogen atoms into the binding pocket. As a case study, a complex formed by HIV-1 protease and a tripeptidic inhibitor (PDB code 1a30), for which the inhibition constants were calculated at seven different pH values [35], was studied and deeply analyzed [36].

The active site presented eight different ionizable groups, four located on the ligand and four placed on protein residues surrounding the binding cavity (Fig. 8). Using the computational titration tool implemented in the latest HINT version, we have modeled the pH-dependent inhibition, building all the possible 4374 unique protonation models, ranging from the most basic (all site deprotonated, global charge –7), to the most acidic (all site protonated, global charge +1). All 4374 models differ only in the number and placement of protons, thus we can define them as

isocrystallographic. The results of the titration analysis are reported in Fig. 9.

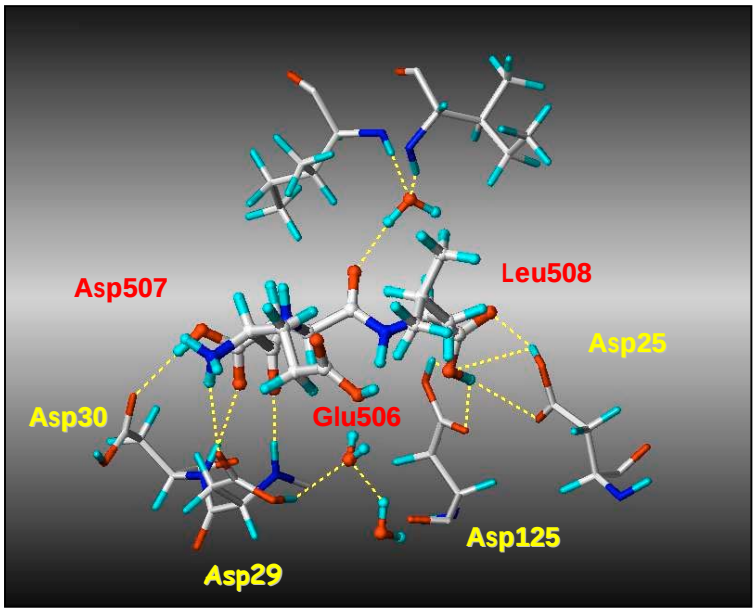


Fig. 8 1a30 active site: protein residues are represented in capped stick, while the ligand is shown in ball and sticks. In this model all the ionizable groups placed on both protein and ligand have been modeled protonated.

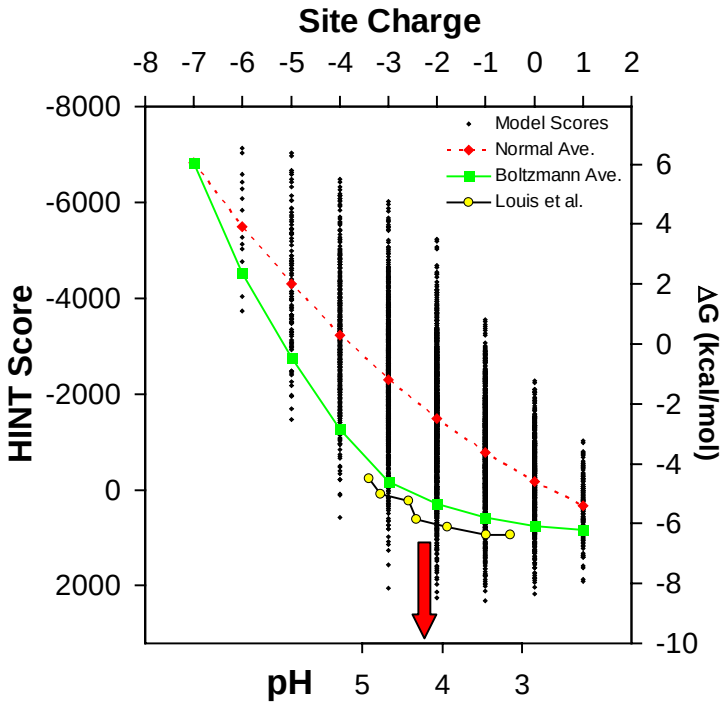


Fig. 9 Computational and experimental titration of 1a30. The arrow indicates the crystallization pH.

The computational and the experimental titration curves are nicely superimposed, in particular, the difference between measured and predicted binding free energy has been estimated to be about 0.6 kcal/mol throughout the experimental pH range.

The capability of finding the most probable protonation model, corresponding to the highest

HINT score value, could represent the discriminating key in designing new potential drugs, when no experimental information about ionizable group state is available.

REFERENCES

1. E. Benfenati, G. Gini, N. Piclin, A. Ronaglioni, M. R. Vari. *Chemosphere* **53**, 1155 (2003).
2. P. A. Carrupt, B. Testa, P. Gaillard. *Rev. Computational Chem.*, Wiley-VCH **11**, 241 (1997).
3. X. Q. Kong, D. Shea, W. A. Gebreyes, X. R. Xia. *Anal. Chem.* **77**, 1275 (2005).
4. A. J. Leo, P. Y. C. Jow, C. Silipo, C. Hansch. *J. Med. Chem.* **18**, 865 (1975).
5. T. Masuda, T. Jikihara, K. Nakamura, A. Kimura, T. Takagi, H. Fujiwara. *J. Pharm. Sci.* **86**, 57 (1997).
6. D. E. Smith, A. D. J. Haymet. *Rev. Computational Chem.* Wiley-VCH, **19**, 43 (2003).
7. M. Totrov. *J. Comput. Chem.* **25**, 609 (2004).
8. L. Xing, R. C. Glen. *J. Chem. Inf. Comput. Sci.* **42**, 796 (2002).
9. G. E. Kellogg, J. C. Burnett, D. J. Abraham. *J. Comput.-Aided Mol. Des.* **15**, 3813 (2001).
10. D. J. Abraham, A. J. Leo. *Proteins* **2**, 130 (1987).
11. K. M. Biswas, D. R. DeVido, J. G. Dorsey. *J. Chromatogr. A* **1000**, 637 (2003).
12. D. J. Abraham, G. E. Kellogg, J. M. Holt, G. K. Ackers. "Hydropathic analysis of the non-covalent interactions between molecular subunits of structurally characterized hemoglobins." *J. Mol. Biol.* **272**, 613 (1997).
13. P. Cozzini, M. Fornabaio, A. Marabutti, D. J. Abraham, G. E. Kellogg, A. Mozzarelli. *J. Med. Chem* **45**, 2469 (2002).
14. D. Eros, I. Kövesdi, L. Orfi, K. Takacs-Novak, G. Acsady, G. Keri. *Curr. Med. Chem.* **9**, 1819 (2002).
15. T. Fujita, J. Iwasa, C. Hansch. *J. Am. Chem. Soc.* **86**, 5175 (1964).
16. A. J. Leo. *Chem. Rev.* **93**, 1281 (1993).
17. R. Rekker. *The Hydrophobic Fragmental Constant*, Elsevier, Amsterdam (1977).
18. R. Rekker, R. Mannhold. *Calculation of Drug Lipophilicity*, VCH, Weinheim (1992).
19. P. Broto, G. Moreau, C. Vandycke. *Eur. J. Med. Chem.* **19**, 71 (1984).
20. K. Iwase, K. Komatsu, S. Hirono, S. Nakagawa, I. Moriguchi. *Chem. Pharm. Bull.* **33**, 2114 (1985).
21. W. J. Dunn III, M. G. Koehler, S. Grigoras. *J. Med. Chem.* **30**, 1121 (1987).
22. N. Bodor, Z. Gabanyi, C. K. Wong. *J. Am. Chem. Soc.* **111**, 3783 (1989).
23. M. K. Kamlet, J. L. M. Abboud, M. H. Abraham, R. W. Taft. *J. Org. Chem.* **48**, 2877 (1983).
24. G. E. Kellogg, D. J. Abraham. *Eur. J. Med. Chem.* **35**, 651 (2000).
25. G. E. Kellogg, G. S. Joshi, D. J. Abraham. *Med. Chem. Res.* **1**, 444 (1992).
26. M. Fornabaio, F. Spyrakis, A. Mozzarelli, P. Cozzini, D. J. Abraham, G. E. Kellogg. *J. Med.*

Chem. **47**, 4507 (2004).

28. H. Jhoti, O. M. P. Singh, M. P. Weir, R. Cooke, P. Murray-Rust, A. Wonacott. *Biochemistry* **33**, 8417 (1994).
27. B. Pillai, K. K. Kannan, M. V. Hosur. *Proteins* **43**, 57 (2001).
29. G. B. Dreyer, D. M. Lambert, T. D. Meek, T. J. Carr, T. A. Tomaszek Jr., A. V. Fernandez, H. Bartus, E. Cacciavillani, A. M. Hassell, M. Minnich, S. R. Petteway Jr., B. W. Metcalf. *Biochemistry* **31**, 6646 (1992).
30. A. L. Swain, M. M. Miller, J. Green, D. H. Rich, J. Schneider, S. B. H. Kent, A. Wlodawer. *Proc. Natl. Acad. Sci. USA* **8**, 8805 (1990).
31. J. P. Priestle, A. Fässler, J. Rösel, M. Tintelnot-Blomley, P. Strop, M. G. Grütter. *Structure* **3**, 381 (1995).
32. N. Thanki, J. K. M. Rao, S. I. Foundling, W. J. Howe, J. B. Moon, J. O. Hui, A. G. Tomasselli, R. L. Heinrikson, S. Thaisrivongs, A. Wlodawer. *Protein Sci.* **1**, 1061 (1992).
33. P. Y. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, C. H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe, S. Erickson-Viitanen. *Science* **263**, 380 (1994).
34. W. Schaal, A. Karlsson, G. Ahlsén, J. Lindberg, H. O. Andersson, U. H. Danielson, B. Classon, T. Unge, B. Samuelsson, J. Hultén, A. Hallberg, A. Karlén. *J. Med. Chem.* **44**, 155 (2001).
35. J. M. Louis, F. Dyda, N. T. Nashed, A. R. Kimmel, D. R. Davies. *Biochemistry* **37**, 2105 (1998).
36. F. Spyrakis, M. Fornabaio, P. Cozzini, A. Mozzarelli, D. J. Abraham, G. E. Kellogg. *J. Am. Chem. Soc.* **126**, 11764 (2004).
37. See the papers by Giurato et al. (Practice II.4.) and by Spyrakis et al. (Practice III.1) for further useful readings on log *P*. <http://www.daylight.com/daycgi/clogp>.

PRACTICAL SESSION:
Log *P* CALCULATION and BINDING FREE ENERGY PREDICTION

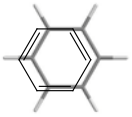
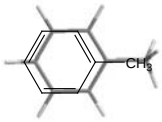
Different molecular modeling softwares allow the users to rapidly calculate log *P* values for several organic molecules. The following exercises have been conceived to be carried out using Sybyl, HINT, Spartan, and CLOGP, but you can try to reproduce them using any different kind of suitable program. (For more details, please contact pietro.cozzini@unipr.it).

Sybyl: http://www.tripos.com/
Spartan: http://www.wavefun.com/
CLOGP: http://www.daylight.com/daycgi/clogp

STEP 1

- Using Spartan software, build the molecules reported in Table 1.

Table 1

	benzene	
	nethyl benzene (toluene)	
	nitro-toluene	
	mino-toluene (para, meta and orto)	
	loro-toluene (para, meta and orto)	
6	change the hydrophobicity of your molecule adding different groups as you prefer…	

Your molecules should look like those reported in column 4.

1. Connect to http://www.daylight.com/daycgi/clogp and calculate log *P* values using CLOGP.

   - Select the Grins button.
   - Select a tool to build your molecule.

- Submit when done.
- Submit again, and you will obtain the log *P* result.

2. In Spartan, try to compute the log *P* values of the designed molecules, using the Ghose/Crippen and Villar options, with the following commands:

   - From the menu bar, select **Setup ⇒ Semi empirical ⇒** choose **Model AM1 ⇒ Save**.
   - Select **Setup ⇒ Properties ⇒** choose **Ghose/Crippen** and/or **Villar** options.
   - Select **Setup ⇒ Submit**.
   - Select **Display ⇒ Output** to analyze the result.

   (You can also build your molecules using the Get Fragment option of the Build Sybyl menu.)

3. Export all the molecule files in .mol2 format and submit to HINT software to compute log *P* calculations.
   - From the menu bar select **eslc ⇒ Hint ⇒ Partition ⇒ Molecule**.
   - In the *Atom expression* window, select the All button to partition the molecule in M1 area.

Fill up Table 2 with the log *P* values obtain from the four different calculation protocols.

Table 2

Molecule	CLOGP	Ghose/Crippen	Villar	HINT
benzene				
toluene				
nitro-toluene				
p-amino-toluene				
m-amino-toluene				
o-amino-toluene				
p-cloro-toluene				
m-cloro-toluene				
o-cloro-toluene				

Compare and comment the results, observing how hydrophobicity changes according to the different substituent chemical nature.

STEP 2

- Using Spartan software, build the aliphatic chains reported in Table 3.

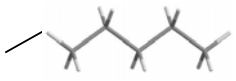
	*n*-pentane
	pentanoic acid
	pentanoic acid (deprotonated form)
	pentyl-amine
	pentyl-amine (protonated form)
	pentanol
6	change the hydrophobicity of your molecule adding different groups as you prefer…

Table 3

Calculate log *P* values using all the different methods and report the results in Table 4.

Table 4

Molecule	CLOGP	Ghose/Crippen	Villar	HINT
n-pentane				
pentanoic acid				
pentanoic acid (deprotonated form)				
pentyl-amine				
pentyl-amine (protonated form)				
pentanol				

Compare and comment the results, observing how hydrophobicity change according to the different substituent chemical nature.

Using the data reported in Tables 2 and 4, build two different graphs to show how log *P* increases or decreases according to hydrophobicity variation, analogous to the plot reported in Fig. 1.

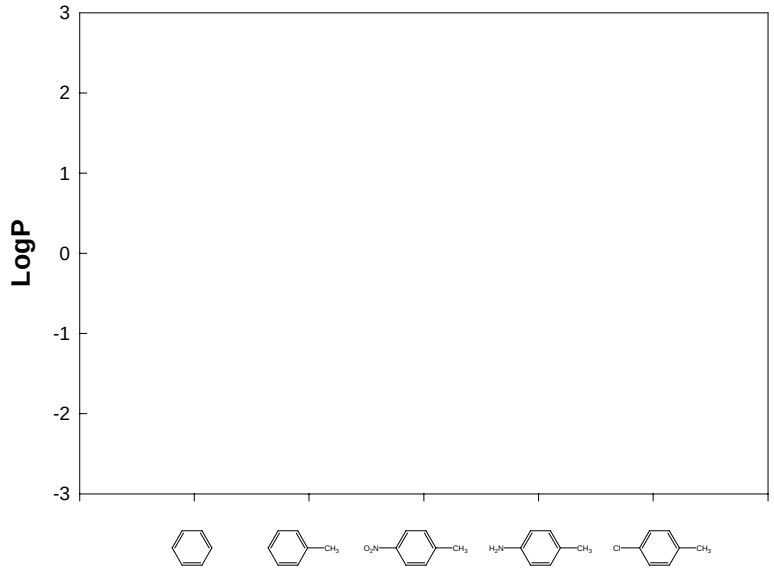


Fig. 1

STEP 3

- Using Spartan software, build several peptides formed by 3–6 amino acids (Fig. 2), changing every time the characteristics of the amino acids (polar, apolar, hydrophobic).

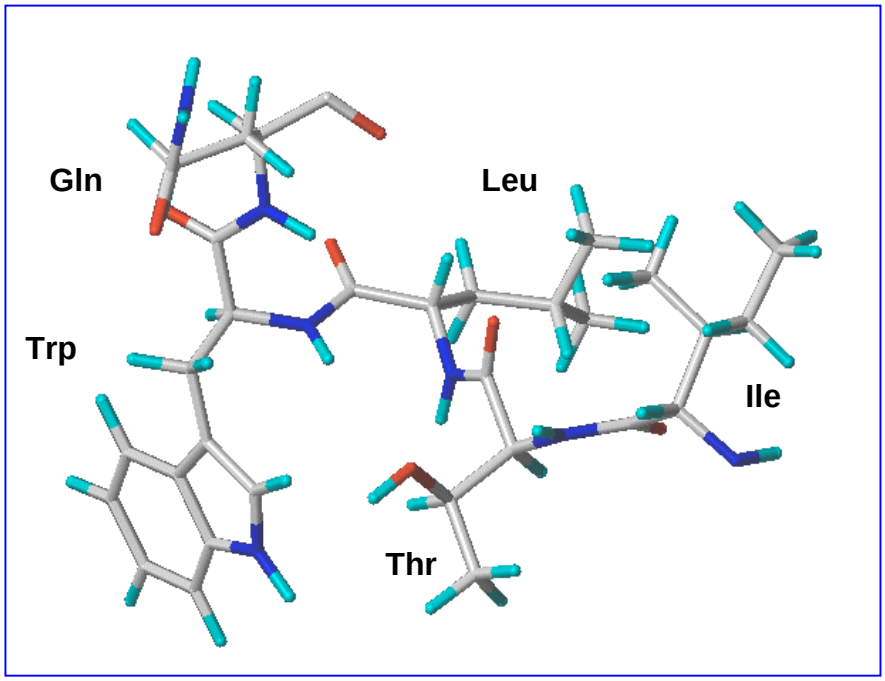


Fig. 2 Example of a peptidic chain formed by four amino acids (Gln-Trp-Thr-Leu-Ile).

- Calculate log *P* with the method you prefer.
- You will observe the variation of the log *P* in according with the variation of the amino acids character.

STEP 4 (Calculation of a protein–ligand complex logp and evaluation of the binding free energy)

- Connect to the Protein Data Bank Web site (http://www.rcsb.org/pdb/)
- Download a protein–ligand complex formed by HIV-1 protease and one of its peptidic inhibitors (PDB code 1a30)

  Your complex will looks like those reported in Figs. 3 and 4.

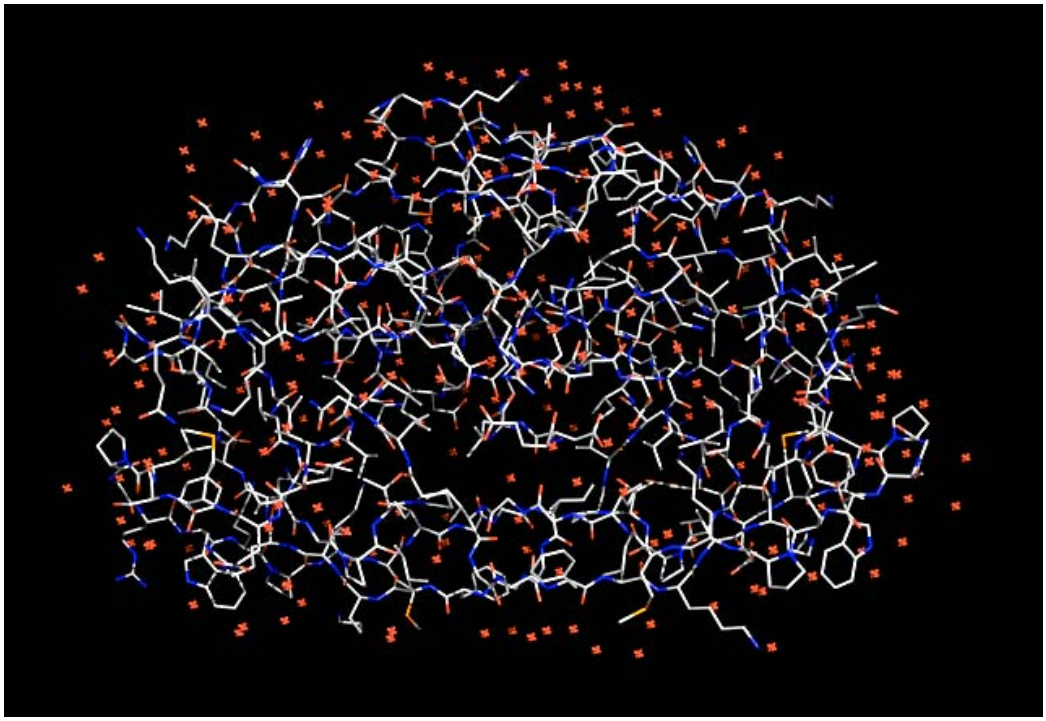


Fig. 3 Representation in sticks of the 1a30 protein–ligand complex.

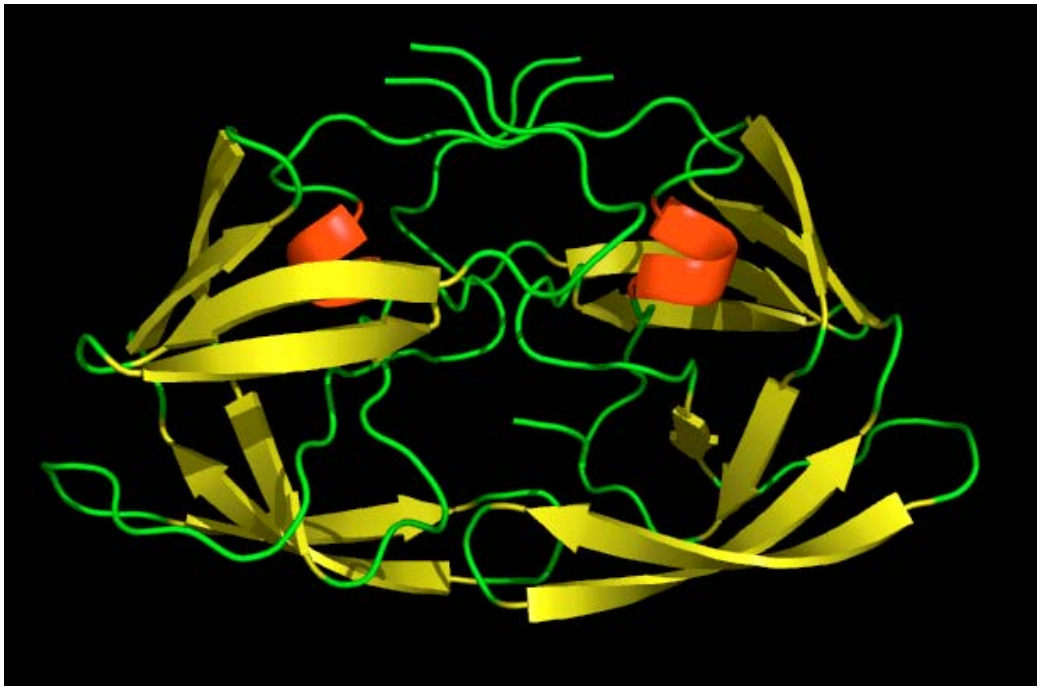


Fig. 4 Representation in ribbon tube of the 1a30 protein–ligand complex.

- Open 1a30.pdb file in Spartan/Sybyl.

- Delete all water molecules.
- Add hydrogens to both protein and ligand.
- Extract the ligand from the complex and save protein and ligand in two different .mol2 format files.
- Export the protein.mol2 and ligand.mol2 files in HINT.
- Calculate the log *P* of both protein and ligand.
- Calculate the HINT score of the protein-ligand complex.
- The calculation will take a few seconds.
- You will obtain an HINT score value directly related to the free energy of binding associated to the complex formation.
- Try to plot your data into the general relationship graph, to find out which should be the $\Delta G°$ value.

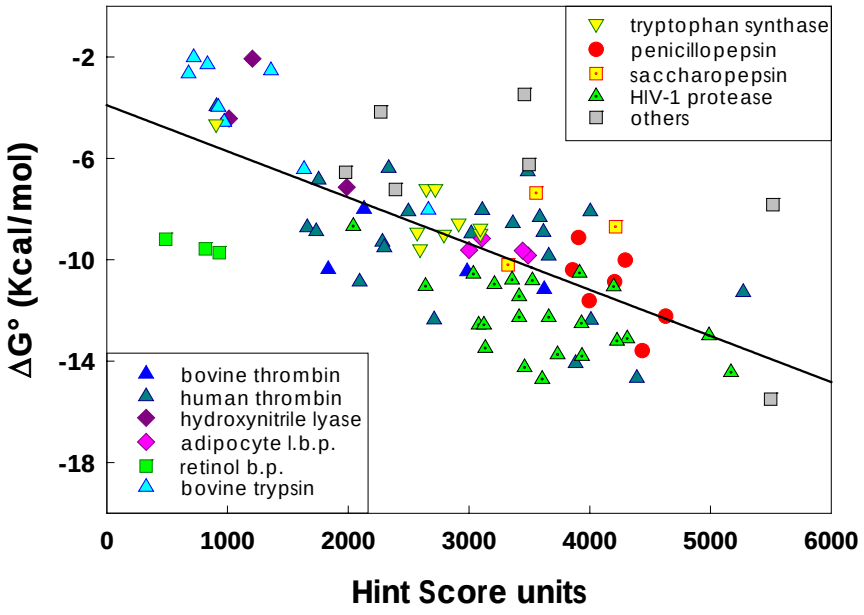


Fig. 5 Plot of experimental $\Delta G°$ vs. HINT score units, for a set of 93 different crystallographic protein–ligand complexes.

(**N.B.** The results obtained with the current 2.35S+ HINT version might be slightly different from those reported in Fig. 4, calculated with the previous 2.35S version.)

STEP 5 (Changing the protonation state of ionizable interacting residues)

In the 1a30 binding pocket, eight different ionizable group are present, respectively, four on the protein residues and four on the peptidic ligand. You can easily identify the carboxylic groups of Asp25, Asp125, Asp30, Asp29 on the protein, of Leu508, Asp507, Glu506 on the ligand and the amino group of Glu506.

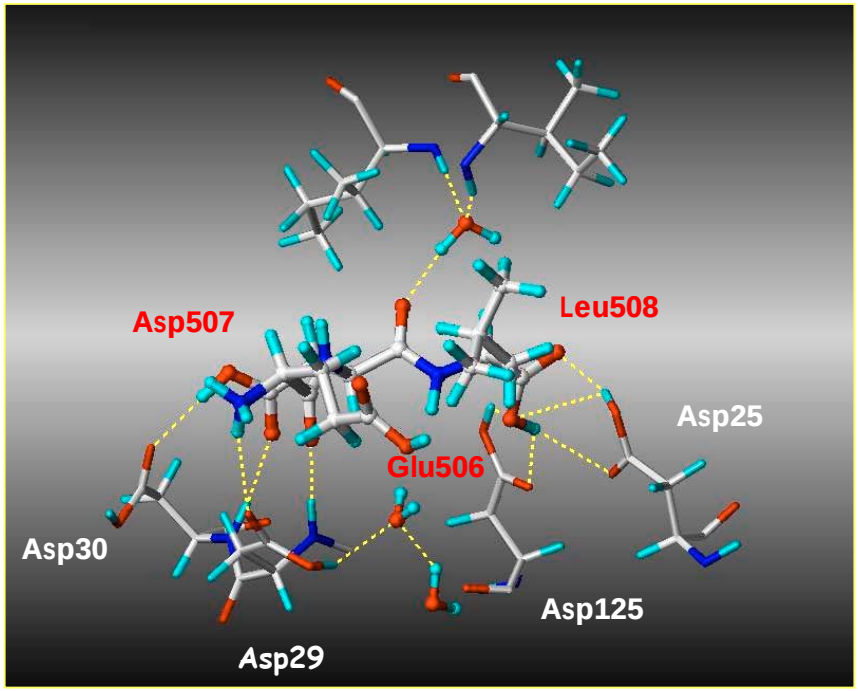
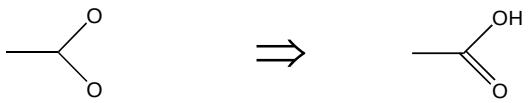


Fig. 6 1a30 binding pocket. All the ionixable groups are modeled protonated.

1. Choose to protonate one of the ionizable groups of the ligand, for example, the carboxylic one of Leu508.

- Open the ligand.mol2 file previously generated in Spartan.
- Identify Leu508.
- Change the atom types of the carboxylic group and add a hydrogen.



- Save the modified ligand as ligandprot1.mol2.
- Read the protein.mol2 and the ligandprot1.mol2 files in HINT.
- Calculate the log *P* of both protein and ligand.
- How does the log *P* value change after the Leu508 carboxylic group protonation?

…… ⇒ ……

- Calculate the HINT score of the new protein–ligand complex.
- How does the HINT score change?

…… ⇒ ……

- Which of the two complexes seem to be more favorable and stable? Why?

2. Choose to protonate one of the ionizable groups of the protein residues, for example, the Asp30 carboxylic one.

- Open the protein.mol2 file previously generated in Spartan.
- Identify Asp30.
- Change the atom types of the carboxylic group and add a hydrogen.
- Save the modified protein as proteinprot1.mol2.
- Read the proteinprot1.mol2 and the ligand.mol2 files in HINT.
- Calculate the log P of both protein and ligand.
- How does the ligand log *P* value change?

…… ⇒ ……

- Calculate the HINT score of the complex.

- How does the HINT score change?

…… $\Rightarrow$ ……

- Compare this result with the previous one obtained after the protonation of Leu508. Which is the more stable complex?

- Try to compute the HINT score between proteinprot1.mol2 and ligandprot1.mol2 and observe the variation

- Starting from proteinprot1.mol2 and ligandprot1.mol2, choose to modify the protonation state of other ionisable groups, adding one proton at a time.
- Calculate the HINT score from each new complex.
- Report all the data in Table 5.
- Compare the difference log *P* and HINT score values between non-protonated, mono-protonated, bi-protonated, etc., models.

Table 5

protein file	ligand file	protein log *P*	ligand log P	HINT score
protein.mol2	ligand.mol2			
protein.mol2	ligandprot1.mol2			
proteinprot1.mol2	ligand.mol2			
proteiprot1.mol2	ligandprot1.mol2			
proteinprot1.mol2	ligandprot2.mol2			
proteinprot2.mol2				

- **Which is the more favorable state and why?**

STEP 6 (Evaluating the water role in protein–ligand binding)

All HIV-1 Protease-peptidic inhibitor complexes are characterized by the presence of a conserved water molecule known as water 301.

Water 301 forms four hydrogen bonds, respectively, two with Ile50 and Ile150 on the protein and two with the ligands (Fig. 6), and its presence is fundamental for the formation and the stabilization of the protein–ligand complexes.

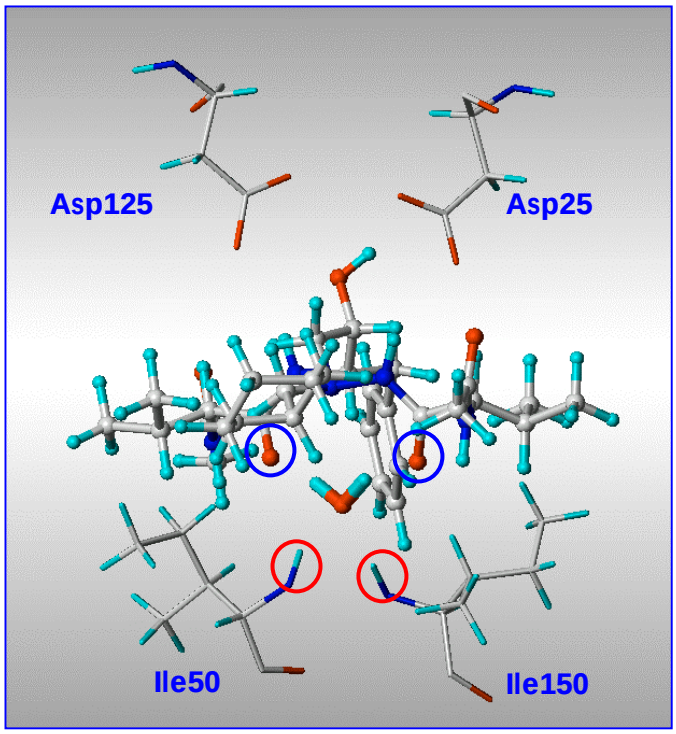


Fig. 6 1hih binding pocket. Water 301 and ligand Cgp 53820 are represented with ball and stick while the protein residues are shown with capped stick. Red circles highlight the two H-bond donators groups present on Ile50 and Ile150, while blue circles identify the two H-bond acceptors carbonilic groups located on the inhibitor.

- Connect to the Protein Data Bank and download another HIV-1 protease-ligand complex identified by the PDB code 1hih.
- Read 1hih.pdb in Spartan.
- Identify the inhibitor Cgp 53820 and water 301 placed at the complex interface.
- Extract the inhibitor from the complex and save it as ligand2.mol2.
- Delete all water molecules from the protein except water 301.
- Extract water 301 and save it as water301.mol2.
- Save the protein as protein2.mol2.
- Export protein2.mol2, ligand2.mol2, and water301.mol2 files in HINT.
- Calculate the log *P* of protein, ligand and water 301.

- Calculate the HINT scores between protein and ligand, protein and water, and ligand and water. (Being water 301 conserved in several complexes and being considered as an extension of the protein binding pocket, the total HINT score is represented by the sum of the protein–ligand and ligand–water contributions.)
- Now you can repeat the calculations downloading other HIV-1 protease–ligand complexes from PDB.
- Observe the variation of the protein–ligand and ligand–water HINT scores and report the data in Table 6. Compare them with those reported in Table 4 in the theoretical section and try to plot the experimental data vs. the computational results, considering and omitting the water contribution.

Table 6

PDB code	HINT score protein–ligand	HINT score ligand–water	HINT score TOTAL
1hih			
1a30			

Pietro Cozzini

Laboratory of Molecular Modelling, Department of General and Inorganic Chemistry, Chemical-Physics and Analytical Chemistry, University of Parma, 43100 Parma, Italy

High standards in safety measures should be maintained in all work carried out in Medicinal Chemistry Laboratories.

The handling of electrical instruments, heating elements, glass materials, dissolvents and other inflammable materials does not present a problem if the supervisor’s instructions are carefully followed.

This document has been supervised by Proff. Pietro Cozzini,Francesca Spyrakis (pietro.cozzini@unipr.it; fspyraki@nemo.unipr.it) who have informed that no special risk

(regarding toxicity, inflammability, explosions), outside of the standard risks pertaining to a Medicinal Chemistry laboratory exist when performing this exercise.

If your exercise involves any “special” risks, please inform the editor.