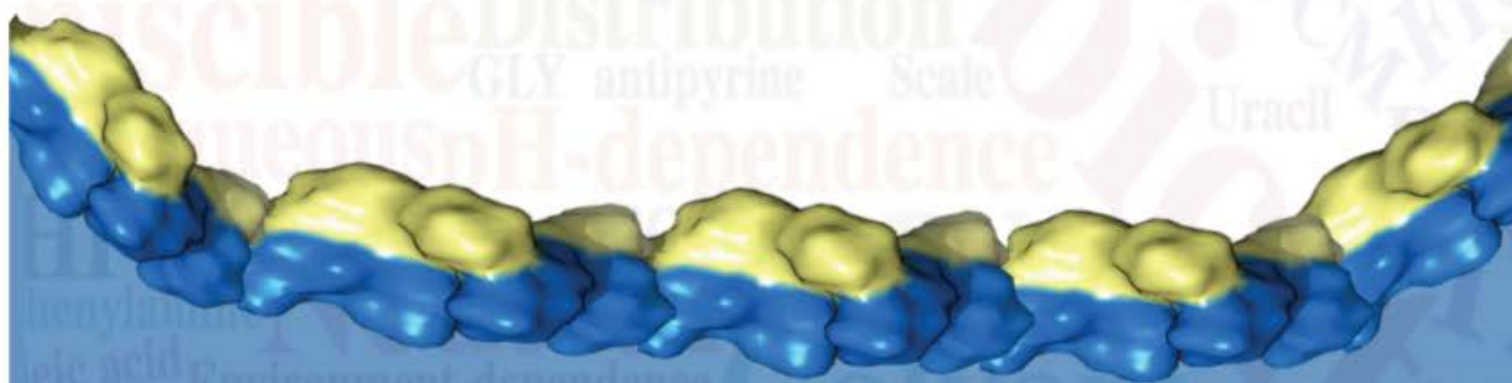# Toward Refined Theoretical Models for the Description of Lipophilicity in Biomolecules

William J. Zamora R.

Doctoral Thesis
Barcelona, 2019

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

PROGRAMA DE DOCTORAT DE BIOTECNOLOGIA

Toward Refined Theoretical Models for the Description of Lipophilicity in Biomolecules
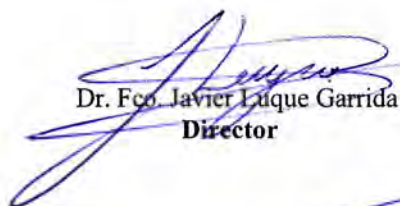
WILLIAM J. ZAMORA RAMÍREZ

BARCELONA, 2019

UNIVERSITAT DE BARCELONA

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

PROGRAMA DE DOCTORAT DE BIOTECNOLOGIA

Toward Refined Theoretical Models for the Description of Lipophilicity in Biomolecules

Memoria presentada por William J. Zamora Ramírez para optar al título de Doctor por la Universidad de Barcelona.

Dr. Fco. Javier Luque Garrida
Director

Dr. Josep Maria Campanera Alsina
Director

William J. Zamora Ramírez
Doctorando

Dr. Josep Maria Campanera Alsina
Tutor

WILLIAM J. ZAMORA RAMÍREZ

BARCELONA, 2019

A mi familia con todo mi amor
À minha família com todo meu amor

Acknowledgments

Barcelona-Brasil-Portugal-El Mundo

Quero agradecer aos amigos queridos que fiz neste período de estudos, e que considero como minha família portuguesa, meu grande amigo Manuel, Mamãe Peu, Sr. Prior, Dona Lena, Joelma, Rui Melro e Carolzinha, obrigado sempre por seus bons conselhos e pelos maravilhosos momentos juntos. A Cira aqui em Barcelona por ser nossa fotógrafa e sempre estar conosco. A minha família brasileira, meu sogro Manoel, minhas cunhadas Silvane e Simone por serem sempre tão amorosas. Meus sobrinhos: Lara, bebezinha titio te ama, Kauã, Kauê, Juninho e Marllon, todos maravilhosos. Acivaldo, Lene, Silvano, Cielly e Milson, Tio Ailton obrigado por toda sua hospitalidade quando me conheceu. Em especial quero agradecer a minha querida sogra Marlete por tratar-me como um filho, cuidar da nossa família da melhor maneira, sem ela definitivamente terminar este trabalho teria sido quase impossível, obrigado sogra és uma mulher incrível.

Barcelona

Continuaram anos de trabalho, físicoquímica, solvatação, lipofilicidade, programação e de repente aquela linda brasileira que mencionei antes lembram? A que foi minha amiga e confidente por muitos anos, me transformou por completo e se converteu em minha companheira de vida. Agora falo português "*caraca*" pois este trabalho sem ela não teria sido o mesmo. Obrigado meu amor por ser essa fonte de inspiração que necessitava em minha vida, além de admirar-te como científica, te admiro como mulher. Sei tudo o que lutaste para conseguir o que eres hoje, obrigado por estar aqui sempre, nos momentos que mas necessitava, por escutar-me e ter paciência, por crer em mim e por dar-me força, espero não decepciona-la e que sinta orgulhosa de mim. Esta conquista não seria o mesmo sem o mais valioso que consegui e me atrevo a dizer conseguimos, nossa *pithukinha* Iara. Iara llegaste cuando más te necesitábamos princesa, fue lindo escribir este trabajo los tres juntos, solo difícil amor cuando me cambiabas algún párrafo o me borrabas alguna ecuación pero ya algún día amor espero enseñarte eso que hace papai en la computadora que por lo que veo te gusta. Eres una princesa y nunca dejes de mirar mis ojos, ahí veras siempre todo mi amor por ti y por mamãe. Gracias mis mujeres hermosas, por hacer de esta etapa de sueño una realidad, esto es para ustedes mis amores, las amo y amaré eternamente, papai.

"Así caí, mi reto a pensar"

Summary

Lipophilicity is a key physicochemical descriptor used to understand the biological profile of (bio)organic compounds, xenobiotics and a broad variety of biochemical, pharmacological, and toxicological processes. This property is estimated from the partition coefficient between aqueous and nonaqueous environments for neutral compounds (P_N) and corrected for the pH-dependence of ionizable compounds as the distribution coefficient (D). In this context, in this doctoral thesis the Miertus–Scrocco–Tomasi continuum solvation model was used to check the suitability of some reported and proposed formalisms to estimate the distribution coefficient for a set of small acidic and basic compounds. The results indicate that in general the simple pH-dependence model of the ionizable compound in water suffices to predict the partitioning at or around physiological pH. However, at extreme pH values, where ionic species are predominant, more elaborate models provide a better prediction of pH-dependent distribution curves of $\log D$ for both acidic and basic compounds as well as for amino acid analogues. New theoretical treatments for the lipophilicity profile of ionizable compounds were proposed to account for the electroneutrality in the phases of the *n*-octanol/water system. In this context, was used the theory of ion-transfer across the interface between two immiscible electrolyte solutions (ITIES). Experimental research is being carried out to see the scope of those formalisms developed in this thesis.

Taking advantage of the successful results in small compounds, a lipophilicity scale adapted to different pH conditions was built for the 20 natural amino. The environment-dependence was introduced from the Dunbrack's backbone-dependent conformational library using two weighting schemes for the rotamers: *solvent-like* (SolvL) and *protein-like* (ProtL) lipophilic schemes. The veracity of our scale was corroborated with successful correlations with other consolidated experimental scales. Characterization of short disordered peptides (retention times in RP-HPLC, $\log P_N$ and $\log D_{7.4}$ values) was best described using the former approach, and biological properties of peptides with available three-dimensional

structure (local context-dependent lipophilicity *e.g* binding free energies) with the second one. Our theoretical lipophilicity scale was thus characterized by its versatility and adaptability, which confers a unifying character. Future studies will address the application of this methodology to the calculation of lipophilic parameters for no proteogenic amino acids, other conformations of the actual residues (proline *cis*) and other fragments relevant to proteins. On the other hand, the applicability of the present versatile scale is vast and promising, including for instance the use as scorings for protein-protein docking protocols, among others.

Resumen

La lipofilicidad es un descriptor fisicoquímico clave utilizado para comprender el perfil biológico de los compuestos (bio)orgánicos, xenobióticos y una amplia variedad de procesos bioquímicos, farmacológicos y toxicológicos. Esta propiedad se estima a partir del coeficiente de reparto entre ambientes acuosos y no acuosos para compuestos neutros (P_N) y corregido para la dependencia del pH de los compuestos ionizables como el coeficiente de distribución (D). En este contexto, en esta tesis doctoral se usó el modelo de solvatación continua de Miertus-Scrocco-Tomasi para verificar la idoneidad de algunos formalismos reportados y propuestos para estimar el coeficiente de distribución para un conjunto de pequeños compuestos ácidos y básicos. Los resultados indican que, en general, el modelo simple de dependencia del pH del compuesto ionizable en agua es suficiente para predecir la partición en o alrededor del pH fisiológico. Sin embargo, a valores extremos de pH, donde predominan las especies iónicas, los modelos más elaborados proporcionan una mejor predicción de las curvas de distribución dependientes del pH de log D tanto para compuestos ácidos como básicos, así como para análogos de aminoácidos. Se propusieron nuevos tratamientos teóricos para el perfil de lipofilicidad de compuestos ionizables para explicar la electroneutralidad en las fases del sistema *n*-octanol/agua. En este contexto, se utilizó la teoría de la transferencia de iones a través de la interfase entre dos soluciones de electrólitos inmiscibles (ITIES por sus siglas en inglés). Se están llevando a cabo investigaciones experimentales para ver el alcance de los formalismos desarrollados en esta tesis.

Aprovechando los resultados exitosos en pequeños compuestos, se construyó una escala de lipofilicidad adaptada a diferentes condiciones de pH para los 20 aminoácidos naturales. La dependencia del entorno se introdujo a partir de la biblioteca conformacional dependiente del *"backbone"* de Dunbrack utilizando dos esquemas de ponderación para los rotámeros: el esquema lipofílico tipo solvente (*SolvL*) y tipo proteíco (*ProtL*). La veracidad de nuestra escala se corroboró con correlaciones exitosas con otras escalas experimentales ya consolidadas. La caracterización de péptidos cortos desordenados (valores de tiempos de retención en

"*RP-HPLC*", log P_N y log $D_{7.4}$) fue mejor descrita utilizando el primer esquema, y las propiedades biológicas de los péptidos con estructura tridimensional disponible (lipofilicidad dependiente del contexto local y energías libres de unión) con la segunda. Nuestra escala teórica de lipofilicidad se caracterizó por su versatilidad y adaptabilidad, lo que le confiere un carácter unificador. Los estudios futuros abordarán la aplicación de esta metodología al cálculo de parámetros lipofilicos para aminoácidos no proteogénicos, otras conformaciones de los residuos actuales (prolina *cis*) y otros fragmentos relevantes para las proteínas. Por otro lado, la aplicabilidad de la escala versátil actual es amplia y prometedora, incluyendo, por ejemplo, el uso como ponderantes para protocolos de acoplamiento de proteína- proteína, entre otros.

General Index

Index

Abbreviations

ADMET	Absorption, Distribution, Metabolism, Excretion and Toxicity
B3LYP	Becke Three-Parameter Lee-Yang-Parr Exchange-Correlation Functional
D	Distribution Coefficient
EC_{50}	Half-Maximal Effective Concentration
$Hphi$	Hydrophilicity
$Hpho$	Hydrophobicity
HPLC	High-Performance Liquid Chromatography
IC_{50}	Half-Maximal Inhibitory Concentration
IEF	Integral Equation Formalism
IEFPCM/MST	Integral Equation Formalism of the Polarizable Continuum Model/ Miertus-Scrocco-Tomasi Solvation Model
ITIES	Interfaces between Two Immiscible Electrolyte Solutions
IUPAC	International Union of Pure and Applied Chemistry
K_d	Dissociation Constant
K_i	Inhibitor Constant
K_{IP}	Constant Formation of the Ion Pair
Lip	Lipophilicity
$LipE$	Lipophilic Efficiency
MD	Molecular Dynamics
MHC	Major Histocompatibility Complex
MST	Miertus-Scrocco-Tomasi
PCM	Polarizable Continuum Model
PDB	Protein Data Bank
P_N	Partition Coefficient
ProtL	Protein-like Lipophilicity
QM	Quantum Mechanics
QM-SCRF	Quantum Mechanics – Self-Consistent Reaction Field
QSAR	Quantitative Structure–Activity Relationship
RP-HPLC	Reversed Phase-High Performance Liquid Chromatography
SASA	Solvent Accessible Surface Area
SCRF	Self-Consistent Reaction Field

SolvL	Solvent-like Lipophilicity
ssNMR	Solid State Nuclear Magnetic Resonance
TLC	Thin-Layer Chromatography
UV/VIS	Ultraviolet–Visible Spectroscopy
3D	Three-Dimensional

1

Chapter

Introduction

1. INTRODUCTION

The physicochemical characterization of various types of compounds, including (bio)organic compounds and xenobiotics, is of utmost significance in environmental, biochemical and pharmaceutical research, because it covers diverse areas in drug discovery and development, such as absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, quantitative structure-activity relationships (QSAR), molecular recognition, and guidelines for agrochemicals. For such purpose, both experimental and theoretical techniques have been constantly improved in recent years to deliver a more detailed description of molecular properties, especially those related to lipophilicity.

Lipophilicity (*Lip*), expressed as the differential solubility of solutes in aqueous and nonaqueous (organic) environments, is regarded as the most important and used physical chemistry descriptor to quantify this property. Thus, the *n*-octanol/water system has served as standard method to quantify lipophilicity in both theoretical[1–4] and experimental methods.[5–7]

This doctoral thesis tackles the computation of lipophilicity by computing the free energy of solvation in both, water and *n*-octanol using the version of the implicit solvation model IEFPCM/MST parametrized in Barcelona. The response in the last solvent was further calibrated in this work for nitrogen-containing heterocyclic molecules as well as a variety of ionic compounds. Also, the theoretical models of lipophilicity were refined and proposing a general formalism where the theory of ion-pairing and the Galvani potential in the interphase of the immiscible solvents were considered. The performance of the refined lipophilicity models was tested by calculations of the distribution coefficient to physiological conditions for a set 35 ionizable compounds. Further, calculations were extended to several drugs and amino acid analogues, which were considered to examine the pH-dependent lipophilicity profiles. Finally, taking advantage of the successful results for computing the distribution coefficients to physiological conditions as well as for

reproducing lipophilicity profiles in amino acid analogues, a new lipophilicity scale was developed for amino acids, which incorporate the effect of the pH but also an environment-dependence by using two weighting schemes for the rotamers given in the Dunbrack's backbone-dependent conformational library. From those schemes were derived the *Solvent-like* (SolvL) and the *Protein-like* (ProtL) lipophilic approaches.

This dissertation, therefore, is structured in various chapters. In Chapter 1, the history of the emergence of partition schemes as descriptors of lipophilicity as well as their state of the art until today are described. Also, the most common theoretical and experimental methodologies to calculate/measure the lipophilicity are reviewed. Finally, besides the application to small molecules, the impact of lipophilicity in the field of amino acid through the so-called "*hydrophobicity scales*" is discussed. Chapter 2 points out the main objective as well as the specific aims of this thesis. The results, in conjunction with new data derived from ongoing work are described in Chapter 3, together with their discussion. The methodology, including the computational and experimental techniques employed in this thesis is described in Chapter 4. Finally, the main conclusions are given in Chapter 5 together with the future perspectives emerging from this doctoral thesis.

1.1 Lipophilicity

1.1.1 History

The differential solubility of solutes in aqueous and nonaqueous (organic) environments, known as lipophilicity, is a fundamental physicochemical property for understanding a wide range of biochemical, pharmacological, and toxicological processes of bioactive compounds.[4,5,7–13] This property has been estimated from the partition coefficient (P_N; eq 1) between aqueous (w) and nonaqueous environments, typically n-octanol (o), for a neutral compound (HX).

$$P_N = \frac{\left[\mathrm{HX}\right]_o}{\left[\mathrm{HX}\right]_w} \tag{1}$$

The first general description for the partition coefficient was presented in 1872 by Berthelot and Jungfleisch[14] and further elaborated for neutral species by Nernst in 1891.[15] This descriptor set the basis for the *lipoid theory of narcosis*, also known as *Meyer–Overton rule*, which stated that not structurally related narcotic compounds must be fat-soluble, its action is more pronounced in cells where lipids are vital (i.e. nerves), and the relative potency depends on its partition coefficient between water and a fatty system.[16,17]
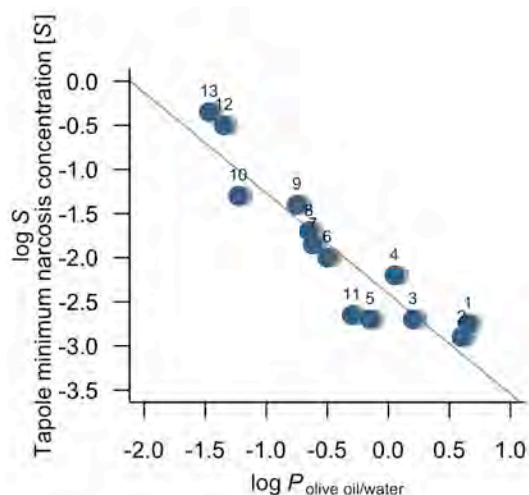
Figure 1. Correlation between the partition coefficients for narcotic compounds (trional (1), tetronal (2), butyl chloral hydrate (3), sulfonal (4), bromal hydrate (5), triacetin (6), diacetin (7), choral hydrate (8), ethyl urethan (9), monoacetin (10), benzamine (11), methyl urethan (12), ethanol (13)) in the olive oil/water system and the minimum narcosis concentration in tapoles.[18-20]

Figure 1 shows the original data that Meyer and Baum used to support their theory based on the partition coefficient between water and olive oil of thirteen narcotic compounds.[18-20] This theory showed that earlier hypothesis that related the narcotic potency of a compound with the number of ethyl groups and its susceptibility to form an active form[21], or to its lower solubility in water[22] were not correct.

The Meyer–Overton rule also explained the permeability of small molecules through membranes using the partition coefficient (eq 2) as long as the mechanism of transport was assumed to be simple diffusion.

$$P_M = \frac{P_N \cdot D_M}{d} \qquad (2)$$

where P_M denotes the permeability coefficient of a solute, P_N and D_M, stand for the partition and diffusion coefficient, respectively, for a solute, and d is the membrane thickness. At present, although violations to this rule have been reported, it continues having great acceptance.[23]

1.1.2 Definitions

In spite of being used in the literature, there is still ambiguity regarding the correct use of the terms lipophilicity and hydrophophicity[24]. According to the International Union of Pure and Applied Chemistry (IUPAC), lipophilicity *"represents the affinity of a molecule or a moiety for a lipophilic environment. It is commonly measured by its distribution behaviour in a biphasic system, either liquid-liquid (e.g., partition coefficient in 1-octanol/water) or solid-liquid (retention on reversed-phase high-performance liquid chromatography (RP-HPLC) or thin-layer chromatography (TLC) system)"*,[25] while hydrophobicity *"is the association of non-polar groups or molecules in an aqueous environment which arises from the tendency of water to exclude non-polar molecule"*.[26]

In this context, lipophilicity (*Lip*) is a more complete and general descriptor than hydrophobicity (*Hpho*), which in fact can be viewed as a part of lipophilicity, as noted in eq 3, which provides a qualitative expression for lipophilicity.[27]

$$Lip = Hpho + polarity + ionic\ interactions \qquad (3)$$

In addition to the above-described definition for lipophilicity, hydrophilicity (*Hphi*) can be defined as *"the tendency of a molecule to be solvated by water"*.[28] From these definitions, lipophilicity can be proposed as the balance betwen hydrophobicity and hydrophilicity, as schematically shown in Figure 2. Accordingly, the lipophilicity represents a balance between the factors that energetically favour affinity by apolar environments and those that do not. In other words, lipophilicity refers to the hydrophobicity of a molecule minus the penalty due to hydrophilic interactions with the polar environment.

Lipophilicity (Lip)

Hydrophobic (Hpho)   *Hydrophilic (Hphi)*

Lip = Hpho - Hphi
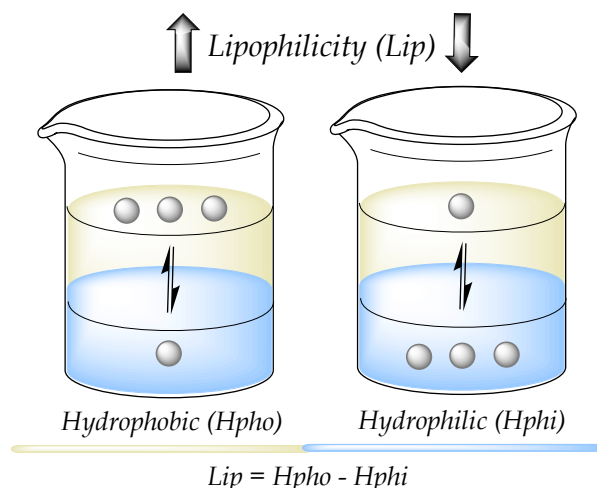
Figure 2. Lipophilicity (*Lip*) can be represented by the differential solubility of solutes in aqueous (blue) and nonaqueous (organic, yellow) environments. Lipophilic compounds (left) present mostly hydrophobic components that favor their preference towards oily environments, whilst poorly lipophilic compounds (rigth) are more likely to be in water because of the predominance of their hydrophilic features.

Despite the efforts made to clarify definitions as lipophilicity, hydrophobicity and hydrophilicity, the relationship between those concepts is still controversial and the division between them is definitely not easy to recognize.[24] This is illustrated in Table 1, which shows some definitions for the terms hydrophobicity and hydrophilicity. For instance, in surface science the differentiation between hydrophobic and hydrophilic surfaces rely on the static water contact angle θ, so that the former term is characterized by $\theta > 90°$ and the last one by $\theta < 90°$. Note that this separation is just given for a change in $2°$ and so, it is not free of controversy because using this angle crossover Teflon surface has been denominated as hydrophilic even when its repulsion by water is evident.[29] An improved definition using the receding angle (θ_R) was recently given by Law[30] (see Table 1). On the other, the distinction between hydrophobic and hydrophilic compounds have also been performed using the free energy of hydration, considering a threshold value of around -27 mcal/m² (see Table 1).[31]

Table 1. Examples of Various Types of Classification for the Terms Hydrophobic and Hydrophilic Used in Surface Science and Physical Chemistry Topics.

Field	Variable	Hydrophilic	Hydrophobic	Superhydrophobic
Surface Science[30]	static contact angle (θ)	$\theta < 90^o$	$\theta > 90^o$	$\theta > 150^o$
	receding angle (θ_R)	$\theta_R < 90^o$	$\theta_R > 90^o$	$\theta_R \geq 145^o$
Physical Chemistry[31]	ΔG_{hyd} (mcal/m^2)	< -27	> -27	NA^a

^{a} *NA*: Not Applicable

From a molecular point of view, the distinction between hydrophobic and hydrophilic properties is fundamental to understand a wide range of properties, such as the formation of micelles, ligand binding and protein folding.[32–35] The *"hydrophobic effect"*, in a thermodynamic perspective, depends on the solute size, the crossover length scale being close to 1 nm.[36–38] Entropy is considered to dominate the hydration of small hydrophobic molecules where water can keep the hydrogen-bond network around them. On the other hand, for large hydrophobic solutes there is a loss of hydrogen bonds, giving rise to an enthalpy-driven contribution that can conduct to aggregation favoured by van der Waals interactions between apolar molecules.[39] In this process, water-mediated interactions can be either attractive, as it has been usually accepted, or repulsive (hydrophobic solutes are driven apart), depending on the solute size, being attractive when the buried water-exposed area buried is larger than 1 nm^2.[34]

Overall, hydrophilicity, hydrophobicity and lipophilicity are physical chemistry descriptors closely related, but the precise understanding at the molecular level is still subject to debate.

1.1.3 Applications

Lipophilicity is a permissive physicochemical concept in the sense that several apolar environments can be adopted to quantify this descriptor (e.g. olive oil[18-20], fatty acids[40], alkanes[41] and cycloalkanes[42], *n*-octanol[43], membrane mimics[44-48] among many others[49]). However, since 1964 Hansch *et al*[43] set down the *n*-octanol/water system as a standard method to quantify lipophilicity (log P_N) in both theoretical[1-4] and experimental methods.[5-7] The relevance and impact of this work has been clearly reflected by Lipinski's rule of five[9] where the partition coefficient was crucial for determining the drug-likeness of compounds. Hence, the partition of a compound in two immiscible phases has represented an essential property for the prediction of ADMET properties.

P_N-based models are not able to explain the partition of ionizable compounds. Figure 3 depicts the composition of (bio)organic compounds with ionizable groups. For individual natural amino acids, 35% of them have an ionizable side chain, but this percentage can reach almost 100% if peptides and/or proteins are considered. Similarly, most drug-like compounds included in chemical libraries contain ionizable groups.[46,50] Therefore, distinct neutral and ionized species may exist at a given pH, and handling the acid/base properties of compounds has added value in fields like drug discovery[51] and in agrochemical studies.[52] For instance, it is well-known that the absorption of bioactive compounds is influenced by the pH changes along the human gastrointestinal tract, with a maximal absorption of weak acids in the jejunum (pH $\approx$ 4.5) and weak bases in the ileum (pH $\approx$ 8.0).[53] Similarly, herbicides with a pK_a < 5 are preferred because in this way relatively high concentrations of the herbicide can be achieved within the phloem sap.[52]
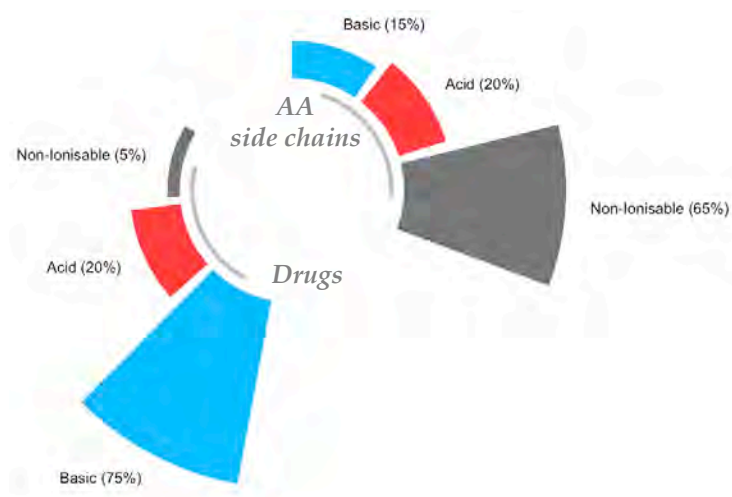
Figure 3. Overall composition of marketed drugs and amino acid side chains.

For an ionizable compound (HX) where both neutral and ionized species may exist at a given pH in aqueous solution, the total partitioning of the compound between aqueous and organic phases is better described by the distribution coefficient (D), which generally considers the equilibrium concentrations of the neutral and ionized forms.[54] Different formalisms have been proposed to derive the formal description of this descriptor, as will be detailed later in this thesis.

Following the considerations outlined above, it is expected that distribution coefficient works better as a lipophilic descriptor for ionizable compounds than the partition coefficient. In fact, drug-like compounds are characterized more efficiently using this descriptor (log $D_{5.5} \leq 5$)[11], chemicals with no concern for acute aquatic toxicity are classified with higher sensitivity using log $D_{7.4}$ (≤ 1.7)[55] and drug distribution among milk fat and skim milk are better predicted employing log $D_{6.8}$ than the pH-independent log P_N.[56–58]

The n-octanol/water distribution coefficient (D) is the most widely lipophilic descriptor used, as an inheritance given by the partition coefficient (P_N), and has a major impact in drug discovery. Figure 4 illustrates the classification of drug-like compounds according to the log D values[59] and their implications in drug development to physiological conditions (log $D_{7.4}$).[60] Thus, it is recommended to maintain log D values comprised between 0 and 3 in order to keep an acceptable

level of in vivo clearance, but also other ADMET properties as solubility, passive permeability and low metabolic liabilities are affected by log D[51,60], reflecting the impact in optimization of ADMET properties.[61]
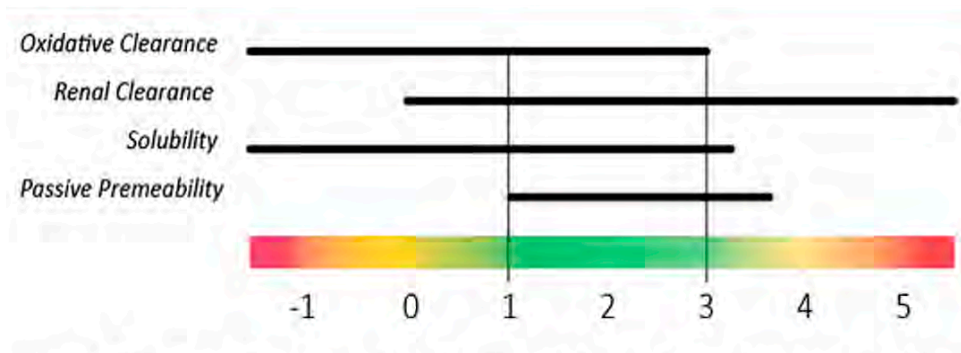
Figure 4. Proposed log D ranges as a guide to success in ADMET properties. [59]

It is worth noting that both n-octanol/water partition (P_N) or distribution coefficients (D) are simple surrogates of biological and/or chemical systems. According to Ribeiro et al.[62], they are rudimentary approaches to lipid vesicle-based methods, which are highly recommended as the best models for the study of lipophilicity[48,62]. However, Sugano et al[59] advocate that these descriptors can still be considered as the *"gold standard"* for lipophilicity due to the good correlation with oral absorption in rats, the fraction of a dose absorbed in humans, and the Caco2 membrane permeation for a varied set of up 500 drugs.

To date, the n-octanol/water system remains alive and this can be seen especially reflected in the research of the pharmaceutical industry sheltered in the concept of lipophilic efficiency (*LipE*; eq 4).[27,51,61,63–66]

$$LipE = -\log\left(potency\right) - Lip \qquad (4)$$

where *potency* can be represented by K_d (dissociation constant), which is usually replaced by K_i (inhibitor constant) or IC$_{50}$ (half-maximal inhibitory concentration) or EC$_{50}$ (half-maximal effective concentration), and *Lip* stands for lipophilicity, generally estimated from calculated or experimental n-octanol/water biphasic framework.[64]

LipE is a metric that normalizes the potency of a compound by its lipophilicity and explicitly considers the balance between these two variables.[27,51] It can be read as the difference between partition/distribution to a specific target (potency) and partition/distribution in a model system (lipophilic measurements).[61] It was conceived with the aim of contrasting different chemical series and assessing the impact in potency of small structural (lipophilic) modifications within series of compounds.[66] The usage of *LipE* in drug discovery and development has been imperative and recommended at all stages of discovery process. As a practical guideline, the best strategy is try to increase or at least maintain potency while lipophilicity is reduced.[27,67,68]

1.1.4 Theoretical Physicochemical Models of Lipophilicity Profile

The transfer free energy of neutral or ionic solutes between water and an organic phase ($\Delta\Delta G_{tr}^{w \to o}$), typically *n*-octanol, can be related to the difference in the solvation free energy upon transfer from the gas phase to the two solvents (ΔG_{hyd}^{w} and ΔG_{solv}^{o}; Scheme 1). Using this approach, it is possible to calculate the partition coefficient of a neutral (P_N) or ionic (P_I) compound.

Scheme 1. Thermodynamic Cycle Used to Determine the Transfer Free Energy of a Neutral (HX) or Ionic (X⁻) Compound between Two Immiscible Solvents.

$$\log P_{N,I} = -\frac{\Delta\Delta G_{tr}^{w \to o}}{RT \ln 10} = -\frac{\Delta G_{solv}^{o} - \Delta G_{hyd}^{w}}{RT \ln 10}$$
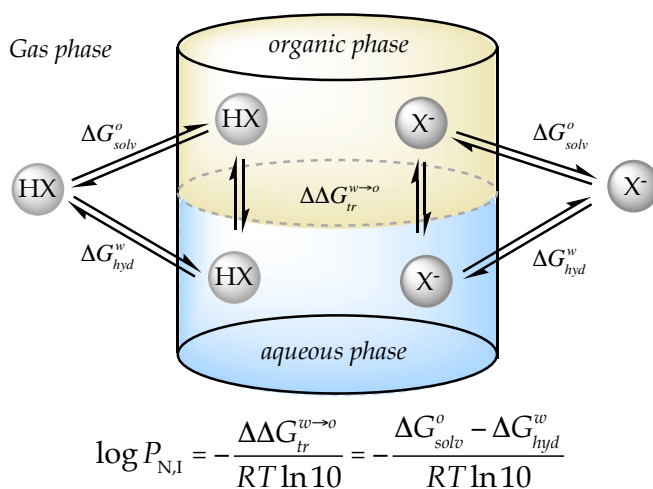
Since the equilibrium between neutral and charged species is a function of pK_a and pH in the aqueous phase, different theoretical formalisms have been proposed to estimate the distribution coefficient (D) for ionizable compounds. For the sake of simplicity, the focus here will be limited to the distribution coefficient for monobasic and monoacid compounds.

In 1940, Jacops[69] established the *pH-partition theory* by noting that the ionization in aqueous phase of weak electrolytes affected the cell permeability. This theory states that for an ionizable compound only its uncharged form can move through a cell membrane by passive diffusion. Hogben and collaborators[70,71] supported this theory in the 1950s from their studies of stomach and intestinal absorption of acidic and basic drugs, which partitioned preferably in conditions of low and high pH, respectively. Indeed, this theory illustrates the simplest and most widely used model to account for the pH dependence on the partition of ionizable compounds (Scheme 2).[54,71,80–84,72–79]

Scheme 2. Mechanism of *n*-Octanol/Water Partition for an Ionizable Neutral Compound (HX).

$$HX$$

$$P_N \quad \begin{array}{c} n\text{-octanol} \\ \sim\sim\sim\sim\sim\sim\sim\sim \\ \text{water} \end{array} \qquad D = \frac{[HX]_o}{[HX]_w + [X^-]_w}$$

$$HX \;\xrightleftharpoons[\text{pH}]{K_a}\; X^{\ominus}$$
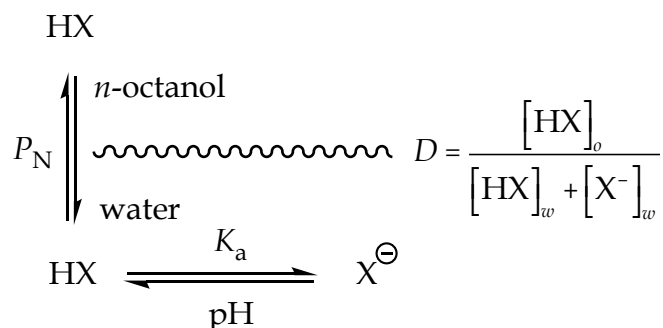
In this model, only the neutral species of an ionizable compound (HX) can partition between water and *n*-octanol, whereas both neutral and ionized species may exist in aqueous solution at a given pH. Under these circumstances, the distribution coefficient (D) of the compound between aqueous and organic phases depends on the pH of the aqueous solution, as noted in eq 5.

$$\log D = \log P_N - \log(1 + 10^{\delta}) \tag{5}$$

where $\delta = pH - pK_a$ for acids and $\delta = pK_a - pH$ for bases.

Nevertheless, the studies by Auerbach[85] and coworkers on the extraction of quaternary ammonium salts from aqueous solutions to organic solvents[74,86–89] demonstrated that ions can cross the interphase between water and an immiscible organic medium. Levine[90] used the same type of compounds to analyse its intestinal absorption and, even though it was poor, they could pass through the biological membranes. Furthermore, Winne and Högerle[91] showed deviations of the *pH-partition theory* in intestinal absorption curves of benzoic acid and aminopyrine, mainly due to the local pH effects but also to the partition of charged compounds. These facts suggest that explaining the partition of ionizable organic species in solvent models or the gastrointestinal absorption on the basis of *pH-partition theory* was not cogent.[72,92]

In agreement with the experimental evidence of the ion partition (P_I), more elaborate models have been proposed to refine the distribution model of ionizable compounds. The most straightforward correction comes from the assumption that a certain amount of the ionic species may also partition between water and *n*-octanol (Scheme 3).[5,44,99–106,49,53,93–98]

Scheme 3. Mechanism of *n*-Octanol/Water Partition for Both Neutral (HX) and Ionic (X⁻) Species
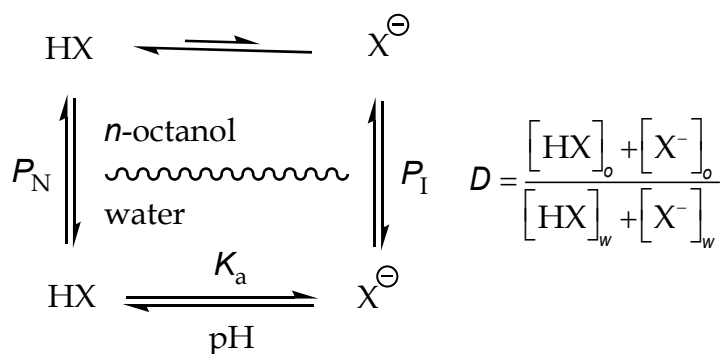
For a monoprotic acid (HX) the total partition of the solute can be expressed in terms of the partition constant of the neutral compound and of the ionic species (see Scheme 1), as noted in eq 6.

$$\log D = \log\left(P_N + P_I \cdot 10^\delta\right) - \log(1 + 10^\delta) \tag{6}$$

This approach challenges both experimental and computational chemistry due to the difficult of obtaining reliable values of partition of ionic species, which are usually largely hydrophilic. On the one hand, classical experimental methods such as shake-flask, pH-metric (potentiometric) and HPLC methods cover ranges of log P between -2, -1 and 0 (log P units), respectively.[107,108] In this regard, electrochemistry methods (lower log P range around -8)[98] are recommended. On the other hand, using thermodynamic cycles (see Scheme 1) challenges the suitability of theoretical models for estimating the differential solvation of ionic species with chemical accuracy. Thus, the solvation free energy of neutral solutes is generally determined from partition coefficients between the gas phase and aqueous solutions, and the experimental uncertainty increases with the solvation free energy, limiting the applicability of this technique to solutes with solvation free energy (in absolute terms) less than −12 kcal/mol.[109] Accordingly, ionic compounds require the use of indirect approaches based on the use of suitable thermodynamic cycles.[110–114]

In the partitioning schemes presented before, no account is made of the presence of the counter ion (C_I^+) for the dissociated organic species (X^-). The most accepted hypothesis states that the transfer of charged species are accompanied by counter ions, reflecting the formation of ion pairs ($X^-C_I^+$). For instance, Colaizzi and Klink[115] concluded that absorption of tetracycline antibiotics in the duodenum of the dog (where those compounds are completely ionized) occurred via the formation of a charge-neutral ion pair. In this context, Scheme 3 may be modified to include the contribution arising from ion pairs (Scheme 4).

Scheme 4. Mechanism of *n*-Octanol/Water Partition for Neutral (HX), Ionic (X⁻) and Ion Pair ($X^-C_I^+$) Species.
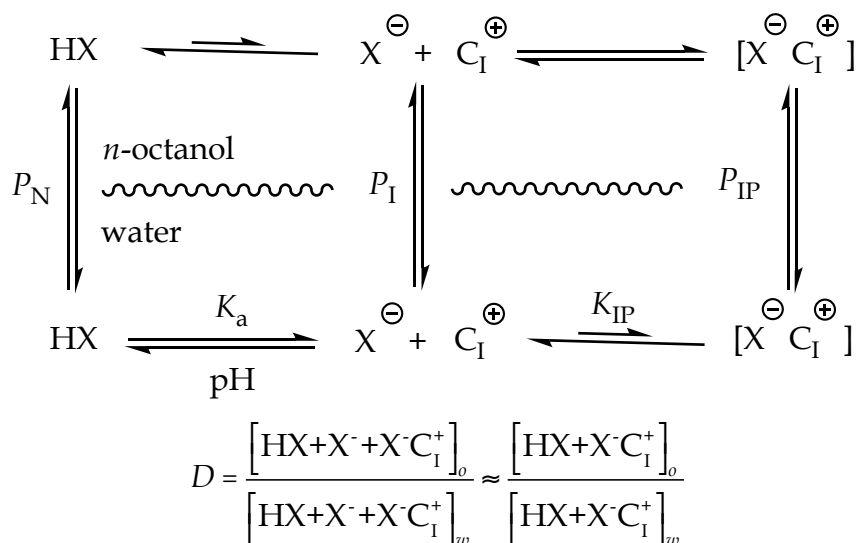
Although at low concentration of the compound the ion pair ($X^-C_I^+$) may dissociate at large extent in aqueous solution, the low permittivity of the organic phase may favour the formation of the ion pair.[116] This makes it necessary to account for the partitioning of this species, P_{IP} (eq 7), and for the formation constant of the ion pair (eq 8).

$$P_{IP} = \frac{\left[X^-C_I^+\right]_o}{\left[X^-C_I^+\right]_w} \tag{7}$$

$$K_{IP} = \frac{\left[X^-C_I^+\right]}{[X^-][C_I^+]} \tag{8}$$

According to Inagi et al.,[117] the log D of a compound can be written as a function of the partition coefficients of the neutral, ionic species, and of the ion pair (eq 9).

$$\log D = \log\left(P_N + \left(P_{IP} \cdot K_{IP} \cdot [C_I^+]\right)\cdot 10^\delta\right) - \log\left(\left(1 + K_{IP} \cdot [C_I^+]\right)\cdot 10^\delta\right) \tag{9}$$

Using the same mechanism exposed in the Scheme 4, Ingram et al.[118] used other considerations and proposed a expression to determine the log D as follows:

$$\log D = \log\left[\frac{P_N \cdot \left(10^\delta + 10^{\left(\delta + \log\left[C_I^+\right] - pK_{IP}\right)}\right) + P_{IP} \cdot 10^{\left(\log\left[C_I^+\right] - pK_{IP}\right)} + P_I}{1 + 10^\delta + 10^{\left(\delta + \log\left[C_I^+\right] - pK_{IP}\right)} + 10^{\left(\log\left[C_I^+\right] - pK_{IP}\right)}}\right] \qquad (10)$$

For practical purposes, however, the application of eq 10 has been largely limited not only by the scarce availability of accurate values of the partitioning constants (mainly P_I and P_{IP}), but also for the association constant of the ion pair (K_{IP}).[118] The most common approximation to this formalism considers that the partition of the ionic compound (P_I) is negligible[67,95,125,117–124] (see Scheme 4), and thus the distribution coefficient can be determined from eq 11.

$$\log D = \log\left(P_N + P_{IP} \cdot 10^\delta\right) - \log(1 + 10^\delta) \qquad (11)$$

1.1.5 Methods to Determine Lipophilicity

Conceptually, as detailed in the models shown in the previous section, in order to determine the lipophilicity for a given compound, it is necessary to measure/compute the partition of the neutral and ionic species of a given compound, the pK_a and takes into account the background salt used. Experimentally, these variables can be extracted from a lipophilic profile curve (apparent partition versus pH). There are several experimental approaches that can be classified as direct (shake-flask and potentiometric) and indirect (RP-HPLC) methods of measuring lipophilicity. Those methods have been reviewed and compared meticulously in the literature.[24,98,108,126] In addition to quantifying lipophilicity, they are the basis for development of high quality predictive in silico models[67] which are helpful at early stages of the drug discovery and development process but also for applications in food[40,58] and xenobiotic[52,55] either in academia, industry and/or government regulations. Some experimental methods will be reviewed in the next sections.

1.1.5.1 Experimental Methods

1.1.5.1.1 Direct: Shake-Flask Method

The shake-flask method (see Figure 5) consists of dissolving the analyte in the biphasic system, usually *n*-octanol and water, inside a test tube. Once both phases are mutually saturated, the system is shaken and left to rest for a few hours until reaching the partition equilibrium. Once the equilibrium between all interacting components is attained, an appropriate analytical method (e.g UV/VIS spectroscopy) is used to determine the concentration of substances dissolved in both phases.[107]
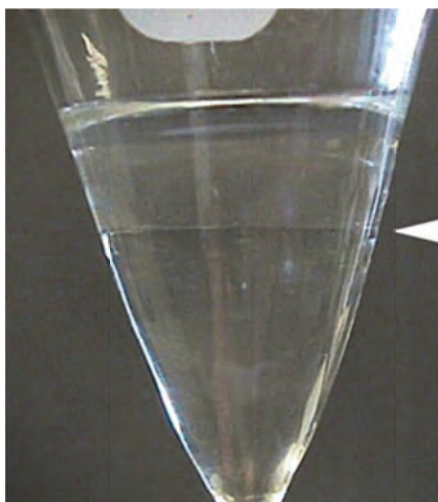
Figure 5. Separator funnel containing two immiscible solvents used for the shake-flask method.

The standard shake-flask measurement is a time-consuming, labour-intensive process, which allows partition/distribution coefficient determination in a narrow range of -2 to 4 (mainly determined by the analytical methods used for concentration measurements). The partition coefficient depends on the relative solubility of a substrate in a polar and nonpolar solvent, and the log P_N has to be corrected for ionization. Difficulties arise with very hydrophilic or very hydrophobic compounds, usually because of solubility issues, emulsion formation, and adsorption onto vessel walls. Another problem concerns compounds that can have tautomerism equilibria or be affected by the conversion between zwitterion and neutral forms.[127]

Using this method is possible to obtain a lipophilicity profile by measuring the partition in a range of pH values. This classical technique is still used as a benchmark method against which other methods are validated.[98]

1.1.5.1.2 Direct: Potentiometric Method

Potentiometric methods are based on the measurement of the activity of ions from the potential of an ion-selective membrane electrode (Figure 6).

Figure 6. Equipment to perform pH-metric and pK_a experiments used in the determination of lipophilicity of compounds. (Sirius Analytical http://www.sirius-analytical.com)

This method is useful for compounds with ionizable groups where pH metric titrations can be performed. Here, the difference in the pK_a values in the two immiscible solvents is exploit to estimate the partition coefficients. Using this methodology is possible to determine partition of neutral and ion-paired compounds[127] inside a range of -1 to 8 (log P units).[98]

In order to obtain both, the log P_N and the apparent log P_I (usually associated to the log P_{IP}) it is necessary to perform sequential assays[94]. Accordingly, one assay would be made without n-octanol, obtaining the pK_a, and the next ones would have differents amounts of this solvent, obtaining n values for $p_oK_a(n)$. Then, in the case of performing three assays ($n = 2$), the partition of the neutral and ionic specie can be obtained using the eq 12 and 13, respectively.

$$P_{\mathrm{N}} = \frac{r_2 \cdot 10^{\left(\mathrm{P}_o K_a(2) - \mathrm{p}K_a\right)} - r_1 \cdot 10^{\left(\mathrm{P}_o K_a(1) - \mathrm{p}K_a\right)} - \left(r_2 - r_1\right) \cdot 10^{\left(\mathrm{P}_o K_a(1) + \mathrm{P}_o K_a(2) - 2\mathrm{p}K_a\right)}}{r_2 \cdot r_1 \cdot \left(10^{\left(\mathrm{P}_o K_a(1) - \mathrm{p}K_a\right)} - 10^{\left(\mathrm{P}_o K_a(2) - \mathrm{p}K_a\right)}\right)} \qquad (12)$$

$$P_{\mathrm{I}} = \frac{r_1 \cdot 10^{\left(\mathrm{P}_o K_a(2) - \mathrm{p}K_a\right)} - r_2 \cdot 10^{\left(\mathrm{P}_o K_a(1) - \mathrm{p}K_a\right)} + r_2 - r_1}{r_1 \cdot r_2 \cdot \left(10^{\left(\mathrm{P}_o K_a(1) - \mathrm{p}K_a\right)} - 10^{\left(\mathrm{P}_o K_a(2) - \mathrm{p}K_a\right)}\right)} \qquad (13)$$

1.1.5.1.3 Indirect: High Performance Liquid Chromatography Methods

High-performance liquid chromatography (HPLC) is a chromatographic method that is used to separate and quantify components of a mixture of compounds in analytical chemistry and biochemistry. The differential elusion time of each compound directly relates to the compound partition between the mobile and the stationary phases. The retention factor (k) can be related to the partition coefficient of the compound between the mobile and the stationary phase, according to eq 14.

$$\log k = \log\left(P_{\mathrm{N}} / D\right) + \log\left(\frac{V_s}{V_m}\right) \qquad (14)$$

where (V_s/V_m) represents the ratio of the stationary and mobile phases to obtain the absolute value of the chromatographic partition coefficient.

There are two major approaches for investigation of lipophilicity using HPLC: isocratic and gradient methods.[126] In the first ones, previous estimation of probable compound lipophilicity and measurements at different mobile phase concentrations are needed.[126,128] The gradient method consists in programmed increase during the chromatographic process of organic solvent in the aqueous mobile phase. Thus, the pool of compounds is eluted of the column (stationary phase) allowing measurements in a range of 0 to 5 of logarithm partition coefficients. In most of cases, impurities rarely affect results and simultaneous analysis of several substances are possible.[129]

1.1.5.2 Theoretical Methods

In addition to experimental methods for quantifying lipophilicity, an extensive variety of computational approaches for estimating this descriptor have also been developed.[1-3,130] These approaches can be classified in substructure and property-based methods, which encompass a wide variety of formalisms, from atomic (e.g ACD/log$D$[131] and ChemAxon/log$D$[132]), fragmental and molecular contributions to quantum mechanical (QM)-based strategies (e.g IEFPCM/MST[133-136] and SMD[137] models), and lately fashionable, machine learning methods. Theoretical methods are imperative when is wished to analyze a huge amount of compounds what would be experimentally unviable.

1.1.5.2.1 Substructure-Based Methods

Beyond simplicity and low expensiveness, fragment-based methods are able to dissect the lipophilicity of a compound, which is a property of the whole molecule, in empirical contributions of atoms and/or fragments. Accordingly, it can be utilized to gain insight into the molecular determinants that govern the interactions between bioactive molecules and receptors. Thus, within the framework of atoms/groups based methods for estimation of partition coefficients, fragmental and atom-based techniques follow a general additive scheme as show in eq 15.

$$\log P_{\mathrm{N}} = \sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{m} b_i F_i \tag{15}$$

where $\log P_{\mathrm{N}}$ is the sum of the weighted (a_i) contribution of each atom/fragment (f_i) and a correction factor ($b_i F_i$).

Leo and Hansch[5] exposed the first fragmental method, named cLOGP. It allowed to extrapolate $\log P_{\mathrm{N}}$ starting from a list of experimentally fitted fragmental contributions to lipophilicity. Then, Ghose[2] proposed the ALOGP method using a list of 120 atom types for carbon, hydrogen, oxygen, nitrogen, sulfur, and halogens.[138-140] Another version of these methods, the XLOGP[141] was based on the summation of

atomic contributions derived from experimental lipophilicity data of 1831 organic molecules, which includes correction factors for some intramolecular interactions.

1.1.5.2.2 QM-based Methods

The paramount importance of lipophilicity in molecular recognition justifies the efforts conducted to develop quantum mechanical (QM)-based strategies for the calculation of lipophilic descriptors. A straightforward strategy for the computation of lipophilicity/hydrophilicity patterns of molecules comes from QM self-consistent reaction field (SCRF) models, which rely on the description of the solvent as a continuum polarizable medium that reacts against the perturbing field created by the charge distribution of the solute.

The Miertus-Scrocco-Tomasi (MST) solvation model has been used to develop 3D distribution patterns of lipophilicity using log P as descriptor. The MST model is a parametrized version of the polarizable continuum model developed by Tomasi and coworkers[142,143] at both semiempirical, Hartree-Fock and B3LYP levels.[133–136] From the solvation free energies in water and n-octanol, one can derive the n-octanol/water partition coefficient, which is a property of the whole molecule. Nevertheless, by decomposing the solvation free energy into atomic contributions, one can obtain the 3D profile of lipophilicity from the corresponding atomic contributions to the log P. For a molecule (M) containing N atoms, this is achieved by decomposing the log P (or the corresponding transfer free energy, $\Delta G_{tr,M}^{w\to o}$) into electrostatic ($\log P_{ele,i}$), cavitation ($\log P_{cav,i}$) and van der Waals ($\log P_{vdW,i}$) components, which can be derived from the polar ($\Delta G_{ele,i}^{w\to o}$) and non-polar ($\Delta G_{cav,i}^{w\to o}, \Delta G_{vdW,i}^{w\to o}$) contributions to the solvation free energy (eqs 16 and 17)

$$\Delta G_{tr,M}^{w\to o} = \sum_{i=1}^{n} \Delta G_{tr,i}^{w\to o} = \sum_{i=1}^{n} \left(\Delta G_{ele,i}^{w\to o} + \Delta G_{cav,i}^{w\to o} + \Delta G_{vdW,i}^{w\to o} \right) \qquad (16)$$

$$\log P_{N,M} = \sum_{i=1}^{n} \log P_{N,i} = \sum_{i=1}^{n} \left(\log P_{ele,i} + \log P_{cav,i} + \log P_{vdW,i} \right) \qquad (17)$$

Partitioning of the electrostatic term into atomic contributions can be made resorting to a perturbation approximation of the coupling between the solute charge distribution and the solvent reaction field[144], leading to eq 18.

$$\log P_{ele,i} = \frac{1}{2} \left\langle \Psi^\circ \left| \sum_{\substack{k=1 \\ k \in i}}^{K} \frac{q_k^w}{|r_k - r|} - \sum_{\substack{l=1 \\ l \in i}}^{L} \frac{q_l^o}{|r_l - r|} \right| \Psi^\circ \right\rangle \tag{18}$$

where Ψ^o is the solute wave function in the gas phase, and K and L stand for the total number of reaction field charges in water (q_k^w) and n-octanol (q_l^o), located at positions r_k^w and r_l^o.

The atomic decomposition of the cavitation and van der Waals terms takes advantage of the linear dependence with the solvent-exposed surface of the atoms in the molecule (eqs 19 and 20).

$$\log P_{cav,i} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G_{P,i}^{w \to o} \tag{19}$$

$$\log P_{vdW,i} = \sum_{i=1}^{N} S_i \cdot \Delta \xi_i^{w \to o} \tag{20}$$

where $\Delta G_{P,i}^{w \to o} = \Delta G_{P,i}^{o} - \Delta G_{P,i}^{w}$, with $\Delta G_{P,i}$ being the cavitation free energy of atom i, $\Delta \xi_i^{w \to o} = \xi_i^o - \xi_i^w$, with ξ_i being the atomic surface tension, and S_i denotes the contribution of atom i to the total molecular surface (S_T).

Other IEFPCM approach widely used in the literature to compute solvation energies and thus, lipophilic descriptors, is the universal solvation model based on solute electron density (D) called SMD continuum solvation model. In this model the standard-state free energy of solvation (ΔG_{solv}°) is computed according to eq 21.

$$\Delta G_{solv}^{o} = \Delta G_{ENP} + G_{CDS} + \Delta G_{conc}^{o} \tag{21}$$

where ΔG_{ENP} denotes the electronic (E), nuclear (N), and polarization (P) terms of the solvation energy; G_{CDS} stands for the changes associated with the cavitation (C), dispersion (D) and local structure (S) of the solvent and ΔG_{conc}^{o} accounts for the concentration change between the gas-phase standard state and the liquid-phase standard state (this term is 0 for 1 mol/L or well 1.89 kcal/mol for 1 atm).[137] This model has been recently used to calculate the lipophilicity of cytisine[104] (nicotinic receptor partial agonist) and thiosemicarbazide derivatives[145] (drug candidates).

1.1.6 Lipophilicity in Amino Acids, Peptides and Proteins

Proteins and their constituents, amino acids residues, perform their task in diverse environments, from water-like to less polar environments. Solvation free energy, as a measure of the energy cost to transfer a molecule from vacuum to solvent, has been extremely useful to unravel the mechanism of protein folding to native structure, protein function and molecular recognition.[146] Derived magnitudes such as transfer free energies and thereby lipophilicity has also been used to describe the balance of a molecule to interact with different solvation environment. So, lipophilicity plays a crucial role in many chemical and biochemical events such as transport and distribution of biological molecules, solubility, molecular recognition, aggregation and protein folding among many other implications.[147] Recently, more than 7000 peptides are known and approximately 140 peptide drugs are currently being considered in clinical trials. In that sense, the rapid and accurate determination of their physicochemical properties is of vital importance in peptide drug discovery.[148] However, the quantification of peptide and protein lipophilicity presents a significant challenge since has a multidimensional nature that depend on the environment conditions like thermodynamic variables (temperature, concentration, pH, pressure), additives (salts, osmolytes) and even on residue sequence (primary structure), surface topography (secondary and tertiary structures) and size. [149,150]

As a consequence of the aforementioned, a manifold of lipophilicity scales have been developed, giving insight into the biological world using this descriptor.

1.1.6.1 Lipophilicity Scales of Amino Acids

Since the quantitative description of solvent accessible surface area (SASA) by Lee and Richards[151] to account for the interaction among proteins and solvent, and the solubility of amino acids in ethanol and dioxane by Nazaki and Tanford[152], several efforts have been made in order to quantify the stability to remove nonpolar amino acids from water to nonpolar environment. This phenomenon is called "*the hydrophobic effect*" and it is well known that this repulsive free energy between water and the nonpolar side chain of amino acids depends on the SASA of the latter. Table 2 reports a set of values for the free energy of transfer of nonpolar medium to water for different models, employing nonpolar side chain of amino acid analogues, amino acid or peptides models.[153-159]

Table 2. Values for the Free Energy of Transfer of Nonpolar Medium to Water for Nonpolar Side Chain of Amino Acid Analogues, Amino Acids or Peptides Models.

Reference	$\Delta\Delta G$ transfer from nonpolar medium to water per nonpolar accessible surface (cal/molÅ^2)
Chothia et al. (1974) [a]	22
Reynolds et al. (1974) [b]	23 ± 2
Fauchère-Pliska (1983) [c]	20.9 ± 2.5
Rose et al. (1985) [a]	18.9 ± 0.7
Einserberg-McLachlan (1986) [c]	16 ± 2
Wimley et al. (1996) [c]	22.8 ± 0.8
Moon-Fleming (2011) [d]	23

[a] By means of amino acid solubility in ethanol and dioxane, [b] hydrocarbons solubility in water, [c] partition between water and *n*-octanol of amino acid and/or peptide systems and [d] partition in a biological system.

Nevertheless, amino acid contains also ionizable side chains and thus not just hydrophobic interactions should be taking into account but also polar and ionic interactions. Again, the lipophilicity reappears as the appropriate descriptor to understand the preferences for certain environments (bulk solvent or buried in a protein) of these biomolecular building blocks.

Several methods have been proposed for the fast and reliable quantification of lipophilicity of peptides, but the so-called "*hydrophobic scales*" based on amino acid contributions has found specially acceptance. The hydrophobic scales present a relative ranking of hydrophobicity for each of the 20 natural amino acids using various experimental, statistical and theoretical measurements. Consequently, scales are usually classified as biological-based, knowledge-based or bulk-solvent-based (Table 3). As expected, those scales have been subjected to many reviews, as illustrated by Simm et al.[160], Peters et al.[161] and MacCallum et al.[162]

Table 3. Lipophilicity Contribution of 20 Coded Amino Acids (Including Two Tautomers for His) Expressed as log P/D Coefficients.

Residue	Scales								
	Bulk-Solvent Adapted Scales				*Biological-Based Scales*		*Knowledge-Based Scale*		*Consensus Scale*
	Fauchère-Pliska	Eisenberg-McLachlan	Hopp-Woods	Wimley et al.	Moon-Fleming	Hessa et al.	Janin et al.	Koehler et al.	Kyte-Doolittle
Ala	0.31	0.49	0.36	0.00	0.00	-0.08	0.30	0.12	1.31
Arg	-1.01	-1.53	-2.19	1.55	-2.71	-1.88	-1.40	-0.40	-3.28
Asn	-0.60	-0.44	-0.15	-0.42	-2.53	-1.50	-0.50	-0.36	-2.55
Asp	-0.77	-0.88	-2.19	-2.43	-2.15	-2.55	-0.60	-0.53	-2.55
Cys	1.54	0.28	0.73	0.26	-0.36	0.09	0.90	-0.01	1.82
Gln	-0.22	-0.16	-0.15	-0.42	-2.20	-1.72	-0.70	-0.34	-2.55
Gln	-0.64	-0.55	-2.19	-2.48	-1.20	-1.96	-0.70	-0.51	-2.55
Gly	0.00	0.00	0.00	0.10	-1.26	-0.54	0.30	0.15	-0.29
His(δ)	0.13	0.47	0.36	0.04	-3.47	-1.50	-0.10	-0.28	-2.34
His(ε)	0.13	0.47	0.36	0.04	-3.47	-1.50	-0.10	-0.28	-2.34
Ile	1.80	1.39	1.31	0.94	1.14	0.44	0.70	0.28	3.28
Leu	1.70	1.39	1.31	1.04	1.32	0.40	0.50	0.22	2.77
Lys	-0.99	-0.42	-2.19	1.18	-3.93	-1.98	-1.80	-0.66	-2.85
Met	1.23	1.75	0.95	0.61	0.55	0.07	0.40	0.15	1.39
Phe	1.79	1.68	1.82	1.32	1.61	0.23	0.50	0.34	2.04
Pro	0.72	0.88	0.00	0.02	1.11	-1.63	-0.30	-0.36	-1.17
Ser	-0.04	0.01	-0.22	-0.01	-1.34	-0.61	-0.10	-0.04	-0.58
Thr	0.26	0.38	0.29	0.06	-1.30	-0.38	-0.20	0.01	-0.51
Trp	2.25	1.90	2.48	1.53	0.28	-0.22	0.30	0.02	-0.66
Tyr	0.96	1.17	1.68	0.58	0.80	-0.50	-0.40	0.09	-0.95
Val	1.22	1.09	1.09	0.54	0.57	0.23	0.60	0.18	3.07

It is important to note that apart from the theoretical or experimental source of data, each scale gives a slightly distinct model of lipophilicity. For instance, in the bulk-solvent adapted scales Fauchère-Pliska[155] used partitioning of *N*-acetyl-*L*-amino-acid amides between *n*-octanol and water in a neutral pH. Similarly, Eisenberg-McLachlan and coworkers[163] build their scale using the summation of the atomic solvation parameter multiplied by the atomic solvent accessible surface area (SASA, for the amino acid X in Gly-X-Gly sequence in a extended conformation) for each atom in an amino acid. The atomic solvation parameter was taken, in fact, from the Fauchère-Pliska scale. Hopp-Woods scale[164] put special attention to charged amino acids by virtue of their role in antigenic determinants, since they are very common in these regions. Wimley et al. scale[165] also used the *n*-octanol/water transfer energy for a pentapeptide model (AcWL-X-LL) but employed a pH = 9.

On the other hand, biological scales are exemplified the studies of Kyte-Doolittle[166], Moon-Fleming[167] and Hessa et al.[168]. In the Kyte-Dolittles scale water-vapour transfer free energies in conjunction with the interior-exterior distribution of amino acid side-chains were considered to build the hydrophaty scale (consensus scale). Moon-Flemings scale was developed using a β-barrel system inside a membrane where the transfer free energy for a given was measured at pH 3.8. Hessa et al. scale is based on the recognition of artificial helices by the Sec61 translocon, thus it gives an estimate of the relative stability of a residue to be inserted into a cell membrane at physiological conditions.

Finally, knowledge-based scales are developed from statistical methods taken information from an average hydrophobicity values of amino acids in folded proteins (Koehler et al. scale[169]) or well from transfer energy from molar fraction of buried and accessible amino acids in proteins (Janin scale[170]).

To sum up, lipophilicity scales have been built for specific purposes. Though there is a significant degree of correlation between the most cited ones, there are conceptual differences, which give rise to discrepancies in the predicted lipophilicities, particularly at non-physiological pH values.

2
Chapter

Aim

2. AIM

The main objective is to refine the theoretical models of lipophilicity estimated as the *n*-octanol/water distribution coefficient in (bio)organic compounds and biomolecules. To this end, a general formalism where the theory of ion-pairing and the Galvani potential at the interphase of the immiscible solvents has been proposed. The models have been applied to a variety of small molecules, including ionizable monoacid and monobasic substances. For the specific case of amino acid analogues, a lipophilicity scale has been developed, which opens the way to explore the recognition and binding in peptides and proteins.

With this general aim, the specific objectives that encompass the research developed in this work are indicated as follows.

2.1 Testing Models for Lipophilic Profiles of (Bio)Organic Compounds.

The first aim is the refinement of the Miertus-Scrocco-Tomasi (MST) continuum solvation model, which relies on the integral equation formalism of the polarizable continuum model (IEFPCM), to account for the solvation free energy of nitrogen-containing heterocyclic molecules, as well as ionic compounds, in *n*-octanol.

Second, it also aims to develop a formalism for predicting the pH-dependent lipophilicity profile, taking into account the effect of counter ion accompanying the ionizable (bio)organic compound. In this context, the aim is to use the theory of ion-transfer across the interface between two immiscible electrolyte solutions (ITIES).

Finally, the refined MST model will be used to determine the lipophilicity profile in conjunction with different physicochemical models for the partition of ionizable compounds.

2.2 Exploring the Effect of Galvani Potential on the Lipophilicity Profile.

Our aim here is to revise the suitability of a general formalism, which includes the effect of both the Galvani potential and ion-pairing, for determining lipophilicity profiles of monoacid and monobasic compounds at different ionic strength conditions, and validated against experimental data.

2.3 Development of a Lipophilicity Scale for Amino Acid Residues.

The final aim is to develop a lipophilicity scale for the natural amino acids using implicit solvation calculations in *n*-octanol and water, which account for the structural (conformational) dependence of residues and adapted to pH conditions. As a potential application, attention will be placed to the analysis of peptides involved in Alzheimer`s disease for a better understanding of the relationship between lipophilicity and toxicity.

3
Chapter

Results and Discussion

3. RESULTS AND DISCUSSION

In this dissertation, we have first refined the parametrization of the MST model for neutral nitrogen-containing aromatic compounds but also for ionic compounds in *n*-octanol. Together with the free energy of solvation in water, these improvemenst have been exploited for the computation of partition coefficients of neutral and ionic compounds. Furthermore, we have studied different theoretical models of pH-dependent lipophilicity profiles based on the *n*-octanol/water distribution coefficient in (bio)organic compounds and biomolecules. In order to refine the formalisms reported in the literature, we have taken into account the effect of counter ion accompanying the ionizable (bio)organic compound using the theory of ion-transfer across the interface between two immiscible electrolyte solutions (ITIES). Thus, we have revisited a new formalism for predicting the pH-dependent lipophilicity profile. The suitability of different formalisms to estimate the distribution coefficient for a wide range of pH values has been examined for a set of small acidic and basic compounds.

For the sake of completeness, a general formalism, which combines the acid dissociation constant in water, pH, background salt and partition of neutral, ionic, and ion-pair species, is proposed and its validity is being tested experimentally.

From these initial studies, an extension of the research line has allowed us to obtain successful results for computing the distribution coefficients to physiological conditions as well as for reproducing pH-adapted lipophilicity profiles in amino acid analogues. Thus, we have elaborated a lipophilicity scale for the 20 natural amino acids from theoretical computations that take into account the structural dependence of the conformational preferences of amino acids as well as the influence of pH in order to provide a consistent description of pH-adapted lipophilicity profiles in peptides and proteins.

Two weighting schemes have been considered to derive *Solvent-like* and *Protein-like* lipophilicity scales, which have been calibrated by comparison with other experimental scales reported in the literature, as well as by examining properties such as the retention time of small peptides, and the recognition of antigenic peptides. Finally, the lipophilicity scale have been applied to the study of the differentiated toxicity of 11 Aβ42 peptides involved in Alzheimer`s disease.

3.1 The Miertus−Scrocco−Tomasi Model: Framework for Continuum Solvation Calculations.

In this thesis, the theoretical computation of solvation free energy in *n*-octanol and water was needed to achieve the objectives of the thesis. This is a considerable challenge since the model has to be capable of describing the specific interactions of the solute in the two solvents, as a preliminary requirement to the computation of the free energy of transfer (Scheme 1).

QM-SCRF continuum models have been one of the most powerful approaches that have succeeded in this context. Among these methods, the MST model has proven to be a robust approach due to the rigorous formalism utilized for describing electrostatic and non-electrostatic interactions, and to the precise parametrization against experimental data, including properties such as solvation free energies, partition coefficients, tautomerism equilibria, ionization, and solvent-induced spectral shifts (Scheme 5).

Within this strategy, the solvation free energy accounts for the reversible work necessary to transfer a molecule from gas phase to a specific solvent at constant temperature, pressure and concentration. This thermodynamic process can be divided into three contributions

$$\Delta G_{solv} = \Delta G_{ele} + \Delta G_{cav} + \Delta G_{vdW} \tag{22}$$

where ΔG_{cav} denotes the cavitation free energy, ΔG_{vdW} is the van der Waals term (considering dispersion and repulsion interactions), and ΔG_{ele} is the electrostatic free energy component.

Scheme 5. Miertus-Scrocco-Tomasi (MST) Model: Framework for Continuum Solvation Calculations (ΔG_{solv}).
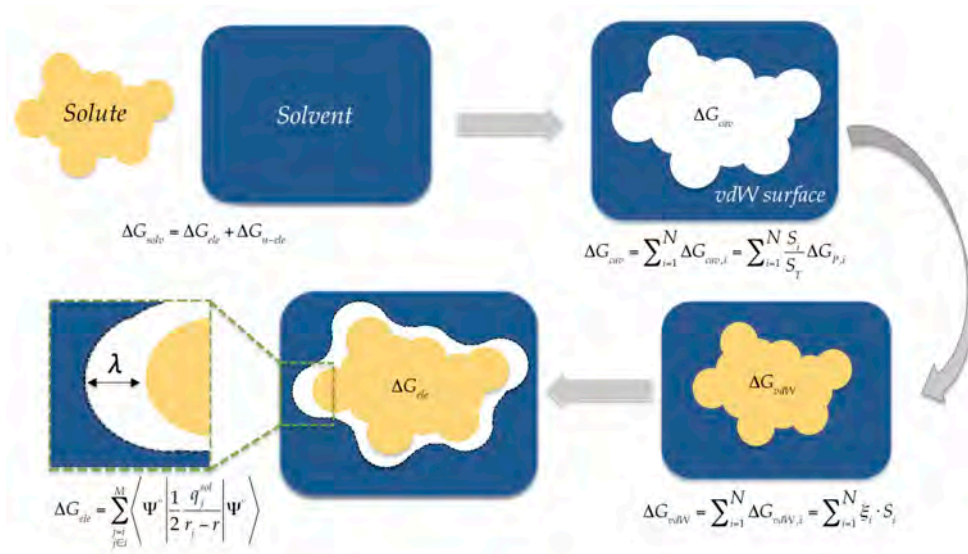
The cavitation free energy (ΔG_{cav}) is computed following Pierotti's scaled particle theory[171] adapted to molecular-shaped cavities according to the procedure proposed by Claverie[172]. In this model, the atomic cavitation free energy is computed according to:

$$\Delta G_{cav} = \sum_{i=1}^{N} \Delta G_{cav,i} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G_{P,i} \qquad (23)$$

where $\Delta G_{P,i}$ stands for the cavitation free energy of the isolated atom i in Pierotti's formalism, S_i is the solvent-exposed surface of such an atom, and S_T denotes the total surface of the atom.

CHAPTER 3. RESULTS AND DISCUSSION

The term ΔG_{vdW} is determined as the sum of the solvent-exposed surface of each atom weighted by a scalar factor determined by parametrization according to the following expression:

$$\Delta G_{vdW} = \sum_{i=1}^{N} \Delta G_{vdW,i} = \sum_{i=1}^{N} \xi_i \cdot S_i \tag{24}$$

where ξ is the atomic surface tension determined by fitting experimental values.

The electrostatic term (ΔG_{ele}) measures the work needed to build up the solute charge distribution in the solvent. Here, the integral equation formalism (IEF) was employed. Under this formalism, ΔG_{ele} can be decomposed into atomic contributions by using the surface-based partitioning method[173] (eq 25), where the fractional electrostatic contribution of a given atom i is determined from the interaction energy between the whole charge distribution of the molecule with the apparent charges located at the surface elements pertaining to the portion of the cavity generated from that atom.

$$\Delta G_{ele} = \sum_{\substack{j=1 \\ j \in i}}^{M} \left\langle \Psi^o \left| \frac{1}{2} \frac{q_{sol}^j}{|r - r_j|} \right| \Psi^o \right\rangle \tag{25}$$

where M is the total number of surface elements j, Ψ^o denotes the wave function of the solute in the gas phase, and q_{sol}^j stands for the apparent charge created on the j surface element j (located at r_j) in response to the fully polarized solute in solution.

A solvent-excluded surface is obtained by scaling the atomic radii by a factor (λ; Scheme 5) of 1.25 for solvation in water, and 1.50 for solvation in n-octanol.[135,136] These values were derived from a systematic analysis that included the comparison between the electrostatic component obtained from MST calculations and the work required to annihilate the solute charge in solution as determined from classical free-energy calculations. However, while this strategy is valuable for describing the solvation of neutral solutes, accounting for the strong electrostatic response of the

solvent induced by ionic species, and the structural perturbation of the solvent molecules in the first hydration shell relative to the bulk solvent, was treated by reducing the solvent-excluded surface in the IEFPCM/MST model. Thus, the optimum cavity for the hydration of charged compounds was defined by scaling the atomic radii of the groups bearing the formal charge by a factor of 1.13, which implies a reduction of ca. 10% relative to neutral solutes.

3.2 Refinement of the MST Model for Solvation of Neutral Nitrogen-Containing Aromatic Compounds in *n*-Octanol.

One of the initial aims of this study was to refine the parametrization of the MST model for neutral nitrogen-containing aromatic compounds (see Figure 7), as they are key structural elements in many biologically relevant molecules and drugs, but were poorly represented in the data set of compounds considered in the original B3LYP/6-31G(d) parametrization of the IEFPCM/MST model.

First, preliminary calculations performed for a subset of 12 heterocyclic organic compounds (**2–7**, **9**, **11–13**, **15**, and **16**; see Figure 7) revealed the need to adjust the surface tension of the pyridine-like nitrogen atom for solvation in *n*-octanol. Thus, the original atomic surface tension assigned to the N-type atom (ξ_N = -0.115 kcal mol^{-1} Å^{-2}) was found to underestimate the solvation free energy in *n*-octanol, and a better agreement with experimental data was achieved upon adjustment to a surface tension of -0.161 kcal mol^{-1} Å^{-2}, which was therefore adopted in the refined version (see Figure 8). Additional studies were performed to check the surface tension for the pyrrole-like nitrogen atom (NH-type), even though in this case the adjustment of the original surface tension (ξ_{NH} = -0.234 kcal mol^{-1} Å^{-2}) to -0.295 kcal mol^{-1}Å^{-2} was found to have a lower effect on the solvation free energy in *n*-octanol (see Figure 8).

$$\Delta G_{vdW} = \sum_{i=1}^{N} \xi_i \cdot S_i$$
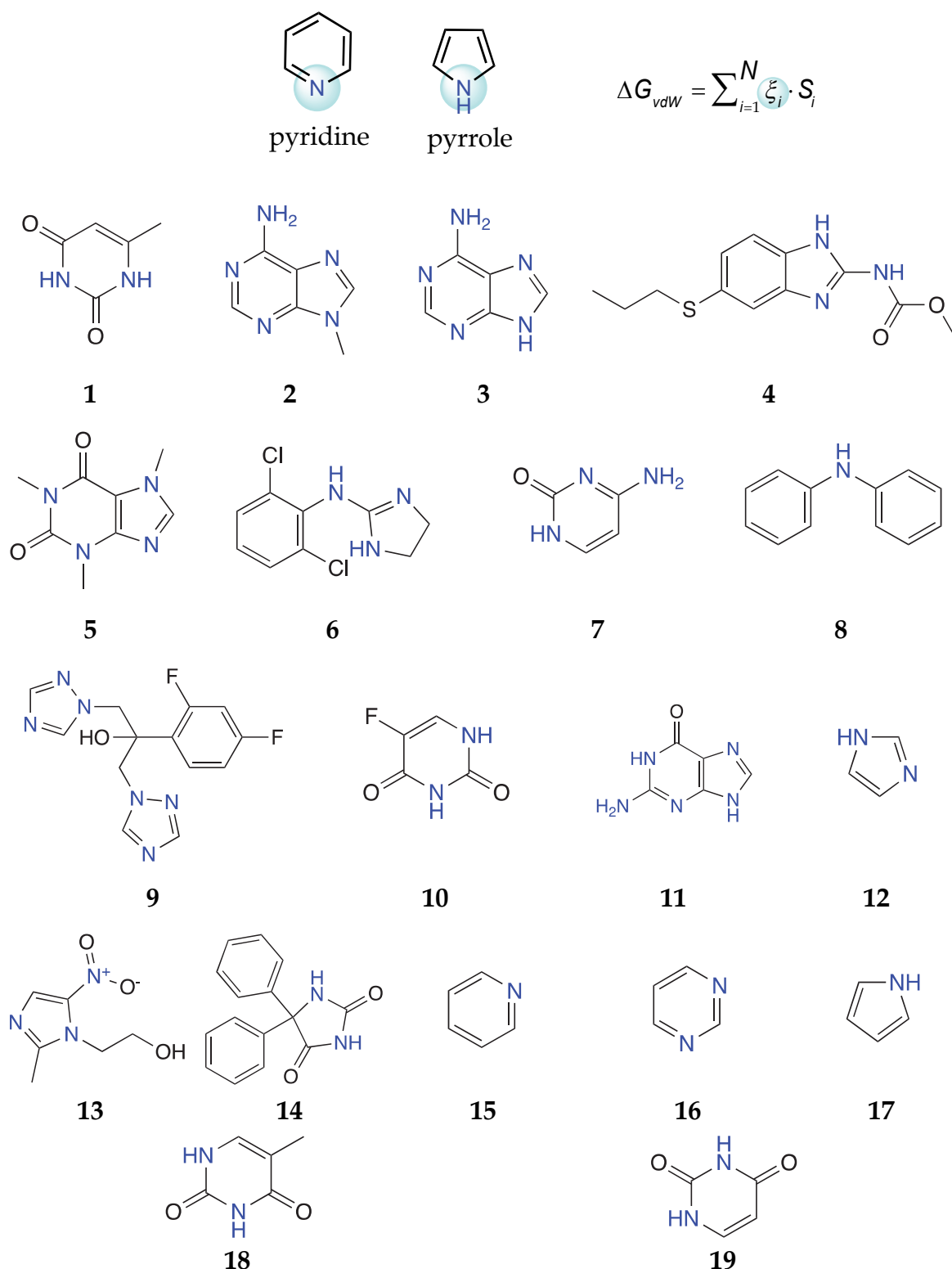
pyridine   pyrrole

Figure 7. Pyridine and pyrrole-like nitrogen atom type which the atomic surface tension (ξ) in the van der Waals component of the free energy of solvation in *n*-octanol was adjusted for the data set of 19 neutral nitrogen-containing aromatic compounds. Nitrogen atoms subjected to reparametrization are shown in blue.

Figure 8. Representation of the change of the mean unsigned error (mue) between the experimental and computed log P_N with the refined IEFPCM/MST model (*n*-octanol) as a function of the factor that scales the default atomic surface tension (ξ) in the ΔG_{cav} term for the pyridine-like (left) and pyrrole-like (right) nitrogen atom in aromatic compounds.

The effect of these refinements is shown in Table 4, which reports the solvation free energies determined with the original and refined parameters, as well as the experimental data[174] for the set of compounds. The adjustment of the surface tension of these two atom types sufficed to improve significantly the ability of the IEFPCM/MST model for predicting the log P_N values of these compounds. This is noted in the reduction of the root-mean square deviation (rmsd) from 1.9 (log P units) in the original parametrization to 0.8 for the refined version (Table 4), as well as in the comparison between experimental and calculated log P_N values, as the refined surface tensions (ξ_N and ξ_{NH}) improve the regression correlation with the experimental values from 0.85 to 0.93 (see Figure 9).

Table 4. Calculated and Experimental *n*-Octanol/Water Partition Coefficient (log P_N) for the Series of Neutral Nitrogen-Containing Aromatic Compounds Used in the Refinement of ξ_N and ξ_{NH} Atomic Surface Tensions for *n*-Octanol.

Compound[a]	Computed log P_N (original)	Computed log P_N (refined)	Exptl.[b]
6-methyluracil (**1**)	-1.7	-0.5	-1.2
9-methyladenine (**2**)	-2.9	-0.3	0.0
adenine (**3**)	-4.1	-1.1	-0.1
albendazole (**4**)	2.0	3.7	2.7
caffeine (**5**)	-0.2	0.9	-0.1
clonidine (**6**)	1.1	2.8	1.6
cytosine (**7**)	-4.3	-2.2	-1.7
diphenylamine (**8**)	3.1	3.7	3.5
fluconazole (**9**)	-1.2	1.1	0.4
fluorouracil (**10**)	-2.2	-0.9	-0.9
guanine (**11**)	-5.9	-2.7	-0.9
imidazole (**12**)	-2.2	-1.1	-0.1
metronidazole (**13**)	-0.9	0.0	0.0
phenytoin (**14**)	2.0	3.2	2.5
pyridine (**15**)	0.4	0.9	0.7
pyrimidine (**16**)	-0.8	0.1	-0.4
pyrrole (**17**)	-0.2	0.5	0.8
thymine (**18**)	-1.8	-0.5	-0.6
uracil (**19**)	-2.3	-1.1	-1.1
mse [c]	1.4	-0.1	
mue [c]	1.4	0.6	
rmsd [c]	1.9	0.8	

[a] See Figure 7 [b] Ref 174. [c] Mean signed error (mse), mean unsigned error (mue), and root-mean square deviation (rmsd) calculated relative to the experimental values are given in log P units.

Figure 9. Comparison between experimental and calculated n-octanol/water log P_N for the series of neutral nitrogen-containing aromatic compounds. Calculated values determined from IEFPCM/MST calculations using the original parametrization of the IEF-MST method (green) and the refined atomic surface tension for N- and NH-atom types (blue).

3.3 Calibration of the MST Model for Ionic Compounds in n-Octanol.

In the MST model the electrostatic contribution to the hydration free energy of charged species is determined by reducing the solvent-exposed cavity of the charged atoms by a factor close to 10%.[175] While this strategy proved to be valuable for calculating the solvation of univalent ionic species in water, its suitability in other solvents has never been checked. Therefore, for our purposes here, it is necessary to calibrate the reliability of this strategy for the solvation of ionic compounds in n-octanol. To this end, calculations were performed for a set of 27 compounds, including 9 anions (**22, 29-31, 36, 37, 39** and **41**; see Figure 10) and 18 cations (**20, 21, 23-28, 32-35, 38, 40, 43-46**; see Figure 10), taking advantage of the availability of partition coefficients for these charged species.[80,95,104,119,123,155,165,176]
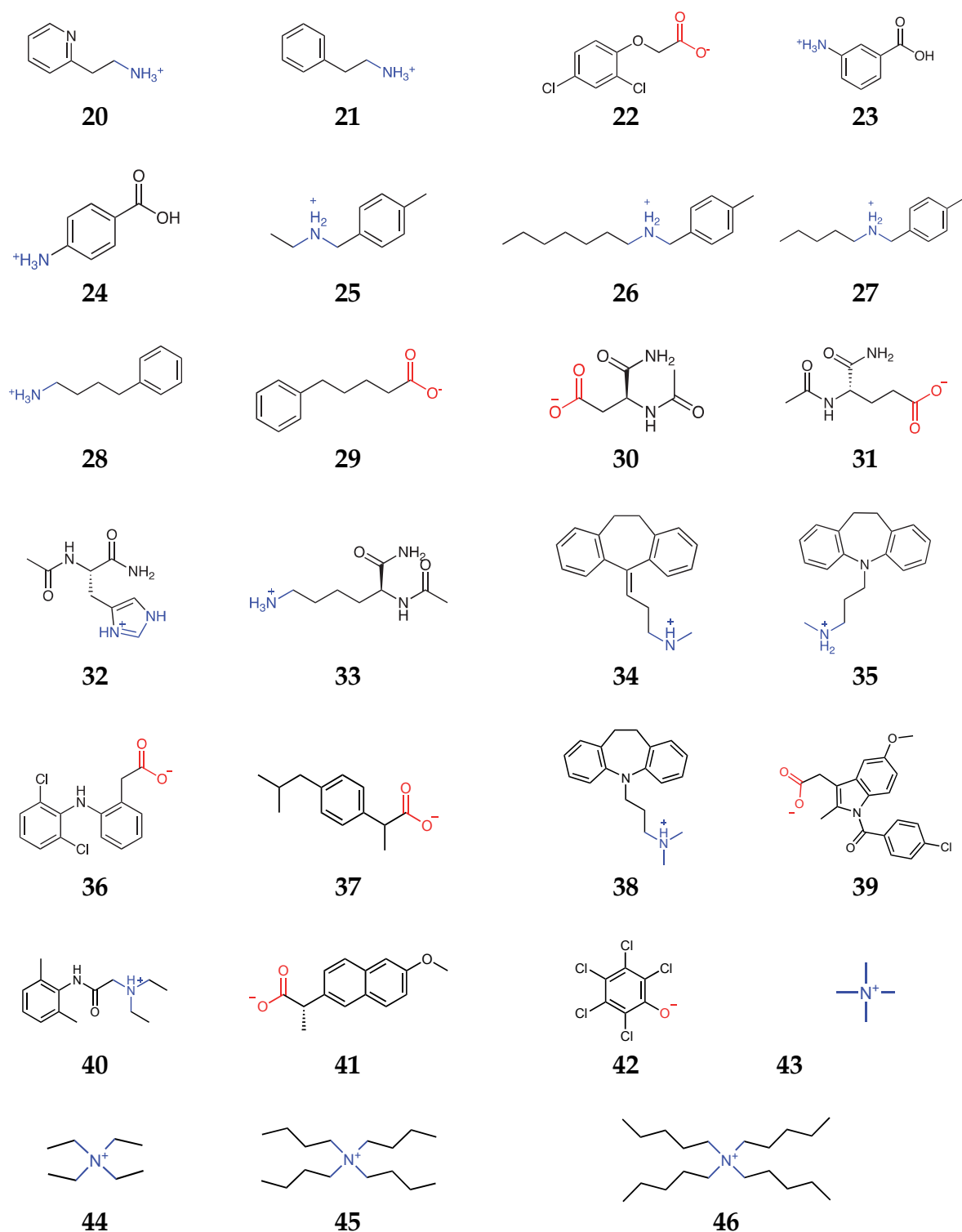
Figure 10. Data set of 27 ionizable compounds used to refine the MST solvation model for solvation in *n*-octanol. Atoms subjected to reparametrization are shown in blue and red for cations and anions, respectively.

Comparison of the calculated and experimental log P_I values determined for these compounds suggested that the optimal scaling factor, λ, for solvation in *n*-octanol must be reduced by around 19%, which implies that the scaling factor used for neutral compounds ($\lambda = 1.50$) must be close to 1.20 for charged chemical groups. This adjustment enhances the contribution of the electrostatic component to the solvation free energy for charged compouds, following the trends reported for the hydration of monovalent ions,[175] an effect interpreted from the balance between the gain in solvent-solute stabilization energy triggered by the solute's electron density redistribution upon solvation and the energy cost associated to distortion of the electron density by the solvent reaction field.[177–179]

Due to the formal simplicity of this correction, the suitability of the atomic surface tension was further checked. In the case of cations with a localized charge on the *sp³* nitrogen atom, it was necessary to enlarge the surface tension of the nitrogen atom (NH atom type) by a factor of 17% ($\xi_{NH} = -0.274$ kcal mol^{-1} Å^{-2}; see Figure 11). This enlargement was also extended to the methylene/methyl groups bound to the protonated nitrogen atom ($\xi_{CHx} = -0.227$ kcal mol^{-1} Å^{-2}; see Figure 11), which may be related to the inductive effect noted in the increased chemical shift observed in ^{1}H-NMR studies (see Table 5).[180] This effect is known to be less important for the carbon atoms bound chemical groups with delocalized charges (i.e., carboxylate anions; see Table 5), where no further adjustment was needed.
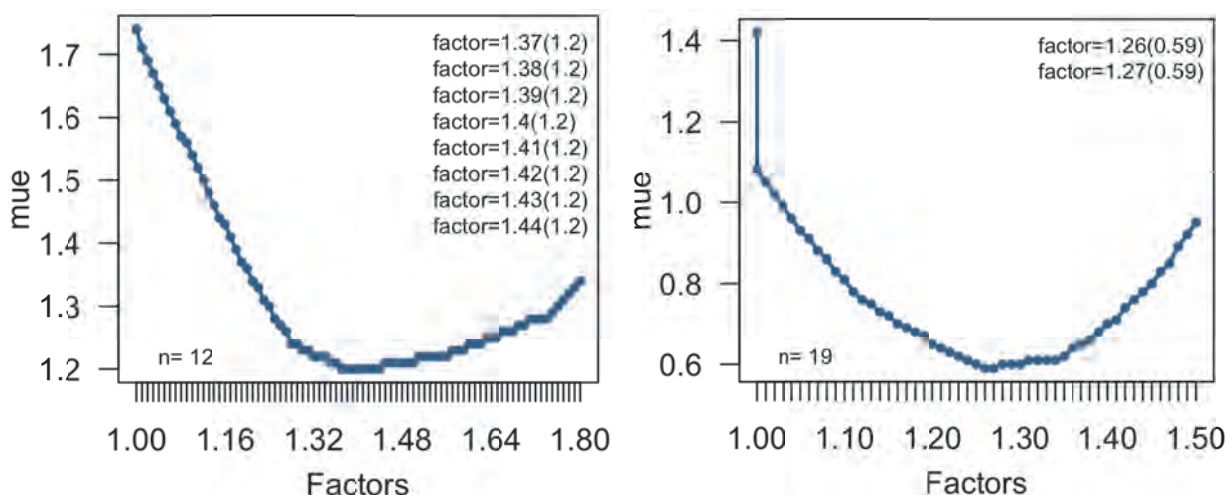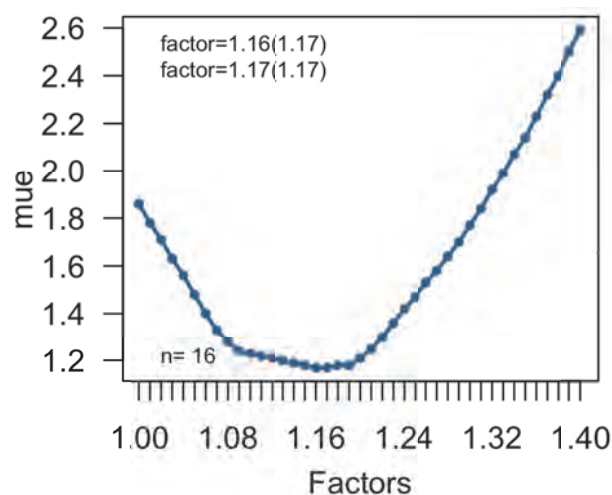
Figure 11. Representation of the change of the mean unsigned error (mue) between the experimental and computed log P_I with the refined IEFPCM/MST model (*n*-octanol) as a function of the factor that scales the default atomic surface tension (ξ) in the ΔG_{cav} term for NH and CH$_x$ atoms for 16 cations (cation **23** and **24** were not included).

Table 5. The Substituent α-Effect in ^{1}H-NMR for Anionic and Cationic Organic Compounds.

Neutral species δ_H (ppm) [a] in α	Charged species δ_H (ppm) [a] in α	diff (ppm) $\delta_{H,charged} - \delta_{H,neutral}$
2.18	2.38	0.2
2.16	2.33	0.2
2.65	3.06	0.4
2.65	3.04	0.4

[a] http://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

The log P_I values calculated for the whole set of ionic organic compounds using the new parameters are presented in Table 6, which also collects the experimental data. The mean signed error was reduced from 4.1 to -0.2 (log P_I units), and the rmsd was decreased from 4.7 to 1.1 (log P_I units) after implementation of the preceding adjustments in the MST model. Furthermore, Figure 12 shows the improved correlation between the refined log P_I values and the experimental ones, which corrected the systematic tendency to overestimate the hydrophilicity of the charged compounds in the original parametrization of the IEFPCM/MST method. For the sake of comparison, it is worth noting that the refined log P_I values are in agreement with the behavior observed for the values estimated by using empirical methods, such as ACD/I-Lab[131] and ChemAxon[132] methods (see Figure 12)

Table 6. Experimental and Computed log P_I Values for the Set of 27 Ionic Compounds Used in the Refinement of the MST Method.

Compound [a]	Computed log P_I (original)	Computed log P_I (refined)	Exptl.
2-2-pyridyl-ethylammonium (**20**)	-1.7	-0.4	-2.3 [b]
2-phenethylammonium (**21**)	-7.1	0.4	-1.6 [b]
2-4-dichlorophenoxyacetate (**22**)	-5.5	-0.7	-0.9 [c]
3-carboxyanilinium (**23**)	-9.2	-1.6	-0.9 [c]
4-carboxyanilinium (**24**)	-9.4	-1.6	-0.4 [c]
4-methyl-*N*-ethylbenzylammonium (**25**)	-6.4	-1.9	-0.8 [d]
4-methyl-*N*-heptylbenzylammonium (**26**)	-3.6	0.1	2.1 [d]
4-methyl-*N*-pentylbenzylammonium (**27**)	-4.7	-0.9	0.8 [d]
4-phenylbutylamine (**28**)	-6.1	1.5	0.7 [e]
5-phenylvalerate (**29**)	-6.1	-0.4	-1.0 [c]
N-acetyl-*L*-aspartic amide (**30**)[f]	-6.8	-2.8	-2.6 [g]
N-acetyl-*L*-glutamic amide (**31**)[f]	-6.2	-3.4	-2.5 [g]
N-acetyl-*L*-histidine amide (**32**)[h]	-7.9	-1.7	-3.4 [g]
N-acetyl-*L*-lysine amide (**33**)[f]	-7.7	-1.8	-2.8 [g]
amitriptyline (**34**)	-1.1	1.3	0.2 [c]
desipramine (**35**)	-2.7	0.0	0.3 [c]
diclofenac (**36**)	-4.3	0.7	0.7 [c]
ibuprofen (**37**)	-5.9	-0.7	-0.2 [c]
imipramine (**38**)	-0.6	1.8	0.5 [c]
indomethacin (**39**)	-2.7	2.1	0.6 [c]
lidocaine (**40**)	-2.6	-0.7	-0.5 [c]
naproxen (**41**)	-5.5	-0.6	-0.2 [c]
pentachlorophenol (**42**)	-1.9	1.8	1.3 [i]
tetrabutylammonium (**43**)	1.2	3.3	2.3 [j]
tetraethylammonium (**44**)	-2.8	-0.8	-0.9 [j]
tetramethylammonium (**45**)	-5.6	-2.7	-2.0 [j]
tetrapentylammonium (**46**)	3.0	5.1	3.8 [j]
mse [k]	4.1	-0.2	
mue [k]	4.1	0.9	
rmsd [k]	4.7	1.1	

[a] See Figure 10. [b] Ref. 123. [c] Ref. 99. [d] Ref. 238. [e] Ref. 95. [f] Values derived from log $D_{7.4}$ data reported in ref. 66, assuming full ionization of the compounds at physiological pH. [g] Refs. 165 and 155. [h] Estimated from additive scheme (see Methods). [i] Ref. 119. [j] Ref. 176. [k]Mean signed error (mse), mean unsigned error (mue), and root-mean square deviation (rmsd) calculated relative to the experimental values are given in log P units.
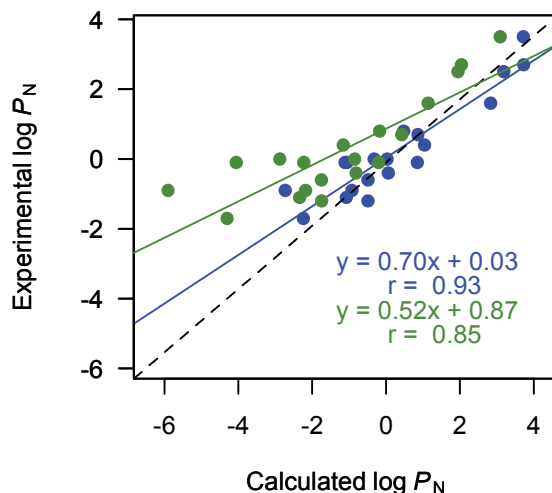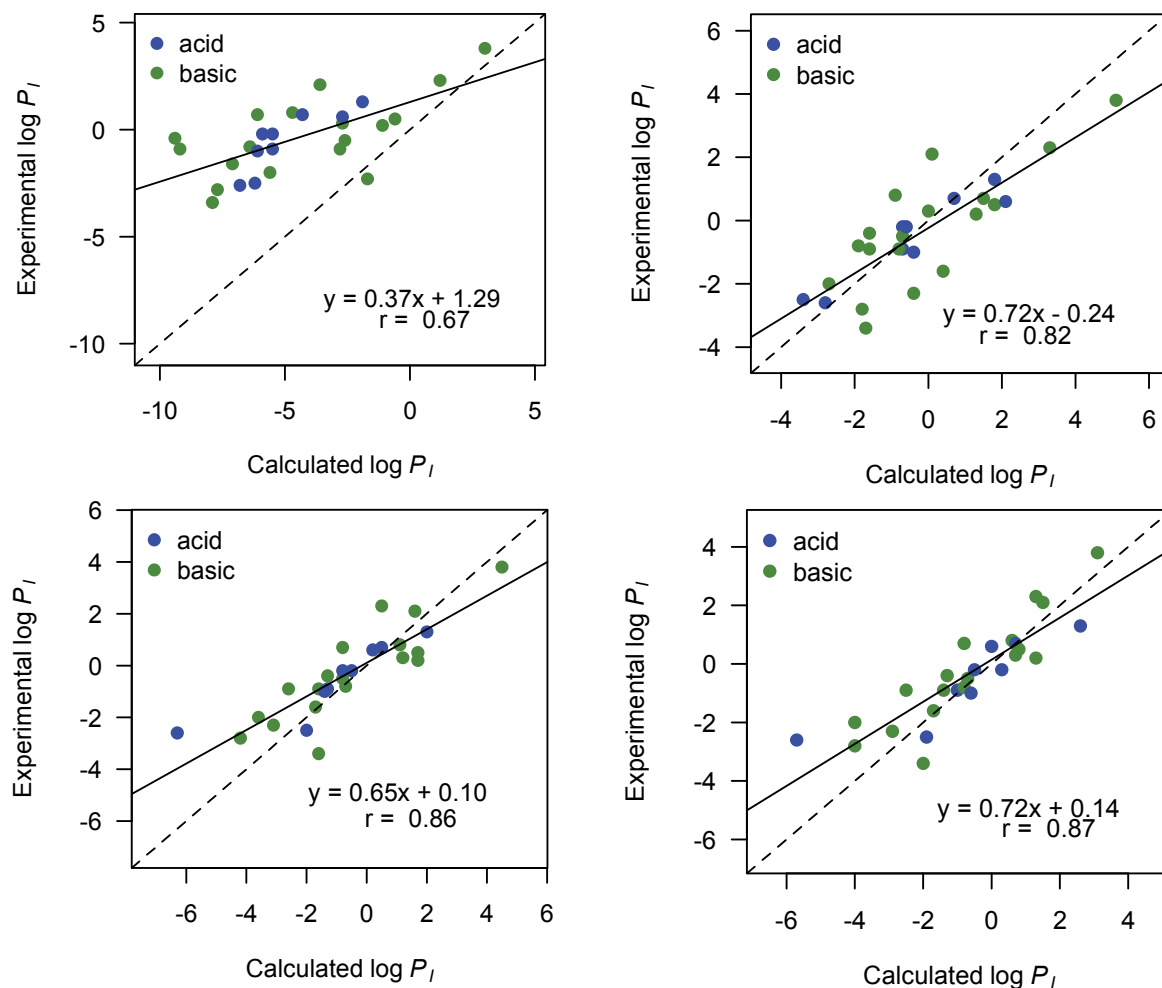
Figure 12. Comparison between experimental and calculated log P_I values for the series of 27 ionic compounds (acidic and basic compounds are shown in blue and green, respectively). Calculated values were determined from IEFPCM/MST computations using the original (top left) and refined (top, right) parameters, as well as ACD/I-Lab (bottom left) and ChemAxon (bottom right).

3.4 Refinement of Lipophilic Profiles of (Bio)Organic Compounds.

In the simulation of pH-dependent lipophilicity profiles based on the n-octanol/water for (bio)organic compounds, as presented in the introduction of this thesis, the log P_N and pK_a values are the minimal data for a theoretical model (eq 5). The degree of refinement of these models explicitly brings about a greater complexity and imposes the need for using more variables (i.e log P_I, and/or log P_{IP}), which should a priori be rewarded by greater accuracy. As mentioned before, the refined MST model enables the calculation of reliable values of log P_N and log P_I, opening the door to the study of the computation of distribution coefficients (D) employing different formalism.

In this work, an alternative theoretical formalism for the lipophilicity profile of ionizable compounds was proposed to account for the electroneutrality in the phases of the n-octanol/water system (Scheme 6).

Scheme 6. Mechanism of n-Octanol/Water Partition for Neutral (HX) and Ionic Species (X^- and C_I^+) Influenced by the Electric Potential at the Interphase ($\Delta_o^w \phi$)

$$D = \frac{\left[HX + X^- + C_I^+\right]_o}{\left[HX + X^- + C_I^+\right]_w}$$
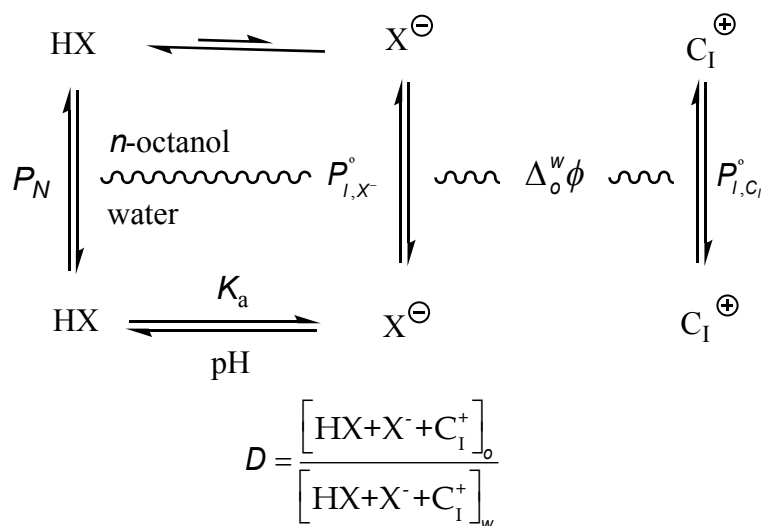
This development was aimed to look upon the effect of the counter ion (C_I^+) accompanying the ionizable (bio)organic compound (X^-). In this context, the most accepted hypothesis reported in the literature relies on the formation of ion pairs ($X^-C_I^+$).[67,95,125,117–124] However, electrochemistry offers another perspective based on powerful methods for studying and predicting the behaviour of ionic species of an ionizable compound in the interface between two immiscible electrolyte solutions (ITIES).

The formalism presented here uses the pioneering theories of Hung[181] and Kakiuchi[182], whose foundation assumes the electroneutrality of each phase and allows to derive expressions for the partition for an ionic compound from the initial concentrations of the ions in the solution, its standard Gibbs free energy of transfer, and the volume ratio of the two phases.

On the basis of the ITIES theory, the standard partition coefficient of a given ionic species i is given by[183]

$$\log P_{I,i}^{\circ} = -\frac{z_i F}{RT \ln 10} \Delta_o^w \phi_i^{\circ} \tag{26}$$

where R is the gas constant, T the absolute temperature, F is the Faraday constant, z_i stands for the net charge of i and $\Delta_o^w \phi_i^{\circ}$ is the standard transfer potential of the ionic specie i, which only depends on the chemical structure.

As expected, there are other ions in the solution and the Galvani potential difference in the interphase ($\Delta_o^w \phi$) is a function of their type and concentration. Thus, for the ionic organic species (X^-) the apparent partition is given by eq 27.[184]

$$\log P_{I,X^-}^{app} = \log P_{I,X^-}^{\circ} + \frac{z_{X^-} F}{RT \ln 10} \Delta_o^w \phi \tag{27}$$

It can be demonstrated that if the two immiscible electrolyte solutions are dilute, and that all ionic species are fully dissociated in both phases (i.e., no ion pair formation), for a generic electrolyte $C_I^+ A^-$ the Galvani potential difference can be rewritten as[182]

$$\Delta_o^w \phi = \left(\frac{\Delta_o^w \phi_{C_I^+}^{\circ} + \Delta_o^w \phi_{A^-}^{\circ}}{2} \right) \tag{28}$$

Employing eq 26 in eq 28, it can be deduced that the apparent partition can be expressed as

$$P_{I,X^-}^{app} = \sqrt{P_{I,X^-}^{\circ} \cdot P_{I,C_I^+}^{\circ}} \tag{29}$$

Using this approach, the distribution coefficient is given by

$$\log D = \log\left(P_N + \sqrt{P_{I,X^-}^{\circ} \cdot P_{I,C_I^+}^{\circ}} \cdot 10^\delta \right) - \log(1 + 10^\delta) \tag{30}$$

which explains in an alternative way to the ion pair theory why the distribution coefficient increases in the presence of a more hydrophobic counter ion.

In addition, if one assumes that the partitioning of the main ionic species corresponds to the generic electrolyte ($C_I^+ A^-$; Scheme 7), as in cases of high ionic strength, the Galvani potential difference in the interphase ($\Delta_o^w \phi$) depends entirely on those species and permits to rewrite eq 29 as noted in eq 31.

$$P_{I,X^-}^{app} = P_{I,X^-}^{\circ} \sqrt{\frac{P_{I,C_I^+}^{\circ}}{P_{I,A^-}^{\circ}}} \tag{31}$$

In this way, the distribution coefficient can be expressed as

$$\log D = \log\left(P_N + P_{I,X^-}^{\circ} \sqrt{\frac{P_{I,C_I^+}^{\circ}}{P_{I,A^-}^{\circ}}} \cdot 10^\delta \right) - \log(1 + 10^\delta) \tag{32}$$

Scheme 7. Mechanism of *n*-Octanol/Water Partition for Neutral (HX) and Ionic Species (X^-, A^- and C_I^+) Influenced by the Electric Potential at the Interphase ($\Delta_o^w \phi$) Formed by the Salt Background ($C_I^+ A^-$).
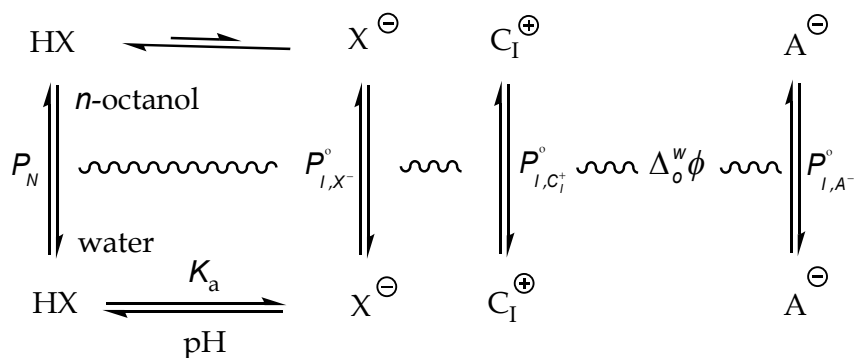
$$D = \frac{\left[\text{HX} + X^- + C_I^+ + A^- \right]_o}{\left[\text{HX} + X^- + C_I^+ + A^- \right]_w}$$

Eq 32 indicates that for an acidic compound (HX), there will be an increase in the lipophilicity when a salt ($C_I^+ A^-$) is added whenever the partition of the cation (C_I^+) increases and decreases to the anion (A^-) independently of the standard partition of its ionic form (X-)

As a final remark, it must be noted that partitioning of a cationic species, X⁺, would give rise to distribution coefficients formally analogous to eqs 30 and 32, which are omitted here for the sake of brevity. It is also worth noting that the preceding formalisms limit the distribution coefficient to the partition of both neutral and ionic species of an ionizable compound.

With the aim to look for a unified model that includes both the effect of the Galvani potential and ion-pairing for determining lipophilicity profiles of monoacid and monobasic compounds, a general theoretical formalism was proposed (Scheme 8), whose validity is currently being tested experimentally in collaboration with Prof. Clara Ràfols and Prof. Martí Rosés, members of the research group PhysChem (Fac. Chemistry) at the University of Barcelona.

Scheme 8. Mechanism of *n*-Octanol/Water Partition for Neutral (HX) and Ionic Species (X^-, A^- and C_I^+) Influenced by the Electric Potential at the Interphase ($\Delta_o^w\phi$) and Ion Pair Formation ($C_I^+A^-$).
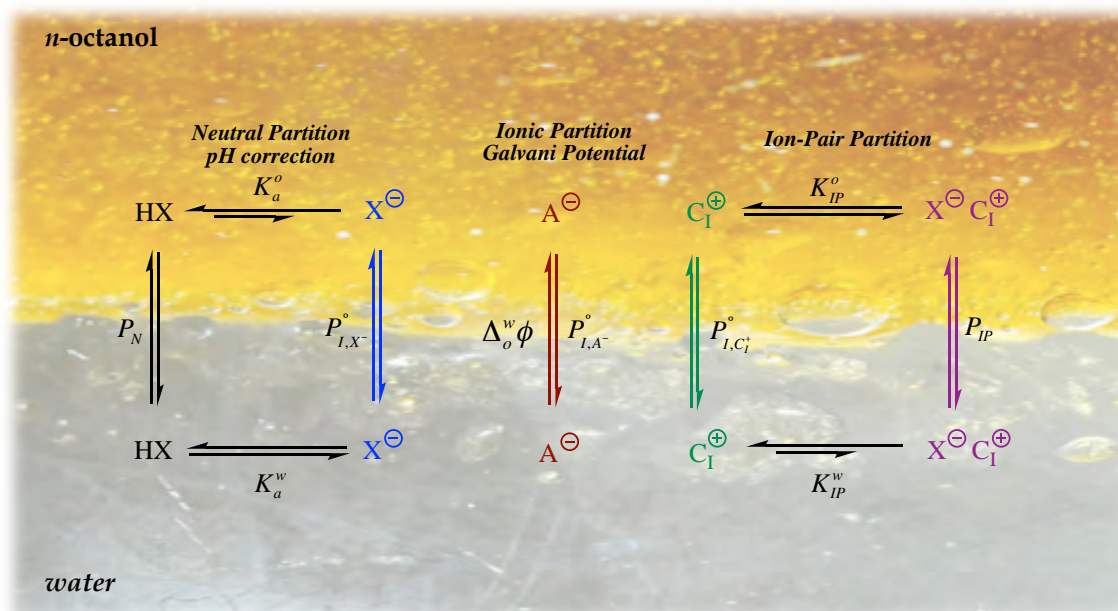
It can be demonstrated that using the refined models presented previously, the distribution coefficient in the framework of Scheme 8 leads to eq 33.

$$\log D = \log\left(P_N + \left(P_{I,X^-}^{app} + P_{IP}\cdot K_{IP}^w\cdot[C_I^+]\right)\cdot 10^\delta\right) - \log\left(1 + \left(1 + K_{IP}^w\cdot[C_I^+]\right)\cdot 10^\delta\right) \quad (33)$$

where P_{I,X^-}^{app} can be reduced to eq 27 or 29 depending on the concentration of the salt added for determining the lipophilicity profile of the ionizable compound (HX).

Base on the formalism proposed in Scheme 8, the P_{I,X^-}^{app} given in eq 29 and 31 would be representative of extreme cases. Thus, eq 29 makes reference when the organic ionic species and its counter ion are the ions that dominate the potential differences in the interphase. On the other hand, eq 31 stands for the situation where the background salt is the major factor that contributes to the potential difference. One

could make an approximation in intermediate cases defining a variable, r_{C_I}, according to eq 34,

$$r_{C_I} = \frac{[X^-]}{[X^-]+[A^-]}$$ (34)

where X^- represents the ionic organic species, and A^- stands for the ionic species with the same charge from the background salt. Here, for the sake of brevity we limit ourselves to the specific case of an acidic compound (HX). So, the P_{I,X^-}^{app} can be approximated using this variable r_{C_I}, although future work is still necessary to calibrate the formal dependence of r_{C_I}, which could adopt either linear (eq 35) or a exponential (eq 36) models.

$$P_{I,X^-}^{app} = P_{I,X^-}^{o} \cdot \sqrt{P_{I,C_I^+}^{o} / P_{I,A^-}^{o}} \cdot \left(r_{C_I}\right) + \sqrt{P_{I,X^-}^{o} \cdot P_{I,C_I^+}^{o}} \cdot \left(1 - r_{C_I}\right)$$ (35)

$$P_{I,X^-}^{app} = \sqrt{\frac{P_{I,C_I^+}^{o} \cdot \left(P_{I,X^-}^{o}\right)^{(r_{C_I}+1)}}{\left(P_{I,A^-}^{o}\right)^{r_{C_I}}}}$$ (36)

From these two last equations, if the r_{C_I} is known, and the standard partition of ech ion in the background salt, the standard partition for the organic ion of interest can be derived.

3.5 Estimation of *n*-Octanol/Water Distribution Coefficients at Physiological Conditions (log $D_{7.4}$).

The log $D_{7.4}$ values compiled for a set of 35 ionizable small molecules (see Figure 13) were used as a test set to calibrate the suitability of the adjustments introduced in the refined IEFPCM/MST model.

47 **48** **49** **50**

51 **52** **53** **54**

55 **56** **57** **58**

59 **60** **61** **62**

63 **64** **65** **66**

| **67** | **68** | **69** | **70** |

| **71** | **72** | **73** | **74** |

| **75** | **76** | **77** | **78** |

| **79** | **80** | **81** |
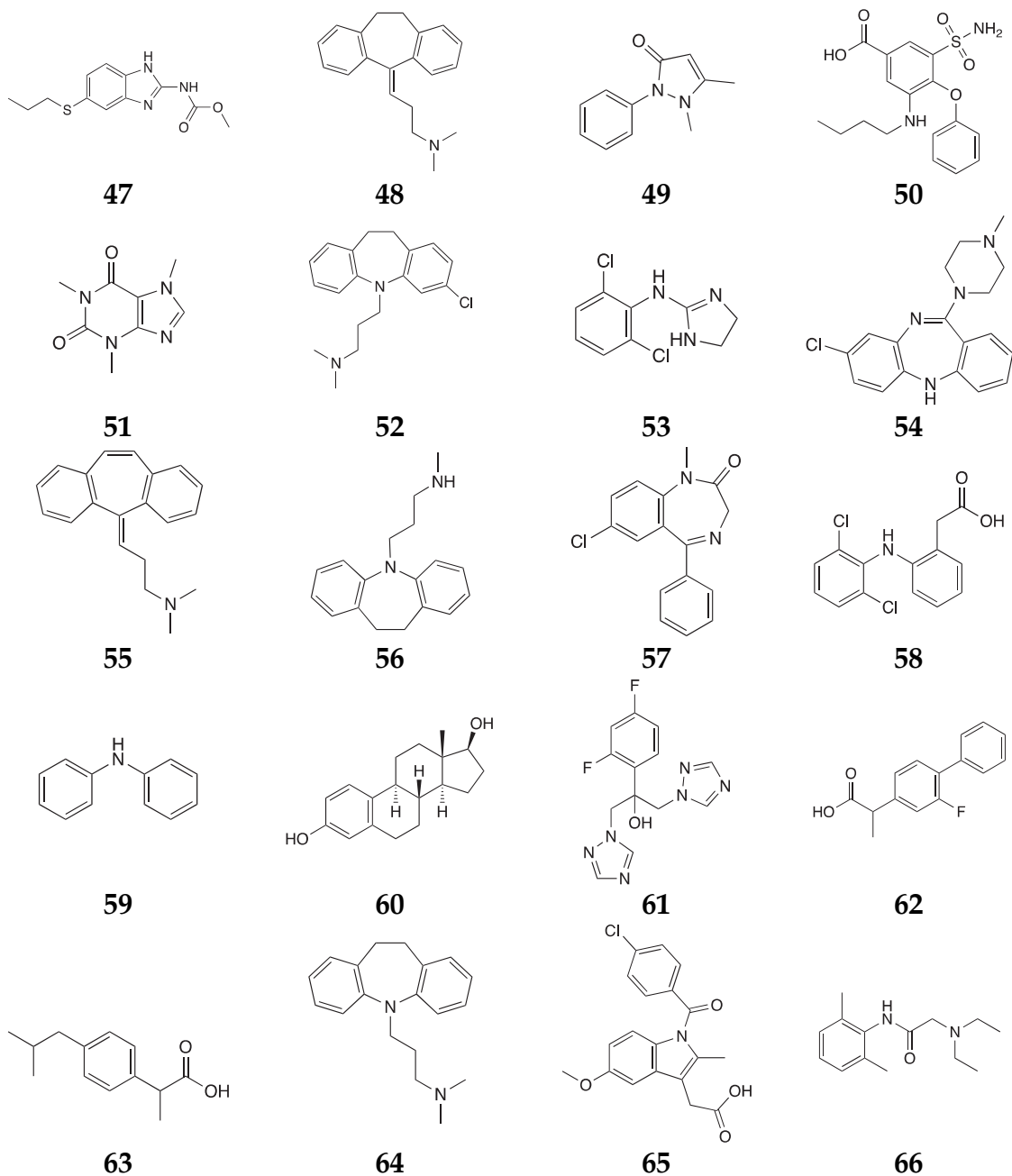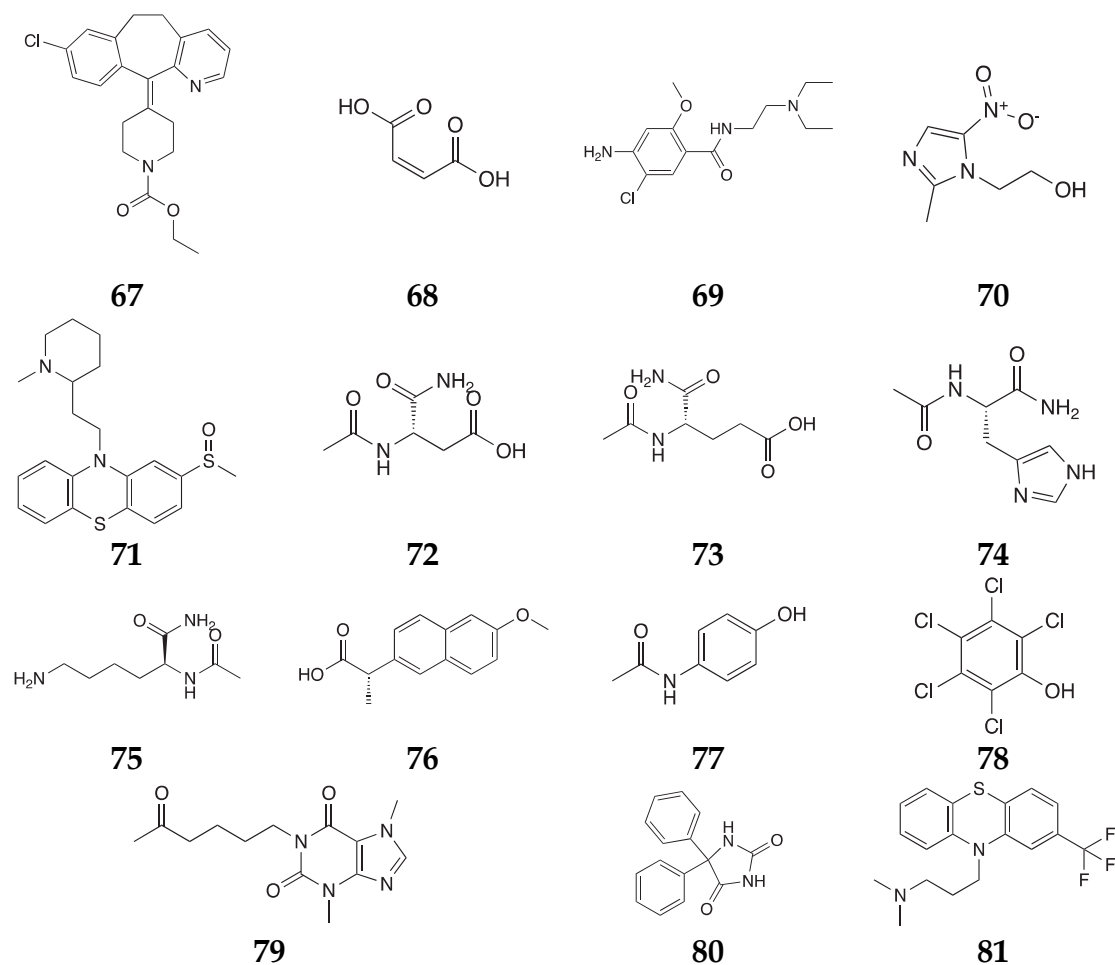
Figure 13. Data set of 35 ionizable compounds used to calibrate the MST solvation model for estimating the distribution coefficient.

These compounds encompass a broad range of chemical diversity in selected physicochemical properties (see Figure 14), such as molecular weigth (up to 400 Dalton), number of rotatable bonds (up to 8), number of aromatic rings (up to 3), and number of hydrogen bond donors (up to 5) and acceptors (up to 7).
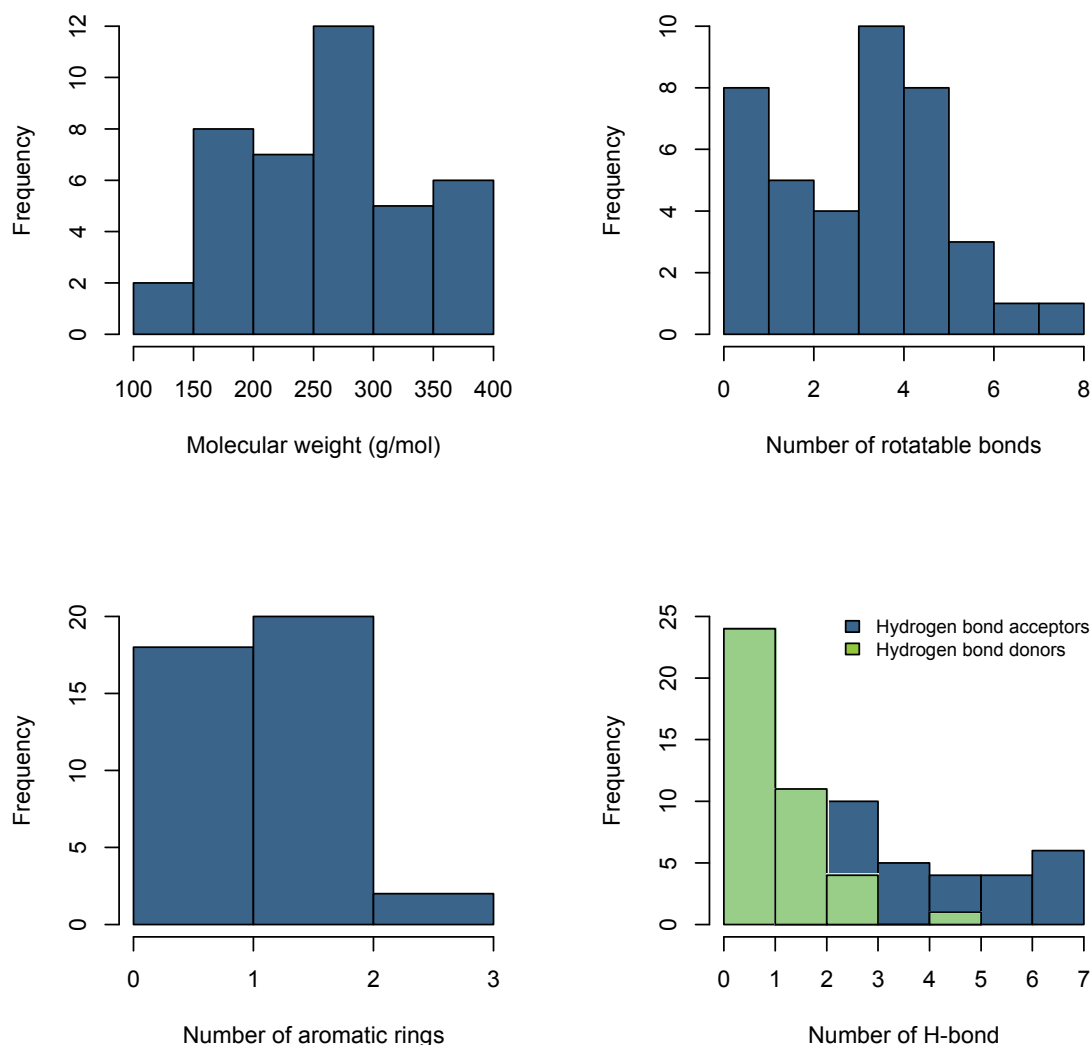
Figure 14. Histograms of molecular properties (molecular weight, number of rotatable bonds, number of aromatic rings and hydrogen bond acceptors and donors) for the set of 35 small molecules.

In order to reproduce the experimental distribution coefficients within the framework of the partition formalisms represented by eqs 5, 6, and 30, log P_N and log P_I values were estimated from MST calculations, whereas the pK_a of these compounds and the partition coefficient of the counter ion were taken from experimental data (see Table 7). We first evaluated the capacity of the refined MST model for predicting the experimental log P_N of these compounds. This comparison

is shown in Table 7 and Figure 15. The rmsd between experimental and calculated values is 1.1 (log P_N units), and the calculated values exhibit a good correlation with the experimental ones ($r = 0.94$). Furthermore, these trends compare well with the values predicted by using empirical methods (ACD/I-lab or ChemAxon; Figure 15), although they exhibit a slightly better correlation with the experimental data, which likely reflects the most extensive parametrization of fragmental contributions that lies behind these methods.[138,185]

Since the distribution coefficient takes into account the partition of both neutral and ionic species of ionizable compounds, it provides an indirect approach to test the reliability of the calculated P_I values. This is more challenging, because the measured log D may be affected by the experimental conditions, such as the nature of the background salt and the concentration of the solution, which would influence the potential difference between the two phases.[186] Moreover, different formalisms have been proposed to combine log P_N, log P_I and pK_a to estimate the log D. Accordingly, the log $D_{7.4}$ was determined using the log P_N and log P_I values determined from IEFPCM/MST computations, and were combined with experimental pK_a values reported for the set of compounds (see Table 7).

Table 7. Experimental and Computed Data for the Set of 35 Ionizable Compounds Used to Analyze the Performance of the MST Solvation Model for Estimating the Partition and Distribution Coefficient (Data taken from refs. 6, 7, 99 and 155).

Compound [a]	Exptl. pK_a	Exptl. $\log P_N$	Comp. $\log P_N$	Exptl. $\log D_{7.4}$	Comp. $\log D_{7.4}$ eq 5	eq 6	eq 30
albendazole (**47**)	4.21	2.7	3.7	3.3	3.7	3.7	3.7
amitriptiline (**48**)	9.40	4.9	6.5	2.8	4.5	4.5	4.5
antipyrine (**49**)	1.40	0.4	2.0	0.2	2.0	2.0	2.0
bumetanide (**50**)	3.60	2.6	2.6	-0.1	-1.2	-0.2	-1.0
caffeine (**51**)	-0.92	-0.1	0.9	0.0	0.8	0.8	0.8
clomipramine (**52**)	9.40	5.2	6.7	3.3	4.7	4.7	4.7
clonidine (**53**)	8.05	1.6	2.8	0.6	2.1	2.1	2.1
clozapine (**54**)	7.50	3.2	5.5	3.0	5.1	5.1	5.1
cyclobenzaprine (**55**)	8.47	5.2	6.3	2.9	5.2	5.2	5.2
desipramine (**56**)	10.4	4.9	5.7	1.4	2.7	2.7	2.7
diazepam (**57**)	3.30	2.8	4.5	2.7	4.5	4.5	4.5
diclofenac (**58**)	4.15	4.5	5.6	1.1	2.3	2.4	2.3
diphenylamine (**59**)	1.03	3.5	3.7	3.4	3.7	3.7	3.7
estradiol (**60**)	10.7	3.7	4.2	4.0	4.2	4.2	4.2
fluconazole (**61**)	2.94	0.4	1.1	0.5	1.1	1.1	1.1
fulbiprofen (**62**)	4.22	4.2	4.2	0.9	1.0	2.2	1.0
ibuprofen (**63**)	4.91	4.0	3.2	1.3	0.7	0.7	0.7
imipramine (**64**)	9.40	4.8	5.9	2.5	3.9	3.9	3.9
indomethacin (**65**)	4.50	4.3	4.9	0.8	2.0	2.4	2.0
lidocaine (**66**)	8.01	2.4	2.8	1.6	2.1	2.1	2.1
loratadine (**67**)	4.58	5.2	7.4	4.4	7.4	7.4	7.4
maleic acid (**68**)	1.83	-0.5	-1.5	-5.0	-7.1	-4.5	-3.5
metoclopramide (**69**)	9.27	2.6	2.2	0.5	0.3	0.3	0.3
metronidazole (**70**)	2.60	0.0	0.0	-0.1	0.0	0.0	0.0
mesoridazine (**71**)	8.89	3.9	6.5	1.8	5.0	5.0	5.0
N-acetyl-*L*-aspartic amide (**72**)	3.90	-2.0	-2.3	-2.6	-5.8	-2.8	-2.6
N-acetyl-*L*-glutamic amide (**73**)	4.20	-1.9	-1.5	-2.5	-4.7	-3.4	-2.9
N-acetyl-*L*-histidine amide (**74**)	7.00	-1.9	-0.9	-3.5	-1.0	-0.9	-1.0
N-acetyl-*L*-lysine amide (**75**)	11.1	-0.8	-0.4	-2.8	-1.8	-1.8	-3.1
naproxen (**76**)	4.15	3.2	2.7	0.3	-0.6	-0.3	-0.5
paracetamol (**77**)	9.38	0.5	-0.1	0.3	-0.1	-0.1	-0.1
pentachlorophenol (**78**)	4.83	5.0	3.8	2.5	1.2	1.9	1.2
pentoxifylline (**79**)	0.28	0.3	1.6	0.3	1.6	1.6	1.6
phenytoin (**80**)	8.33	2.5	2.0	2.2	1.9	1.9	1.9
triflupromazine (**81**)	9.40	5.5	6.6	3.4	4.6	4.6	4.6
mse [b]			-0.6		-0.6	-0.9	-0.8
mue [b]			0.9		1.3	1.1	1.1
rmsd [b]			1.1		1.6	1.4	1.4

[a] See Figure 13. [b] Mean signed error (mse), mean unsigned error (mue), and root-mean square deviation (rmsd) calculated relative to the experimental values are given in log P units.
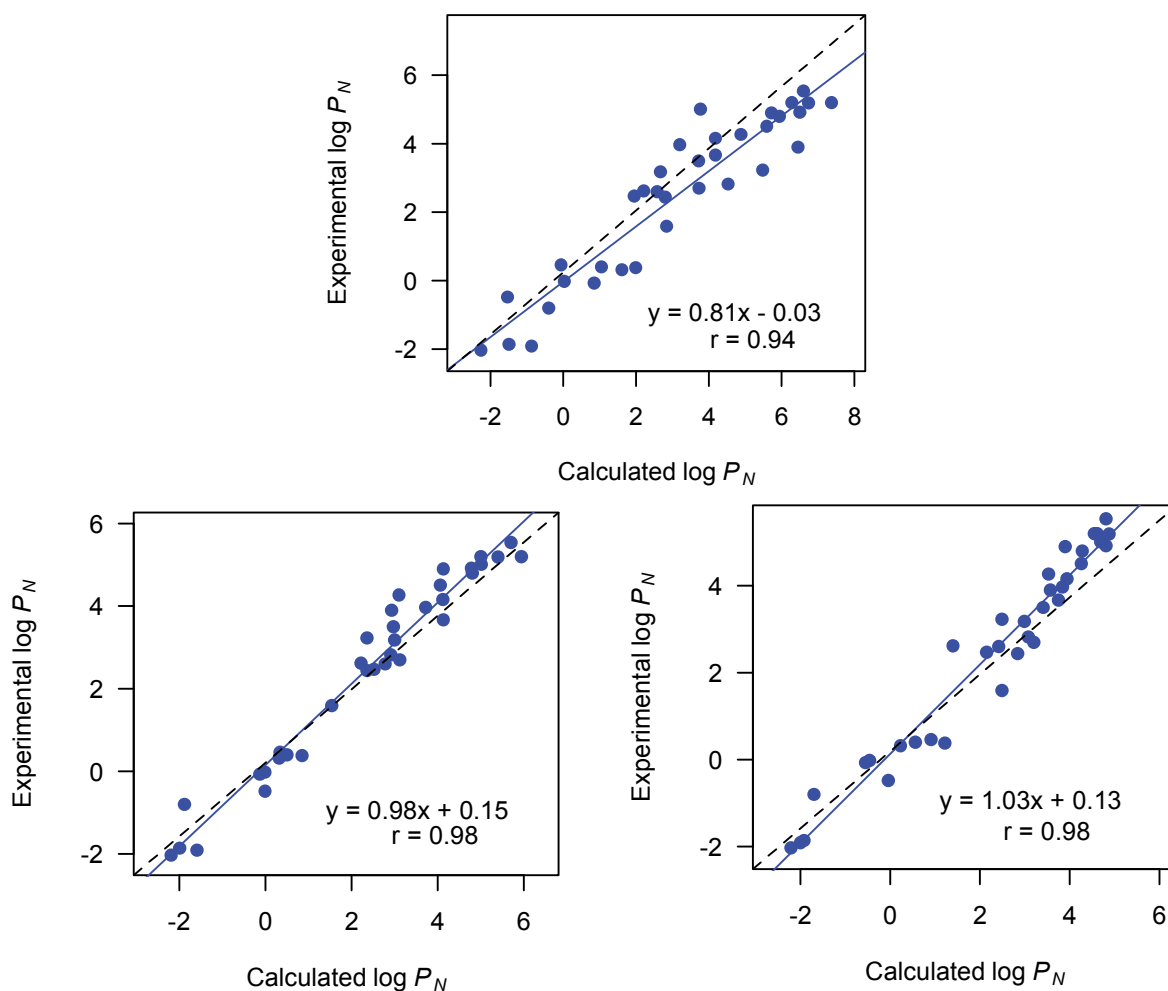
Figure 15. Comparison between experimental and calculated log P_N for a set of 35 small molecules. Computed values were determined by using the refined IEFPCM/MST calculations (top), and by using ACD/I-Lab log D (bottom left) and ChemAxon (bottom right).

Table 7 reports the log $D_{7.4}$ values obtained from IEFPCM/MST calculations. In general, there is a slight tendency to overestimate the hydrophobicity of the compounds, as noted in the mean signed error (mse) of ~ -0.7 (log D units) found for eqs 5, 6, and 30, while the rmsd amounts to ~ 1.5 (log D units). The performance of the IEFPCM/MST model is similar for the three formalisms examined in this study, with a slightly larger rmsd when eq 5 is used. Similar regression equations between calculated and experimental log $D_{7.4}$ values are also found ($r \approx 0.92$), albeit Figure 16 shows a slightly larger deviation from the perfect linear regression for the values obtained with eq 5. For this latter formalism the largest errors (given in log D units in parenthesis) are found for the subset of amino acid analogues N-acetyl-L-aspartic

acid amide (+3.2), *N*-acetyl-*L*-glutamic acid amide (+2.2) and *N*-acetyl-*L*-histidine amide (+2.5), as well as for mesoridazine (-3.1), loratadine (-3.0), cyclobenzaprine (-2.3), clozapine (-2.1), and maleic acid (+2.1). Nevertheless, when the partition of the ionic species is taken into account (Eqs. 4 and 9), the deviation found for aspartic and glutamic analogues, and for maleic acid is largely reduced, whereas the value predicted for *N*-acetyl-*L*-histidine amide remains unaffected.
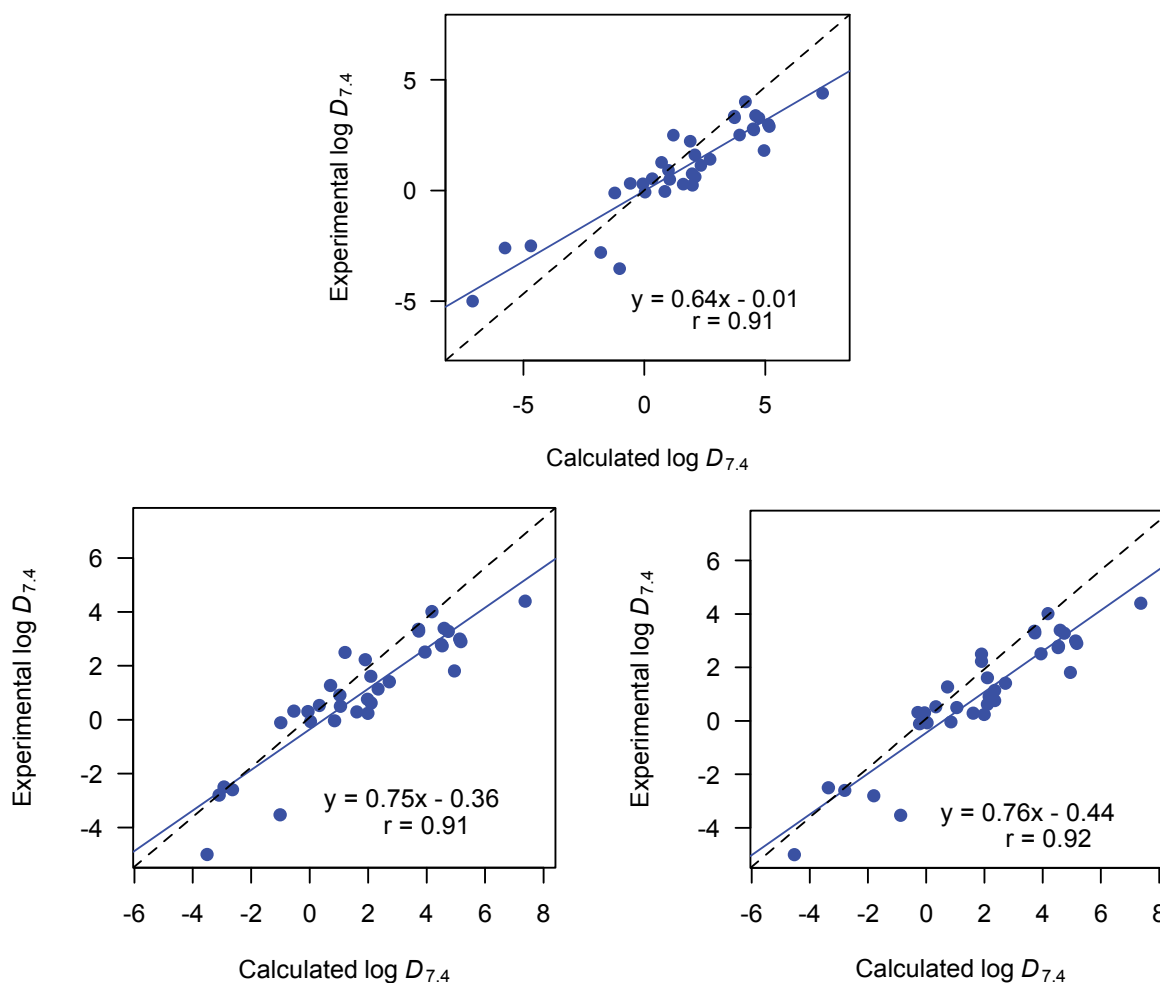
Figure 16.   Comparison between experimental (determined by the shake-flask method) and calculated log $D_{7.4}$ for 35 ionizable small molecules. Computed values were derived from IEFPCM/MST (blue) calculations using eqs 5 (top), 6 (bottom left), and 30 (bottom right).

To further check the reliability of the log P_I values, the calculated log P_N values were replaced by the experimental ones, and the log $D_{7.4}$ was determined using eqs 5, 6, and 30. The distribution coefficients obtained by limiting the IEFPCM/MST calculation to the partition coefficient of the ionized species (log P_I) reduces the rmsd between predicted and experimental data to ~ -0.8 (log D units), and the correlation coefficient is increased to 0.96 (Table 8). Note that these statistical parameters compare well with the values estimated using empirical methods (ACD/I-Lab, ChemAxon), as noted in Table 8 and Figure 17.

Table 8. Statistical Parameters of the Comparison between Experimental and Calculated log $D_{7.4}$ Values for the Series of 35 Small Molecules.

Method	mse	mue	rmsd	r
ACD/ I-Lab	0.0	0.5	0.8	0.95
ChemAxon	0.2	0.5	0.8	0.95
IEFPCM/MST, eq 5	-0.6	1.3	1.6	0.91
IEFPCM/MST, eq 5 (exptl. log P_N)	0.1	0.5	0.8	0.96
IEFPCM/MST, eq 6	-0.9	1.1	1.4	0.92
IEFPCM/MST, eq 6 (exptl. log P_N)	-0.3	0.4	0.6	0.96
IEFPCM/MST, eq 28	-0.8	1.1	1.4	0.91
IEFPCM/MST, eq 28 (exptl. log P_N)	-0.2	0.4	0.5	0.97
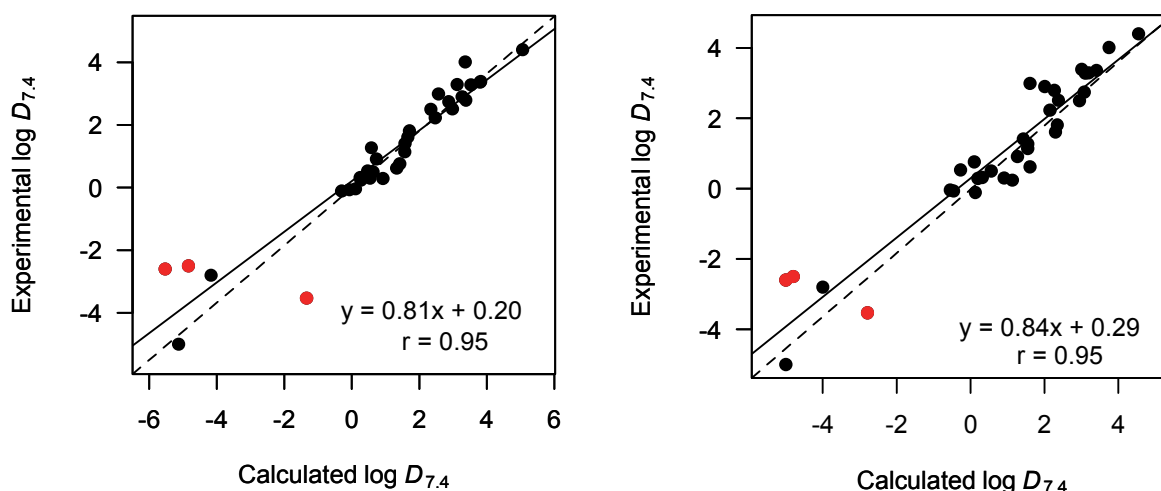
Figure 17. Comparison between experimental and calculated log $D_{7.4}$ for the set of 35 small molecules. Values were obtained using (left) ACD/ I-Lab and (right) ChemAxon. *N*-acetyl-*L*-aspartic acid amide, *N*-acetyl-*L*-glutamic acid amide and *N*-acetyl-*L*-histidine amide are shown as red dots.

Overall, these results give confidence to the partition values of ionic species determined with the refined IEFPCM/MST method, especially taking into account the limited extension of the model refinement, and the single-conformation approach adopted in present calculations. Nevertheless, the use of representative conformational ensembles may be required to obtain more accurate estimates of log $D_{7.4}$ in flexible molecules able to form distinct patterns of intramolecular interactions.[187–189]

3.6 Simulation of pH-Dependent Lipophilicity Profiles.

While the preceding results support the refined IEFPCM/MST method, there is generally little difference between the distinct formalisms previously tested (eqs 5, 6, and 30) for calculation of log $D_{7.4}$. This may reflect the fact that all molecules are approved drugs with high log P_N values, and that the log D was calculated at physiological pH, while the contribution of ionic species may be expected to be more

relevant at extreme pH values. Hence, we decided to determine the lipophilicity profile of 7 drugs and 4 amino acid analogues between pH 2 and 12, taking advantage of the experimental data about the pH-dependent partitioning of these compounds.[80,99,119,155,165]

For ibuprofen, warfarin and pentachlorophenol, the three formalisms give similar log P_N values at low pH, where the neutral species predominates (see Figure 18). However, the profiles diverge at intermediate pH values (between 6 and 8), following the increased population of the anionic species. Eq 5, which does not take into account the partition of the ionic species, gives rise to a profile that decreases steadily with increasing pH. In contrast, eqs 6 and 30 show an asymptotic behavior at basic pH.

For imipramine, amitriptyline, desipramine, and lidocaine all the methods exhibit the same log D at pH ~ 10, which arises from the partition of the neutral species. The pH-dependence of the profiles is similar up to acidic solutions (pH < 5), where the contribution of the cationic species is more important. Again, eq 5 shows a continuous decrease in log D with decreasing pH, whereas the profiles obtained from eqs 6 and 30 show the appearance of an asymptotic behavior at low pH. Note, however, that the appearance of this asymptotic value occurs at lower pH values for eq 30, leading generally to a larger deviation with regard to the experimental profile compared with the results obtained from eq 6, which reproduces well the general trends of the experimental sigmoidal profile. On the other hand, it is worth noting that the IEFPC/MST profiles obtained with eq 6 compare well with the pH-dependent profiles obtained from empirical methods (ACD/I-Lab and ChemAxon; see Figure 19).
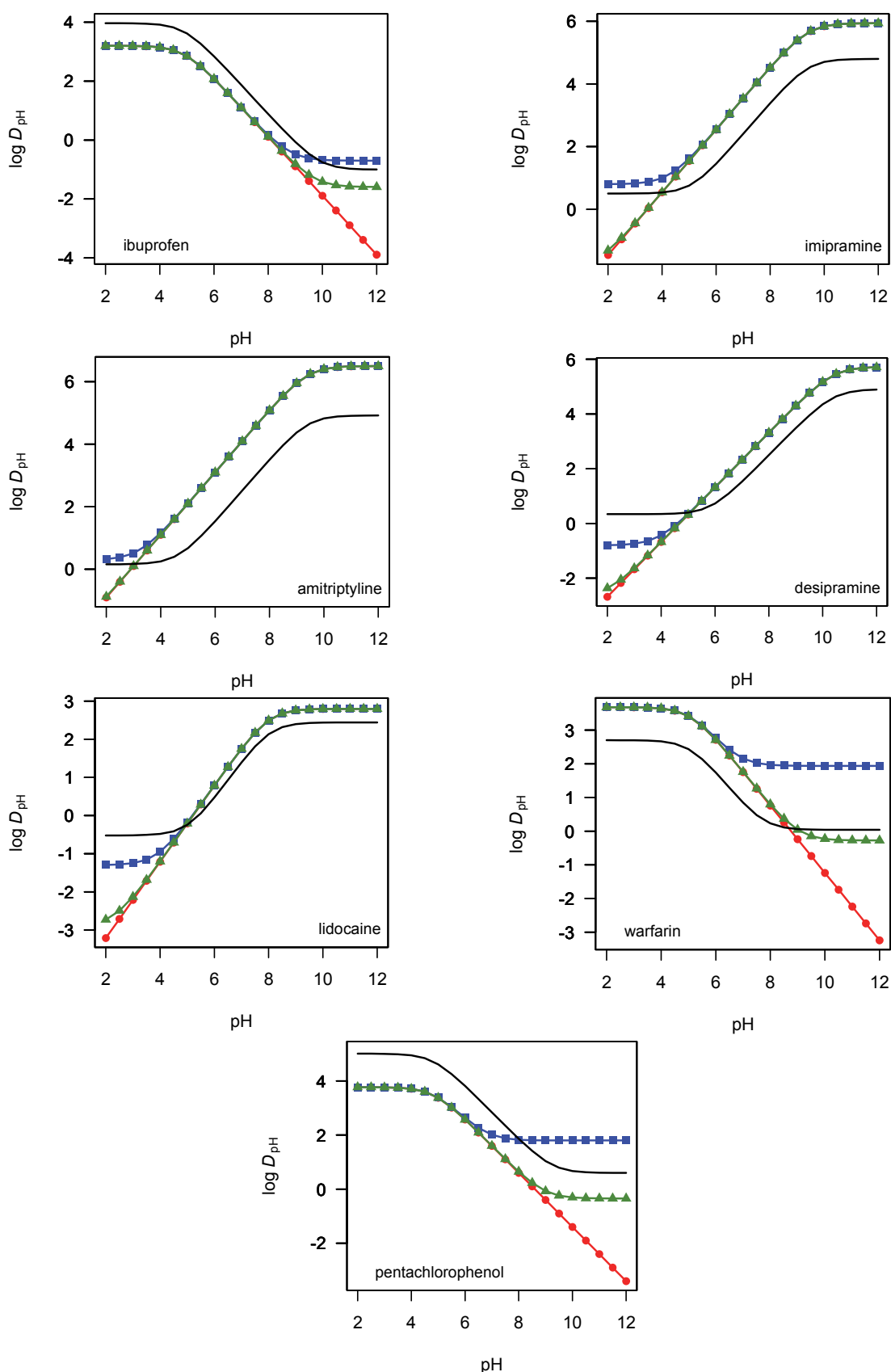
Figure 18. Comparison of *n*-octanol/water distribution coefficient profiles of selected drugs using eq 5 (red), 6 (blue), and 30 (green). The experimental data are shown in black.
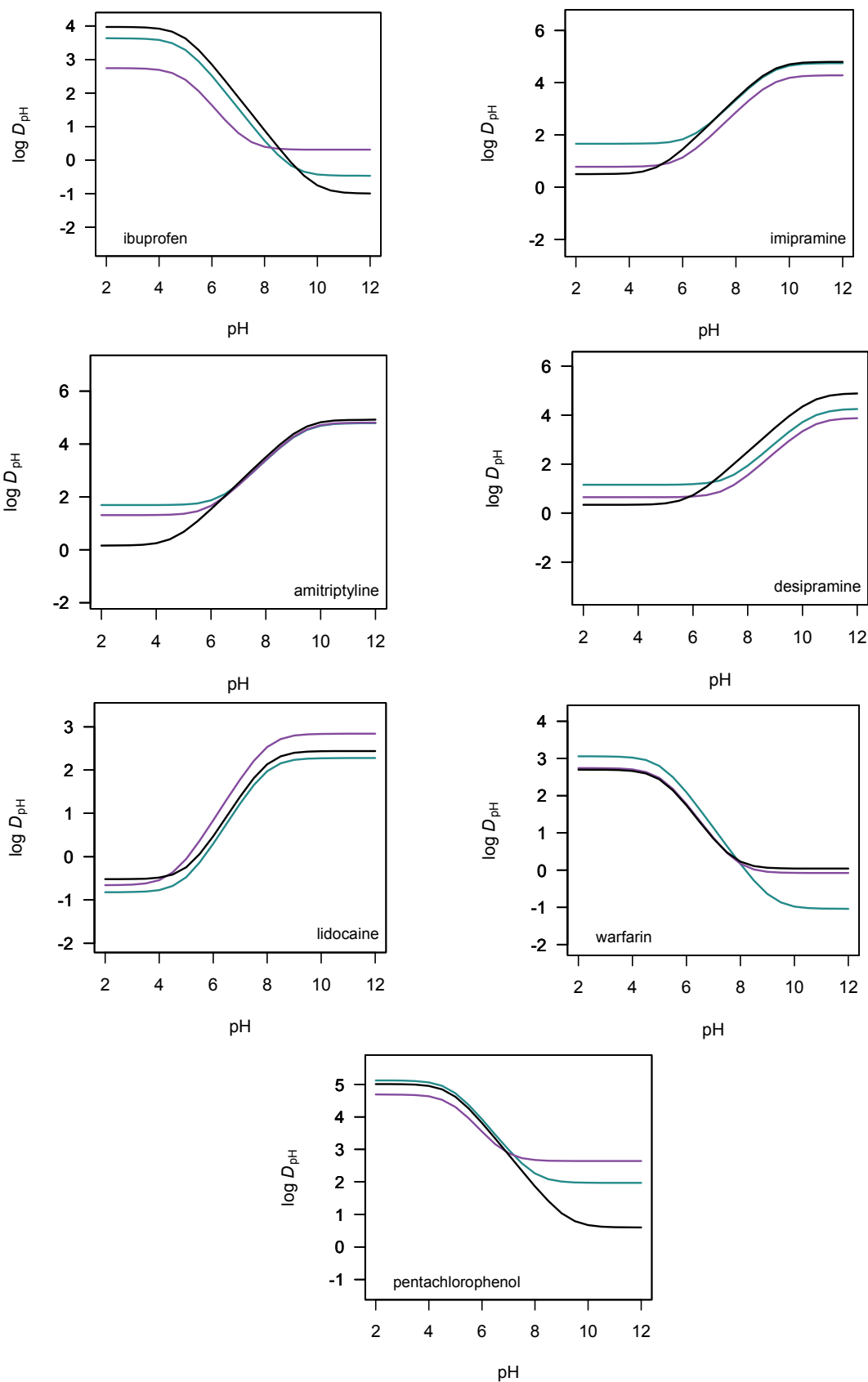
Figure 19. Comparison of *n*-octanol/water distribution coefficient profiles of selected drugs using ACD/I-Lab (cyan), ChemAxon (purple) and experimental data (black).

The pH-dependent lipophilicity profile determined for the set of amino acid analogues is shown in Figure 20.
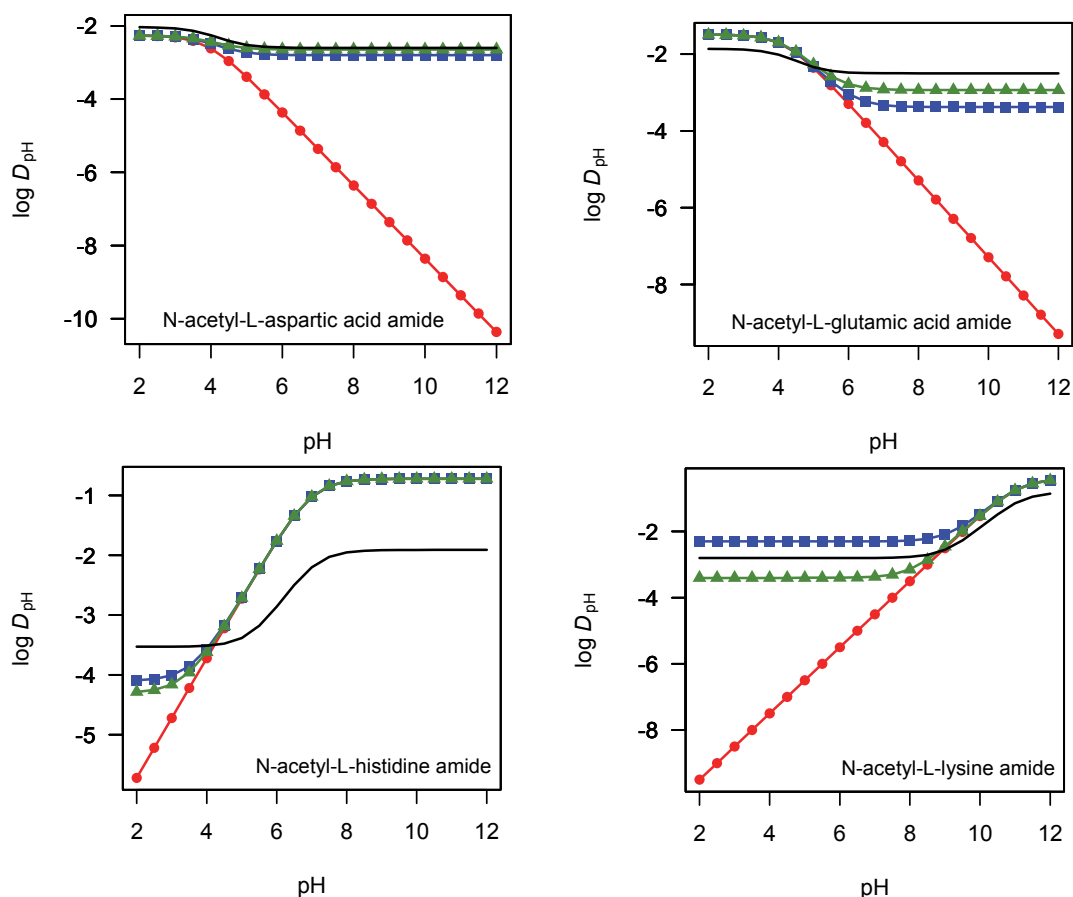
Figure 20. Comparison of *n*-octanol/water distribution coefficient profiles of selected amino acid analogues using eq 5 (red), 6 (blue), 30 (green), and experimental data (black).

As noted above, eq 5 works worse at extreme pH, as expected from the neglect of the contribution arising from the partition of ionic species. On the other hand, eqs 6 and 30 give similar profiles that reproduce the experimental values for the whole range of pH values. Furthermore, for the analogues of aspartic and glutamic acids the IEFPCM/MST results obtained for eqs 6 and 30 improve the pH-dependent profiles obtained from empirical methods, which predict a much higher hydrophilic behavior for pH values larger than 5 and lower than 9 for aspartic/glutamic acid and lysine, respectively (compare Figure 20 and Figure 21).
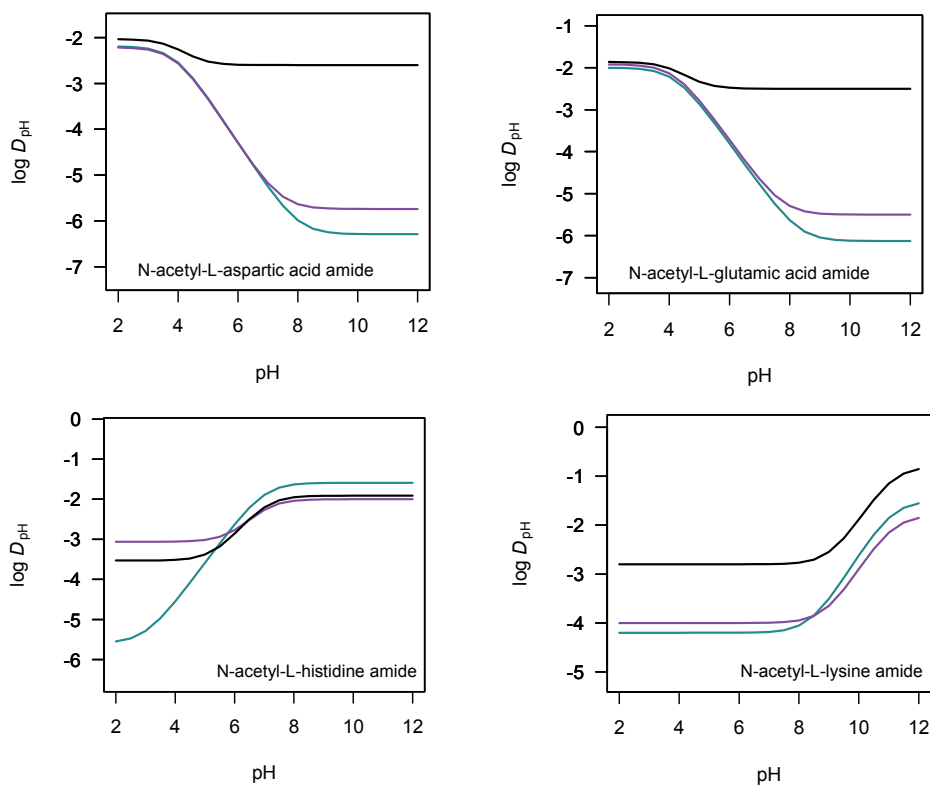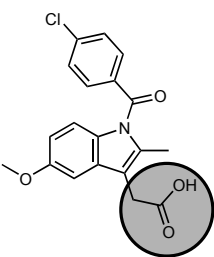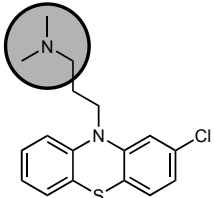
Figure 21. Comparison of *n*-octanol/water distribution coefficient profiles of selected amino acid analogues using ACD/I-Lab (cyan), ChemAxon (purple) and experimental data (black).

Overall, the results support the suitability of eq 6 for estimating the pH-dependence of the distribution profiles of ionizable compounds. The limited success found for eq 30 is surprising, especially when one takes into account the results obtained for amino acid analogues, but it may reflect the marked influence of inorganic ions on the experimental measurements of the distribution coefficient of ionized compounds.[118,121,124,190,191] In general, shake-flask experiments are performed in wet *n*-octanol/water systems using 0.15 mol/L KCl or NaCl, and reliable values for the inorganic standard partition coefficients have been reported. [25,60,61] However, it is also known that the distribution coefficient can be expected to increase substantially when a more hydrophobic cation is added in excess to the system, as has been reported for both ibuprofen and pentachlorophenol.[80,99,119] Finally, the potential contribution due to the formation of ion pairs may also have a significant effect of the distribution coefficient of ionized compounds, especially when the salt concentration is large enough relative to the ionized compound.[118,125]

3.7 Verification Experimental of the Effect of Background Salt Concentration in the Partition of Ionic Species.

The effect of the background salt in the determination of distribution coefficients has been actually discussed before (see Scheme 8). Here, we have analyzed the effect of the concentration of KCl in the experimental determination of log P_N and log P_I for indomethacin (acid compound; **65**) and chlorpromazine (basic compound; **82**) using the potentiometric method (see Table 9).

Table 9. Potentiometric Determination of the log P_N and log P_I Using Different Background Salt Concentrations (KCl) in the Aqueous Phase for Deriving the Standard Partition of the Organic Ionic Species.

Compound	Aqueous KCl (mol/L)	log P_N	log P_I	Derived log P_I^o (r_{C_I})	
				eq 35	eq 36
(65)	0	4.27±0.03	n.d[a]	- (0)	- (0)
	1x10^{-5}	4.47±0.03	n.d[a]	- (≈0)	- (≈0)
	1x10^{-3}	4.45±0.09	1.01±0.18	0.2 (0.6)	1.0 (0.6)
	0.15	4.59±0.10	1.47±0.13	0.5 (1.0)	0.5 (1.0)
(82)	0	5.05±0.03	n.d[a]	- (0)	- (0)
	1x10^{-5}	5.13±0.02	n.d[a]	- (0)	- (0)
	1x10^{-3}	5.20±0.01	0.70±0.03	2.0 (0.6)	2.8 (0.6)
	0.15	5.02±0.04	1.14±0.07	2.1 (0.6)	2.1 (0.6)

[a] no detectable (n.d; log P_I < -1)

This work is being performed in collaboration with the PhysChem Group (Faculty of Chemistry, University of Barcelona), and we only report preliminary results collected up to now. The partition of the neutral species is poorly affected by the absence or presence of different concentrations of KCl in the aqueous phase. However, as expected, the partition of ionic species was more sensible to the background salt, it being affected as a function of the concentration of KCl, especially at high concentrations (0.15 and 1×10^{-3} mol/L). In the case of low salt concentrations, no partition for the ionic species was detected.

At the time of submitting this thesis, present results do not suffice to calibrate the formalism that exploit the partition of the ionic species for both indomethacin and chlorpromazine. Additional potentiometric measurements (see Methods) will be recorded in future studies in order to obtain accurate values of apparent log P_I in all background salt concentrations studied.

3.8 Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations.

Taking advantage of the successful results for computing the distribution coefficients to physiological conditions as well as for reproducing lipophilicity profiles in amino acid analogues, it was planned to develop a new lipophilicity scale for amino acids. There are many lipophilicity scales reported in the literature, which have generally been designed under specific conditions for extracting structural information (i.e., distribution probability of amino acids for some kind of secondary structures) or preferences towards diverse phases (e.g., solution, stationary phase or membrane environments) from a given sequence, which may restrict the capacity of being extrapolated to other environments (e.g., solvents, sorbents, membranes or proteins) and conditions (e.g., pH or ionic strength).

The lipophilicity scale was built from theoretical computations that take into account the structural dependence of the conformational preferences of amino acids as well as the influence of pH in order to provide a consistent description of pH-adapted lipophilicity profiles in peptides and proteins. Here attention was paid to the set of natural amino acids, but the methodological strategy is intended to be easily adapted to nonstandard residues, such as nonproteinogenic residues, or to chemical modifications, such as phosphorylation, sulphonation and nitrosation, which regulate enzyme activity and signalling processes. Here, the theoretical formalism represented by eq 5 was selected to describe the distribution coefficient (D) of each residue using as model system the corresponding *N*-acetyl-*L*-amino acid amides.

Two schemes were explored for weighting the contribution of each conformational state to the differential solvation in the two solvents. In one case, P_N and P_I were determined using a Boltzmann`s weighting scheme to the relative stabilities of the conformational species of a given residue in the two solvents, leading to the *solvent-like* scale (SolvL). In the second scheme, named *protein-like* scale (ProtL), the contribution of each conformation was directly taken from the population distribution reported in the backbone-dependent conformational library. Therefore, these weighting schemes are expected to yield scales better suited for reflecting the lipophilic balance of amino acids well exposed to bulk solvent or in a protein-like environment, respectively. Finally, the effect of pH on the log D values was introduced from the experimental pK_as of ionizable residues in peptide models in aqueous solution[192,193] and in folded proteins[194,195] for the SolvL and ProtL scales.

The values of these lipophilicity scales for the amino acids at physiological pH are shown in Table 10 and Figure 22. ProtL data are averages of the log$D_{7.4}$ values determined separately for α-helix and β-sheet structures, which are depicted in Figure 22 (top). Taken Gly as reference, the ProtL scale comprises log$D_{7.4}$ values ranging from -3.91 (Arg) to 3.99 (Phe), reflecting the extreme values of hydrophilic residues (Arg, Asp, Glu and Lys), and hydrophobic ones (Trp, Phe) (see Figure 22). These trends are also found in the SolvL scale, even though the distribution of log$D_{7.4}$ values vary from -1.35 (Glu) to 2.62 (Phe). This trait is also found in other scales, as

knowledge-based methods generally give rise to a narrower range of lipophilicites compared to other experimental scales.[169] In our case, this arises from the distinct weighting factors used in ProtL and SolvL scales, leading to larger differences in the $\log D_{7.4}$ values of polar and ionizable amino acids, which show a preference for extended conformations (see Figure 23), likely reflecting the formation of stabilizing interactions (*e.g* salt bridges) or the solvent exposure to bulk water in proteins.[196,197]

Table 10. Solvent-like (SolvL) and Protein-like (ProtL) Lipophilicity Scales Based on the log D Values Determined for *N*-Acetyl-*L*-Amino Acid Amides at Physiological pH. The Experimental pK_a of Side Chain Ionizable Groups, and Calculated Partition Coefficients of Neutral (log P_N) and Ionized (log P_I) Residues are also Given.

Residue	Exp. pK_a		log P_N		log P_I		log $D_{7.4}$ [a]	
	SolvL	ProtL	SolvL	ProtL	SolvL	ProtL	SolvL	ProtL
Ala	-	-	-1.16	-2.47	-	-	-1.16 (0.85)	-2.47 (0.66)
Arg	*12.5[b]*	*12.5[b]*	*-2.86*	*-3.66*	*-2.99*	*-7.38*	*-2.99 (-0.98)*	*-7.04 (-3.91)*
Asn	-	-	-2.98	-3.97	-	-	-2.98 (-0.97)	-3.97 (-0.84)
Asp	*3.90[c]*	*3.50[d]*	*-2.26*	*-3.18*	*-2.80*	*-8.54*	*-2.80 (-0.79)*	*-5.87 (-2.74)*
Cys	*9.83[e]*	*6.80[d]*	*-0.16*	*-1.47*	*-4.19*	*-5.78*	*-0.16 (1.85)*	*-2.17 (0.96)*
Gln	-	-	-2.22	-4.00	-	-	-2.22 (-0.21)	-4.00 (-0.87)
Glu	4.20[c]	4.20[d]	-1.49	-3.79	-3.38	-6.20	-3.36 (-1.35)	-5.96 (-2.83)
Gly	-	-	-2.01	-3.13	-	-	-2.01 (0.00)	-3.13 (0.00)
His (δ)	*7.00[c]*	*6.60[d]*	*-1.20*	*-4.67*	*-4.06*	*-5.97*	*-1.35 (0.66)*	*-4.56 (-1.43)*
His (ε)	*7.00[c]*	*6.60[d]*	*-0.72*	*-4.98*	*-4.06*	*-5.97*	*-0.87 (1.14)*	*-4.97 (-1.84)*
Ile	-	-	-0.50	-0.38	-	-	-0.50 (1.51)	-0.38 (2.75)
Leu	-	-	0.05	-1.36	-	-	0.05 (2.06)	-1.36 (1.77)
Lys	*11.1[c]*	*10.5[d]*	*-0.40*	*-2.19*	*-3.24*	*-6.81*	*-3.18 (-1.17)*	*-5.08 (-1.95)*
Met	-	-	-0.51	-1.83	-	-	-0.51 (1.50)	-1.83 (1.30)
Phe	-	-	0.61	0.86	-	-	0.61 (2.62)	0.86 (3.99)
Pro	-	-	-0.77	-1.44	-	-	-0.77 (1.24)	-1.44 (1.69)
Ser	-	-	-2.04	-4.12	-	-	-2.04 (-0.03)	-4.12 (-0.99)
Thr	-	-	-1.22	-3.01	-	-	-1.22 (0.79)	-3.01 (0.12)
Trp	-	-	0.33	0.16	-	-	0.33 (2.34)	0.16 (3.29)
Tyr	*10.3[c]*	*10.3[d]*	*-0.49*	*-1.80*	*-4.21*	*-9.59*	*-0.49 (1.52)*	*-1.80 (1.33)*
Val	-	-	-0.93	-1.68	-	-	-0.93 (1.08)	-1.68 (1.45)

[a] Values for ionizable residues are shown in italic. Log $D_{7.4}$ values relative to glycine are given in parenthesis. [b] Ref 192. [c] Ref 193. [d] Ref 194. [e] Ref 195.
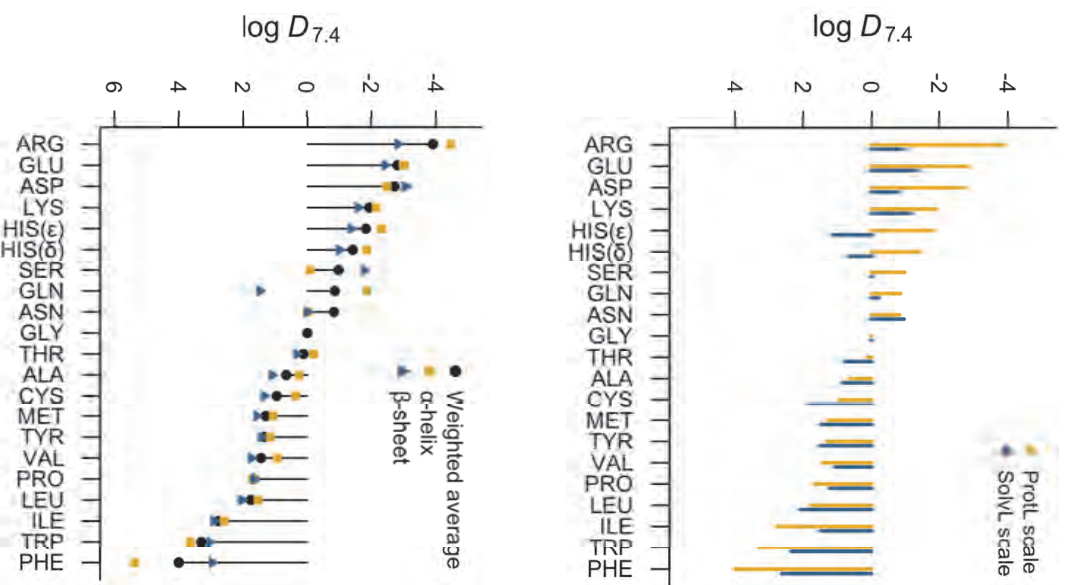
Figure 22. Representation of SolvL (blue) and ProtL (yellow) lipophilicity scales (values relative to Gly) at physiological pH (top). ProtL scale values for twenty-one amino acid residues (relative to Gly). Black circle, orange box and green triangle represent the $\log D_{7.4}$ values for total, α-helix and β-sheet conformers respectively (bottom).
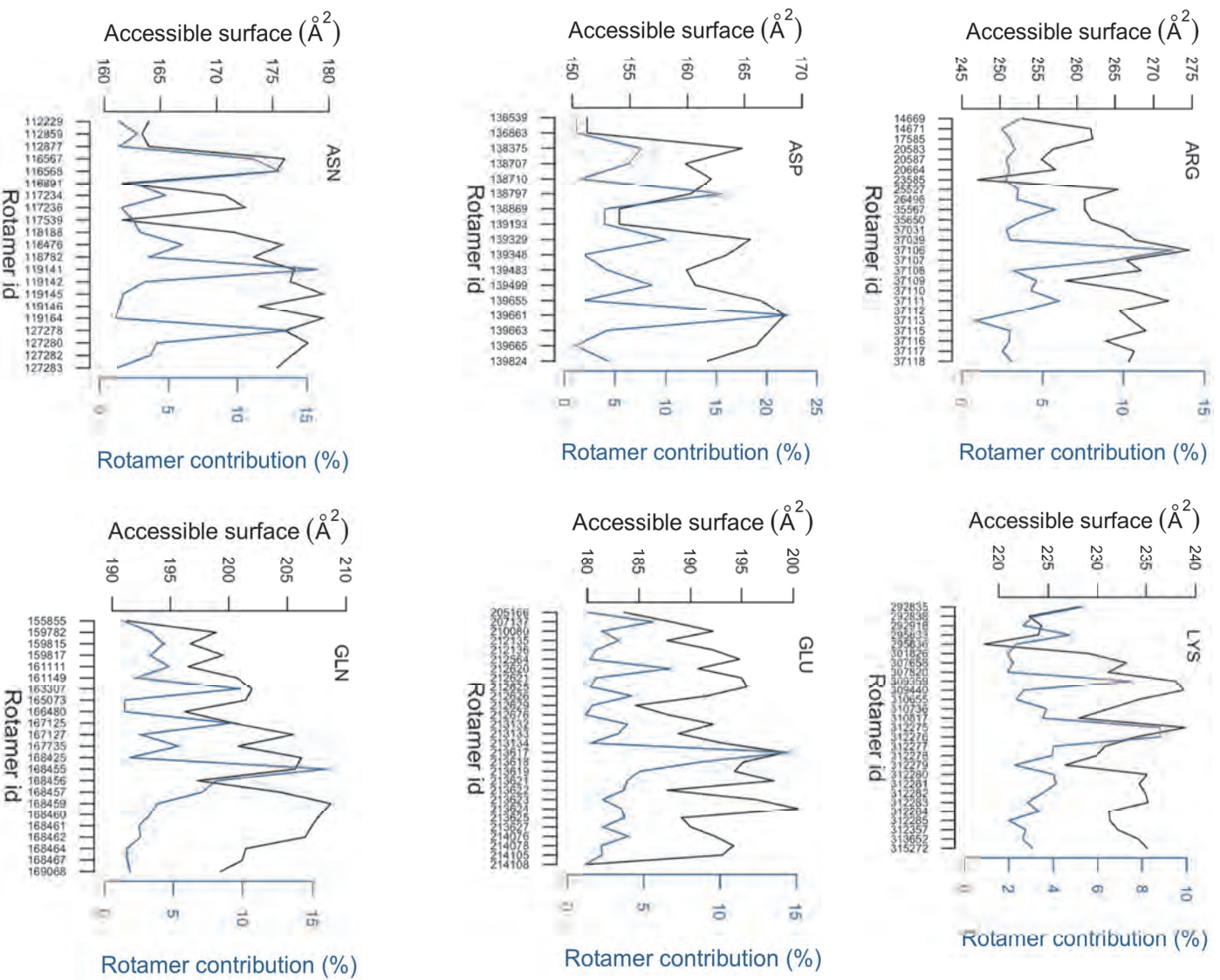
Figure 23. Distribution of the accessible surface area (Backbone + Side Chain) for conformational species of Arg (top, left), Lys (top, right), Asp (middle, left), Glu (middle, right), Asn (bottom, left) and Gln (bottom, right) using the Dunbrack's backbone-dependent conformational library. The weigth of each rotameric species in this conformational library is also shown (blue line).

The sensitivity of the lipophilicity of ionizable residues to pH changes is shown in Figure 24, which compares the log D values at pH 2.1, 7.4 and 9.0, chosen as representative values of the pH changes along the gastrointestinal tract. The hydrophilicity of acid/basic amino acids is enhanced at basic/acidic pHs, as expected from the predominance of the ionic species. In the SolvL scale, it is worth noting the hydrophilic nature of protonated His at acidic pH, and the slight hydrophobicity of protonated Glu. In contrast, the ProtL scale exhibits a higher sensitivity to pH, as noted in the large changes in the log D values of Asp and Glu, which are decreased 2-3 log D units upon deprotonation, the reduced hydrophilicity of Lys at basic pH, and the change from hydrophobic (at acid and physiological pH) to hydrophilic (at basic pH) of Cys. This reflects the ability of these scales to reflect the pH influence on the lipophilicity of ionizable residues, which may be affected by the local environment in proteins.[198,199]
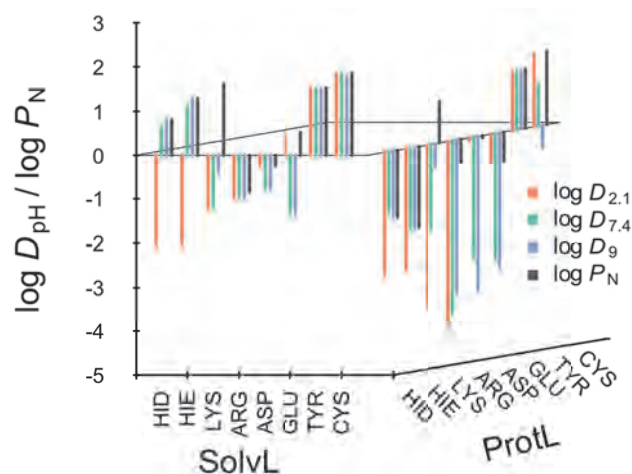
Figure 24. Representation of the pH dependence of the SolvL and ProtL lipophilicity scales for ionizable amino acids (values relative to Gly). Values determined at pH of 2.1, 7.4 and 9.0 are shown in orange, green and gray, respectively, and the values of the neutral Species (log P_N) are shown in black.

To calibrate the suitability of these scales, comparison was made with the log $D_{7.4}$ values reported by Fauchère and Pliska,[155] which were experimentally determined from the partitioning of *N*-acetyl-*L*-amino acid amides between *n*-octanol and water at physiological pH (Figure 25).
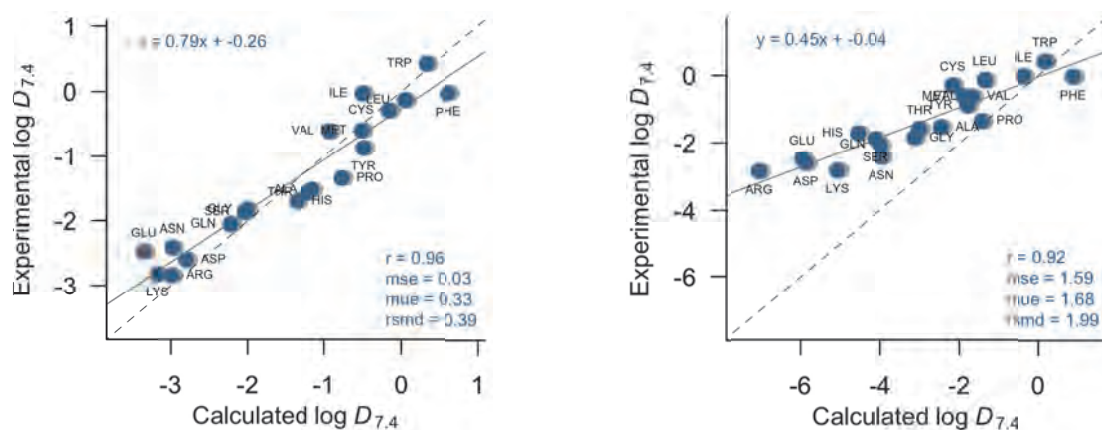
Figure 25. Comparison between (left) SolvL and (right) ProtL lipophilicity scales derived from the IEFPCM/MST solvation model (expressed as log $D_{7.4}$) and Fauchère-Pliska experimental values for the twenty *N*-acetyl-*L*-amino acid amides (*r*: pearson correlation coefficient; mse: mean signed error; mue: mean unsigned error; rmsd: root-mean square deviation).

Comparison with the SolvL values gives satisfactory results, as noted in a correlation coefficient (*r*) of 0.96 and a mean unsigned error (mue) of 0.33 log$D_{7.4}$ units for a set of experimental values ranging from -3.36 to 0.61. The correlation coefficient is slightly worse (*r* = 0.92) and the mue increases to 1.68 for the ProtL scale. For the sake of comparison, the same analysis was performed by using log$D_{7.4}$ values obtained from computations with the SMD solvation model,[137] in conjunction with the two weighting schemes, and the results also revealed a better performance for the solvent-adapted scheme (*r* = 0.85, mue = 0.83; Figure 26). On the other hand, the SolvL scale also performed better than the empirical estimates of log$D_{7.4}$ obtained from ACD/ILab[131] (*r* = 0.88, mue=0.60) and ChemAxon[132] (*r* = 0.92, mue=0.65) when compared with the experimental values reported by Fauchère and Pliska (Figure 26).

Figure 26. Comparison between Fauchère-Pliska experimental log $D_{7.4}$ values and theoretical estimates obtained by using the SMD model with (top left) solvent-adapted, (top right) protein-adapted weighting factors, (bottom left) ACD/I-Lab and (bottom right) ChemAxon for the twenty N-acetyl-L-amino acid amides (r: Pearson correlation coefficient; mse: mean signed error; mue: mean unsigned error; rmsd: root-mean square deviation).

Table 11 shows the comparison of the SolvL and ProtL lipophilicities with other experimental scales, including four *bulk solvent*-based scales (Fauchère-Pliska,[155] Eisenberg-McLachlan,[157] Hopp-Woods,[164] Wimley *et al.*[165]), two biological-derived (Moon-Fleming,[167] Hessa *et al.*[168]) and two knowledge-based (Koehler *et al.*,[169] Janin *et al.*[170]) scales, and a consensus (Kyte-Doolittle[166]) one. The bulk solvent-based scales rely on experimental measurements of the transfer between *n*-octanol and water (Fauchère-Pliska, Eisenberg-McLachlan) at physiological pH or at basic conditions (pH = 9.0; Wimley *et al.*), and between ethanol and the vapor phase (Hopp-Woods). Excellent correlations are found with Fauchère-Pliska, Eisenberg-McLachlan, and Hopp-Woods scales ($0.89 < r < 0.92$), whereas a worse correlation ($r \approx 0.60$) is found with Wimley *et al.* scale. However, this can be attributed to the formation of salt bridges between Arg/Lys residues with the terminal carboxyl group in *n*-octanol for the AcWL-X-LL pentapeptides used as model systems, as noted by ^{13}C-NMR studies.[158] Exclusion of Arg and Lys enhances the correlation coefficient to 0.87. On the other hand, the bulk solvent-based lipophilicities are consistently closer to the values collected in the SolvL scale (mue of 0.36-0.92 log *P/D* units) than to the ProtL ones (mue of 0.84-1.24 log *P/D* units).

The correlation coefficients obtained with biological-, knowledge-based and consensus scales are still satisfactory ($0.74 < r < 0.94$), but tend to be lower than the values obtained with the bulk solvent-based transfer scales. This is not unexpected keeping in mind that the lipophilicites are derived from statistical analysis of topological distributions of residues in proteins (Koehler *et al.*, Janin *et al.*), or from complex biochemically-adapted assays, such as the transfer of amino acids from water to a phospholipid bilayer (Moon-Fleming), the recognition of artificial helices by the Sec61 translocon (Hessa *et al.*), or the combination of water-vapor transfer free energies with the interior-exterior distribution of amino acids in the consensus (Kyle-Doolittle) scale.

Table 11. Statistical Parameters of the Comparison of the SolvL and ProtL Scales with other Lipophilicity Scales. Comparison Was Made Using the Values Adapted to the Specific pH of each Scale and Relative to Gly.

Scale[a]	SolvL				ProtL			
	mse[b]	mue	rsmd	*r*	mse	mue	rsmd	*r*
Bulk-Solvent Adapted Scale								
Fauchère-Pliska	-0.20	0.36	0.46	0.94	0.36	0.98	1.28	0.92
Eisenberg-McLachlan	-0.20	0.44	0.57	0.90	0.36	1.08	1.35	0.91
Hopp-Woods	-0.49	0.60	0.74	0.91	0.07	0.84	1.08	0.89
Wimley et al.[c]	-0.60 (-0.87)	1.02 (0.92)	1.16 (1.03)	0.59 (0.87)	0.04 (-0.30)	1.24 (1.03)	1.64 (1.25)	0.61 (0.87)
Biological-Based Scale								
Moon-Fleming	-0.12	0.57	0.67	0.94	0.24	0.72	0.93	0.91
Hessa et al.	-0.92	0.93	1.18	0.79	-0.36	1.08	1.46	0.82
Knowledge-Based Scale								
Koehler *et al.*	-0.91	1.10	1.33	0.78	-0.35	1.55	1.87	0.80
Janin et al.	-1.06	1.11	1.32	0.78	-0.51	1.36	1.71	0.74
Consensus Scale								
Kyte-Doolittle	-0.81	1.43	1.71	0.72	-0.25	1.13	1.41	0.78

[a] A physiological pH was considered in all cases, but for Wimley et al. and Moon-Flemin., since the corresponding pH was fixed at 9.0 and 3.8 following the specific experimental conditions.
[b] mse: mean signed error, mue: mean unsigned error, rmsd: root-mean square deviation, *r*: Pearson correlation coefficient. mse, mue and rmsd are given in log P_N/D units.
[c] Values in parenthesis obtaining upon exclusion of Arg and Lys. Since this scale was built up using model pentapeptides (AcWL-X-LL) at pH 9.0, Arg and Lys formed a salt bridge with the terminal carboxyl group in *n*-octanol as noted by ^{13}C-NMR studies.[158]

It is worth noting that by using the relatively simple *n*-octanol/water system for deriving the lipophilicity of amino acids, even subject to a controversial capacity as a mimic of biological environments, it was possible to correlate efficiently present values with those reported by Moon-Fleming and Hessa et al., which were compiled from more complex biological systems (Figure 27). Surprisingly, the Moon-Fleming scale, where the chemical equilibrium occurs between the unfolded hydrated protein to those that folds in a β-barrel inside a membrane, was practically emulated. The outlier is represented by the residue arginine (Arg) whose lipophilicity is greater in

the Moon-Fleming scale than in our ProtL scale, maybe due to stabilization of other parts of the protein and/or a deformation of the membrane to reduce its hydrophobic thickness.[167,200] In the case of Hessa et al. scale, proline (Pro) introduces the biggest deviation in the trend. This may be rooted to the nature of the experimental measurements made by Hessa et al., which represents a translocon (protein complex) to bilayer partitioning. The segment insert into the membrane (H-segment) adopts a α-helix structure, and therefore Pro is being classified as hydrophilic residue because is a strong helix breaker. In the Elofsson review[161] about hydrophobic scales, it is stated that better scales classify Pro as a hydrophilic residue. However, Rossky[201] advocates that Pro should be a hydrophobic residue. The lack of agreement between authors highlights the different nature of each hydrophobic scale. As a matter of fact, Elofsson do their comparison of scales on the preferences of transmembrane helices by certain amino acids, as a result of a biological perspective of hydrophobicity. On the other hand, the approach by Rossky et al. is based on studies at atomic level to represent protein surfaces, cavities and pores. Our scale considers Pro to be slightly hydrophobic.
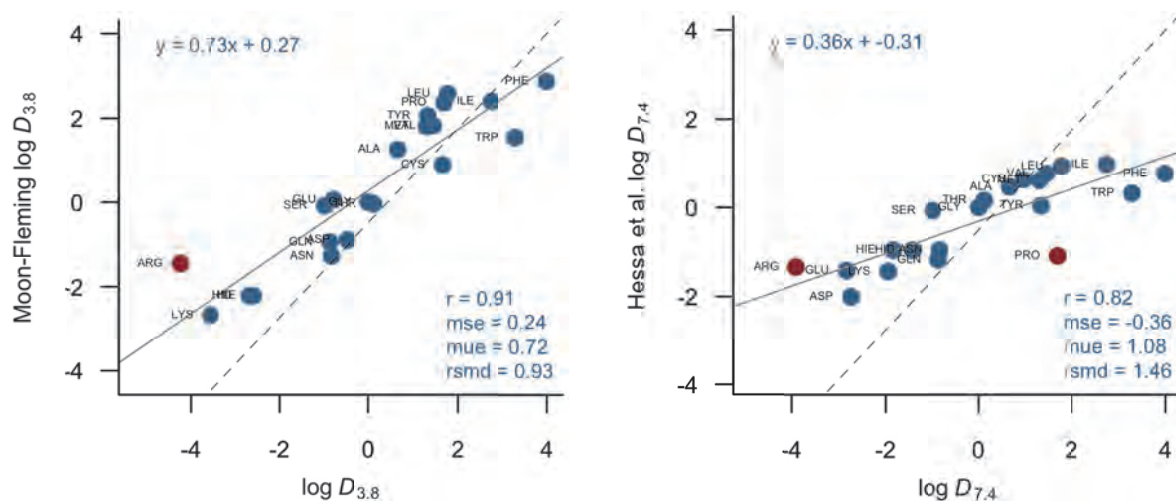
Figure 27. Comparison between *ProtL* Scale derived from the IEFPCM/MST solvation model for the twenty *N*-acetyl-*L*-amino acid amides and Moon`s (left) and Hessa`s (rigth) experimental biological scales. All values are expressed as log *D* relative to Gly. (*r*: Pearson correlation coefficient; mse: mean signed error; mue: mean unsigned error; rmsd: root-mean square deviation).

The sensitivity of the results to the pH was examined by extending the comparison to the lipophilicities determined for the SolvL and ProtL scales at pH values of 3.8, 7.4, and 9.0 (note that the acidic and basic pH values were chosen in the studies reported by Moon-Fleming and Wimley et al., respectively). In general, there is little difference between the correlation coefficients obtained at pH 7.4 and 9.0 (Figure 28). However, a larger effect is found in the comparison of the log $D_{3.8}$, as there is a general decrease in the correlation coefficient, which is remarkable for the bulk solvent-based transfer scales, especially in the case of Hoop-Woods and Wimley et al. The only exception is found in the comparison with the Moon-Fleming scale, as the highest correlation coefficient is found for the ProtL values corrected at pH 3.8. These findings support the suitability of the SolvL/ProtL scales to account for the pH influence on the lipophilicity of amino acids.
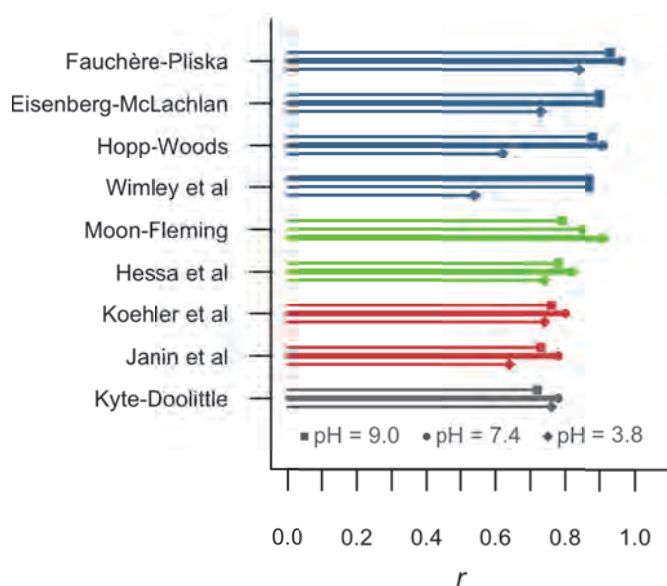
Figure 28. Representation of the Pearson correlation coefficient in the comparison of the SolvL scale with bulk solvent-based scales (blue lines), and ProtL scale with biological-based (green lines), knowledge-based (red lines) and consensus (gray lines) lipophilicity scales at pH 3.8, 7.4, and 9.0.

3.9 Validation of the Hydrophobic Effect Using the Theoretical Lipophilicity Scale.

In addition to determine a ProtL scale, the Dunbrack's backbone-dependent conformational library also allows us also to determine the solvent accessible surface area (SASA) for all rotamers using the program naccess[202]. The entire SASA for each rotamer was decomposed into the backbone and side chain contributions in order to examine the effect of transferring the nonpolar side chain of each apolar residue from *n*-octanol to water as a function on its SASA, the aforementioned *"hydrophobic effect"*. Furthermore, since the IEFPCM/MST model (see Scheme 5) permits to decompose the total free energy of transfer between *n*-octanol and water into electrostatic (ΔG_{ele}^{ow}) and non-electrostatic components (cavitation ΔG_{cav}^{ow} and van der Waals ΔG_{vW}^{ow} terms), we also evaluated the non-electrostatic cost of transfering the side chain for all amino acids ($\Delta\Delta G_{cav+vW}^{ow}$) per accessible surface area (see Figure 29).
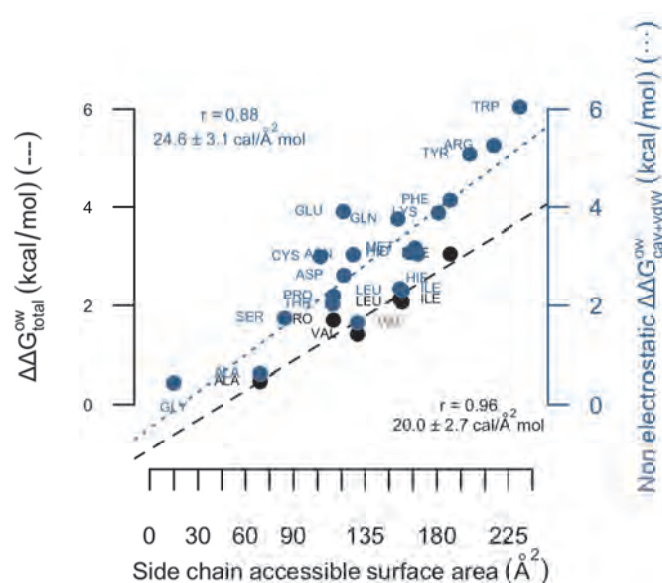
Figure 29. Correlation of the average transfer free energy of *n*-octanol to water for the side chain of amino acids with the average of solvent accessible surface area in the *N*-acetyl-*L*-amino acid amides conformers. (Dashed line depicts the average transfer free energy of water to *n*-octanol for the side chain of nonpolar amino acids and dotted line depicts the average non-electrostatic term in the transfer free energy of *n*-octanol to water for the side chain in the twenty natural amino acids).

Our computations using those two approaches reveal that the hydrophobic effect amounts to 20.0 ± 2.7 cal/mol Å² and 24.6 ± 3.0 cal/mol Å² to the stability of removing apolar side chains or non-electrostatic terms from water to *n*-octanol, respectively. Our average hydrophobic effect (22.3 cal/mol Å²) matches previously experimental values (see Table 2). Therefore, despite of explaining the stability of proteins in solution or inside membranes (Moon-Fleming), the hydrophobic effect also works in the stabilization of amino acid analogues-determinate experimentally (Fauchère- Pliska[155]) or computed (our approach) in solvent systems to an equivalent degree.

3.10 Application of the Lipophilicity Scale to Small Peptides Properties.

The reliability of the SolvL/ProtL scales has been calibrated by comparing the cumulative lipophilicity with the (RP-HPLC) retention time determined for different sets of peptides.[203,204] Given the small size of the peptides (≤ 13 residues) and the lack of well defined secondary structures, non-additivity effects can be expected to play a minor role.[205] Accordingly, the cumulative lipophilicity was determined assuming an additive scheme (see Methods).

The first test comprises eight 10-mer peptides with equal charge that differ in the content of hydrophobic residues (see Table 12).[206]

Table 12. Experimental RP-HPLC Retention Time for Eight Model Decapeptides and Cumulative Lipophilicity Determined with the SolvL and ProtL Lipophilicity Scales.

Peptide [a]	Sequence	Retention factor k` (min)	$\log D_{7.4}$	
			SolvL	ProtL
Pep1Leu	DKDKGGGGLG	4.80	-17.09	-34.04
Pep2Leu	DKDKGGGLLG	11.97	-15.03	-32.27
Pep3Leu	DKDKGGLLLG	16.22	-12.97	-30.50
Pep1Cys	DKDKGGGGCG	0.52	-17.30	-34.85
Pep1Ile	DKDKGGGGIG	4.73	-17.64	-33.06
Pep1Met	DKDKGGGGMG	2.27	-17.65	-34.51
Pep1Phe	DKDKGGGGFG	6.11	-16.53	-31.82
Pep1Val	DKDKGGGLVG	1.86	-18.07	-34.36

[a] Ref. 206

The SolvL cumulative lipophilicity yields a correlation coefficient of 0.96 (Figure 30), which compares with the value estimated from the hydrophobic surfaces of peptides derived from molecular dynamics simulations (r = 0.97),[206] whereas a slightly lower correlation was found for the ProtL scale (r = 0.91; see Table 13). For this simple set of homogeneous peptides, most of the experimental lipophilicity scales generally yielded correlations higher than 0.9 (see Table 13).

Figure 30. Relationship Between the Cumulative Lipophilicities Determined from the SolvL Scale Versus (A) the Retention Time for Eight 10-mer Peptides (pH 7.4; Ref. 38), (B) 248 Unique 13-mer Peptides (pH 2.1; Ref. 207,208), (C) log P_N for 118 Random Peptides (Ref. 42), and (D) log $D_{7.4}$ for 116 Random Peptides (Ref. 209).

A more challenging test is the set of 248 analogues of the influenza virus hemagglutinin 13-mer peptide (98–110) Ac-YPYDVPDYASLRS-NH$_2$, with equal length, but different net charge at the experimental acidic conditions (pH = 2.1),[207,208] comprising 36 peptides with two charged amino acids (Arg combined with His or Lys), 105 peptides with a single charged residue (Arg, Lys, or His), and finally 17 neutral peptides. The SolvL cumulative lipophilicity correlates satisfactorily with the retention time determined for the whole set of peptides (r = 0.85; Figure 30B). Among bulk solvent-based scales, Fauchère-Pliska, Eisenberg-McLachlan and Hopp-Woods also provided reasonable correlations coefficients (0.74 < r < 0.85; Table 13 and Figure 31), but a worst correlation was found for Wimley et al., although this may be attributed to the different pH used in this latter scale (pH = 9.0) and the experimental assay conditions (pH = 2.1). The performance of biological-, knowledge-based and consensus scales was also worse (0.55 < r < 0.64; Table 13 and Figure 31), but for Moon-Fleming (r = 0.78), likely reflecting the acidic pH conditions considered in the derivation of this lipophilicity scale.

Table 13. Correlation of Retention Time for 8 Model Decapeptides with the Same Charge,[206] and for 218 Peptides[207,208] with Three Different Charge States Using the Cumulative Hydrophobicity with Our Adaptive Hydrophobicity Scale and with Others Experimental Scales.

Scale	Set of 10-mer Peptides (pH 7.4)	Set of 13-mer Peptides (pH 2.1)
Fauchère-Pliska	0.96	0.85
Eisenberg-McLachlan	0.95	0.79
Hopp-Woods	0.99	0.74
Wimley et al.	0.99	0.36
Moon-Fleming	0.99	0.78
Hessa et al.	0.96	0.61
Koehler et al.	0.76	0.64
Janin et al.	0.39	0.55
Kyte-Doolittle	0.93	0.60
ProtL	0.96	0.85
SolvL	0.91	0.77

Toward Refined Theoretical Models for the Description of Lipophilicity in Biomolecules
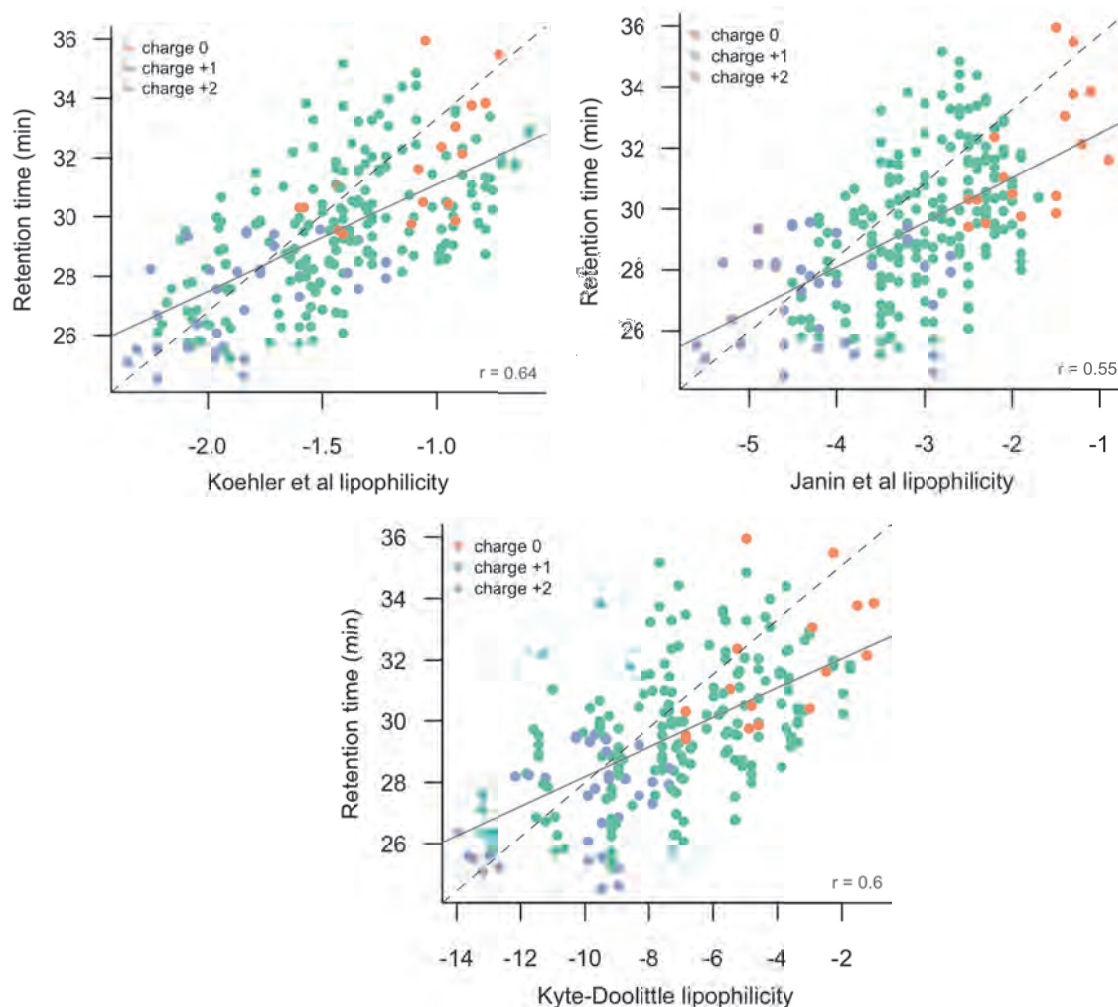
Figure 31. Representation of the RP-HPLC retention time of 248 13-mer peptides (Refs. 207,208) versus the lipophilicity determined from experimental scales in Table 13.

Then, given the relevance of partition (log P_N)/distribution (log $D_{7.4}$) coefficients for ADMET properties of peptides,[148] the suitability of the SolvL scale was further checked for reproducing the differences in log P_N/log $D_{7.4}$ of a set of random peptides.[209] The SolvL-based additive scheme yielded promising results, as noted in r values of 0.93 and 0.83 in reflecting the experimental range of logP_N and log$D_{7.4}$ for sets of 118 and 116 peptides, respectively (Figure 32 C, D). Compared to experimental scales, a similar predictive power was attained for Fauchère-Pliska and Eisenberg-McLachlan scales ($r \approx 0.90$) for the set of 118 log P_N data, and for Hopp-Woods ($r = 0.88$) for the set of 116 log$D_{7.4}$ values, but with a larger mue (around 2.3 versus 0.7 for the SolvL scale; Tables 14 and 15).

Table 14. Statistical Parameters of the Comparison[a] of the SolvL and ProtL Scale with Others Hydrophobicity Scales Against log P_N Values for 118 Random Peptides.

Scale	r	mse	mue	rsmd
Fauchère-Pliska	0.90	-2.53	2.53	2.64
Eisenberg-McLachlan	0.89	-2.29	2.29	2.38
Hopp-Woods	0.74	-2.07	2.11	2.31
Wimley et al.	0.70	-1.54	1.67	1.81
Moon-Fleming	0.69	-0.80	1.12	1.34
Hessa et al.	0.22	0.29	0.98	1.29
Koehler et al.	0.45	-0.35	0.87	1.12
Janin et al.	0.38	-0.65	1.08	1.28
Kyte-Doolittle	0.50	-2.85	3.00	3.60
ProtL	0.60	1.35	1.68	2.00
SolvL	0.93	-0.55	0.71	0.94

[a] mse: mean signed error, mue: mean unsigned error, rmsd: root-mean square deviation, r: Pearson correlation coefficient. mse, mue and rmsd are given in log P_N/D units.

Table 15. Statistical Parameters of the Comparison[a] of the SolvL and ProtL Scale with Others Hydrophobicity Scales Against log $D_{7.4}$ Values for 116 Random Peptides.

Scale	r	mse	mue	rsmd
Fauchère-Pliska	0.76	-2.76	2.76	2.88
Eisenberg-McLachlan	0.75	-2.58	2.58	2.69
Hopp-Woods	0.88	-2.32	2.33	2.43
Wimley et al.	0.52	-1.94	1.94	2.23
Moon-Fleming	0.79	-1.16	1.24	1.48
Hessa et al.	0.72	-0.22	0.60	0.73
Koehler et al.	0.76	-0.90	1.01	1.19
Janin et al.	0.61	-1.12	1.21	1.38
Kyte-Doolittle	0.52	3.04	3.17	3.76
ProtL	0.79	1.46	1.82	2.11
SolvL	0.83	-0.52	0.73	0.95

[a] mse: mean signed error, mue: mean unsigned error, rmsd: root-mean square deviation, r: Pearson correlation coefficient. mse, mue and rmsd are given in log P_N/D units.

In these test cases, the ProtL scale performed worst ($0.60 < r < 0.91$; Figure 32) than the SolvL one, suggesting that the Boltzmann-weighting scheme is better suited for describing the lipophilicity of residues in structureless peptides.
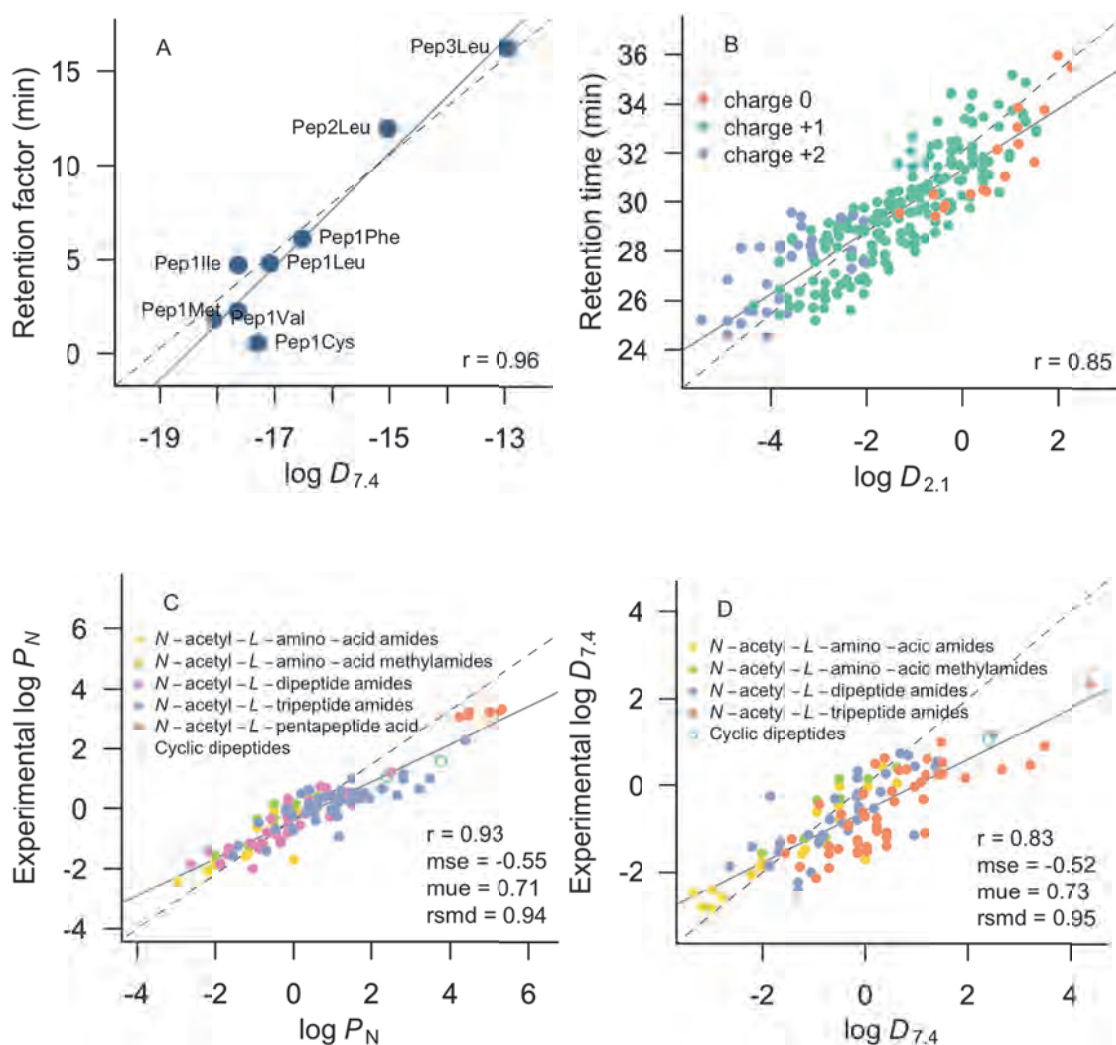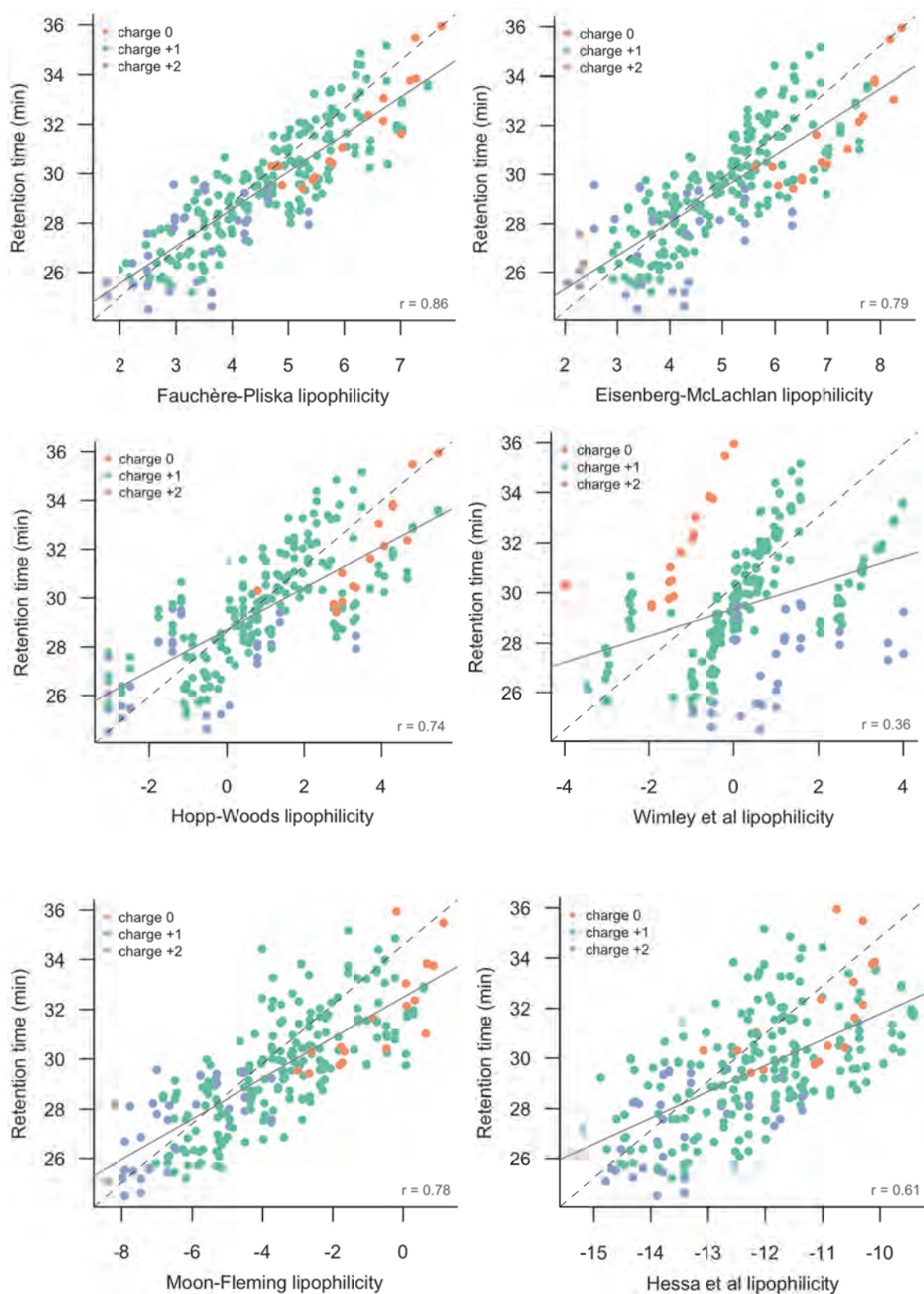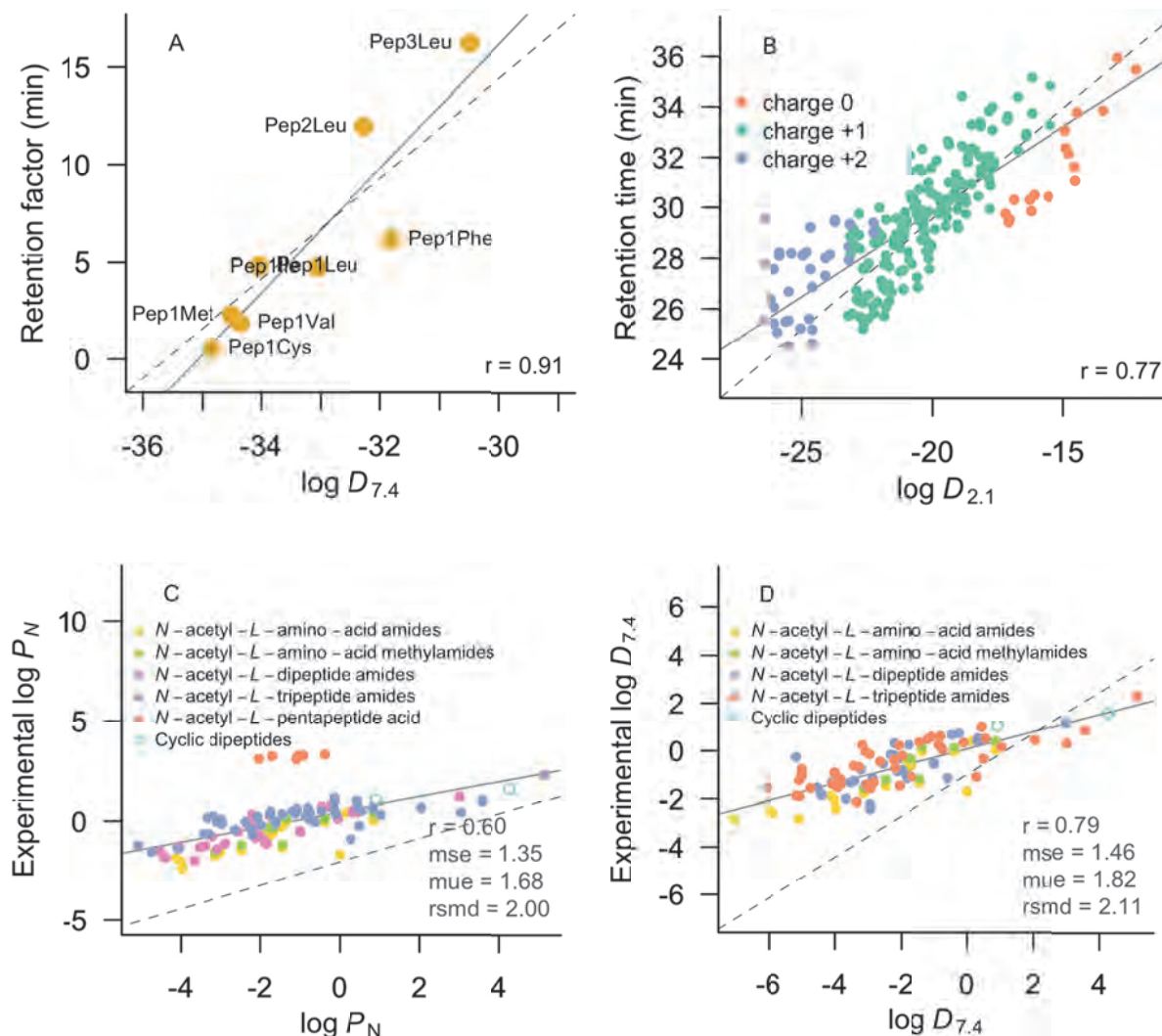
Figure 32. Representation of the cumulative lipophilicities determined from the ProtL scale versus (A) the retention time for eight 10-mer peptides (pH 7.4; Ref. 206), (B) 248 unique 13-mer peptides (pH 2.1; Ref. 207, 208), (C) log P_N for 118 random peptides (Ref. 209), and (D) log $D_{7.4}$ for 116 random peptides (Ref. 209).

Finally, our SolvL scale also can be applied in colloid chemistry. Here, it has been demonstrated that sea spray aerosols, represented commonly by sodium chloride (NaCl), can contain important quantities of organic compounds. Interestingly, NaCl-amino acid aerosols mixtures have shown a differentiated hydration in relation to the amino acid present. In the Dar`s study[210], a hydrophobicity scale served qualitatively as a gauge of hygroscopicity for 7 aerosols studied. For the sake of comparison, our

SolvL scale was tested obtaining a quantitative relation with the hygroscopicity of the aerosols (see Figure 33).
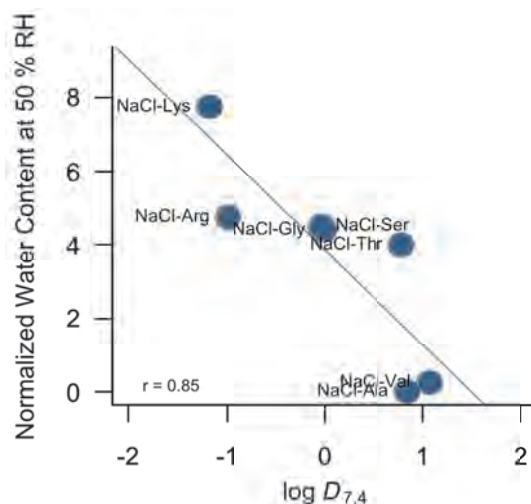
Figure 33. Representation of the normalized water content at 50 % of relative humidity (RH) for 7 NaCl-amino acid aerosols mixtures as a function of the log $D_{7.4}$ using the SolvL scale relative to Gly.

3.11 Application of the Lipophilicity Scale to Local Context-Dependent Lipophilicity of Peptides.

On the other hand, our ProtL scale was applied in the determination of the local context-dependent lipophilicity, which must be able to differentiate peptides or proteins with the same sequence but different topology. In this respect, although literature has shown the usefulness of hydrophobicity scales[211], these approaches have been criticized[150,212] because they just have been used as simple summations of individual hydrophobic descriptors, making them unable to find differences when the identity and quantity of amino acids is the same even when their order along the peptide sequence is variable.

We developed other additive scheme where the cumulative lipophilicity was determined taking into account the fraction of solvent-exposed area of the peptide residues, supplemented with two correction parameters that account for the

contribution due to the involvement of the backbone in hydrogen bonds,[213] and to the burial of apolar residues from water to hydrophobic environments[167] (See Methods). Thus, using this model our scale would have a potential application in differentiating either sequences of the same composition with different order, which should lead to a different molecular arrangement, as well as identical sequences with conformational differences (e.g structures obtained from a molecular dynamics simulations, NMR experiments or different crystallization conditions).

At first exploration, the tridimensional structure of the transmembrane segment of the influenza M2 protein that includes residues 25 to 46 was considered. This segment has been obtained using different experimental methods: an ensemble of structures derived from ssNMR experiments refined with MD simulations in water[214] (hydrophilic environment), and other from a detergent-solubilized state with octyl-*D*-glucoside crystalized with the vapour diffusion hanging drop method[215] (hydrophobic environment).

As expected, Figure 34 (left) illustrates that the X-ray average structure is more hydrophobic than the average ssNMR one, highlighting the effect of the environment on the final three-dimensional arrangement. To reveal the reason of the hydrophobic differences, we analysed the individual residue contribution to the global hydrophobicity in each average structure and we found that the exposure of ionizable residues (Asp44 and Arg45) determines the hydrophobic variance between those structures. As is shown in Figure 34, the X-Ray structure presents a cation-π interaction between Trp41/Arg45, which is more favoured in protein environments than in aqueous solutions,[216,217] increasing the lipophilic profile of this structure. On the other hand, the ssNMR structure present Asp44 and Lys45 directed to the solvent, enhancing its hydrophilicity, especially at physiological pH where both residues are present as charged species.
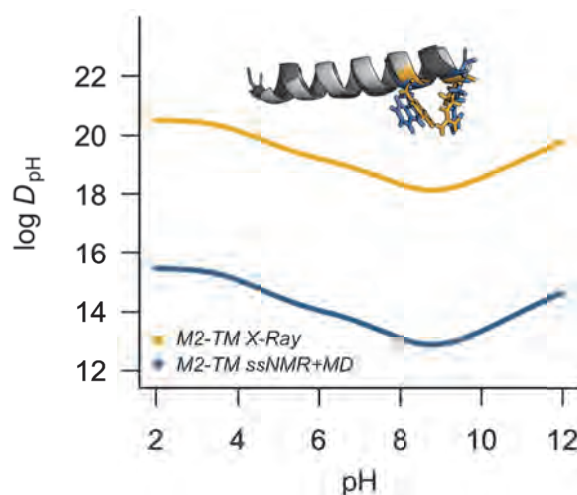
Figure 34. Lipophilicity profile (ProtL scale) for two average structures from the transmembrane segment of the influenza M2 protein (residues 25-46) derived of ssNMR experiments refined with MD simulations in water (PDB code 2l0j, blue line) and X-ray using the vapor diffusion hanging drop method (PDB code 4rwc, yellow line)

We also tested a protein crystalized with the same technique, but under different solvent conditions, which may influence the protein structure.[218,219,220] Figure 35 compares the lipophilicity profile for two crystal structures of the B1 immunoglobulin-binding domain of *Streptococcal* Protein G[221], the orthorhombic from (1pga) crystalized using 20 % isopropanol (IPA) and 50% of 2-methyl-2,4-pentanediol (MPD) and the trigonal (1pgb) using 70 % of MPD, both under acidic conditions (pH 4.0-4.5). In acidic conditions the orthorhombic crystal generated a more hydrophobic profile than the trigonal one, which is in agreement with the lower global polarity of the solvent used in the crystallization of 1pga and the greater number of crystal contacts in this type of crystal (expected lower hydration[221]). We demonstrate that the main difference observed in both hydrophobic profiles is due to the exposure of Lys residues in the crystals. Figure 35 (bottom) revels that Lys residues (marked in blue in the sequence) are less exposed to the solvent and for that reason they are less hydrophilic.
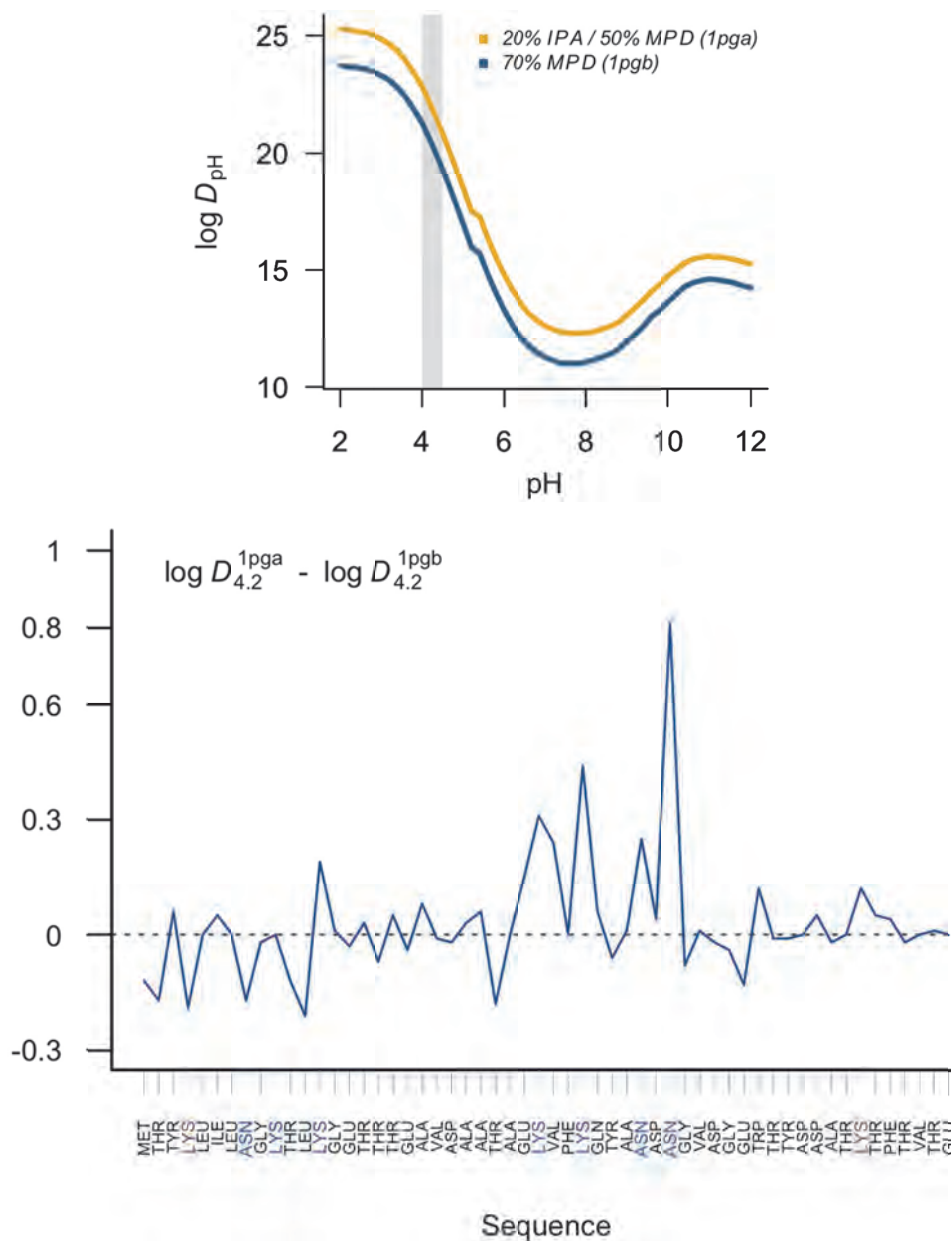
Figure 35. Lipophilicity profiles (ProtL scale) for two structures of the B1 immunoglobulin-binding domain of *Streptococcal* Protein G (PDB code 1pga, yellow line; 1pgb, blue line) crystalized under different solvent conditions (top). Differences for reside-level lipophilicity under cristalization conditions (pH 4.2) are also depicted (bottom).

Although more extensive analysis are required, these preliminary results suggest that, unlike previous applications of hydrophobic scales, our methodology can be used to explore the effect of subtle differences due to the usage of different experimental conditions.

Finally, one might expect an improved performance of the ProtL scale in the analysis of the lipophilic complementarity in peptide-protein and protein-protein complexes with regard to the SolvL approach. To this end, we have examined the relationship between the ProtL cumulative lipophilicity and the experimental binding free energies of 19 peptides to MHC (HLA-A*02:01 allele) proteins (see Table 16). These peptides were chosen subject to the availability of (i) a precise structural information of the peptide-protein complex in the Protein Data Bank,[222] and (ii) an estimate of the binding affinity in the Immune Epitope Database and Analysis Resource[223] (see Table 16).

Table 16. Length (L), Net Charge (Q) and Cumulative Lipophilicity Determined Using Protl, SolvL and Experimental Lipophilicity Scales of Peptides and Experimental Binding Affinities (BA; kcal/mol) Toward MHC (HLA-A*02:01 allele) from the Immune Epitope Database and Analysis Resource (Ref. 223).

| PDB | Sequence | L | Q | BA [a] | Lipophilicity Scale | | | | | | | | | | |
					SolvL	ProtL	Fauchère-Pliska	Eisenberg-McLachlan	Hopp-Woods	Wimley et al.	Moon-Fleming	Hessa et al.	Koehler et al.	Janin et al.	Kyle-Doolittle
2BST	SRYWAIRTR	9	3	-6.4	-9.29	-22.50	2.51	0.75	-0.67	7.75	-8.55	-6.99	-0.72	-3.60	-7.95
3BO8	EADPTGHSY	9	-1	-7.1±0.3	-3.12	-17.72	0.93	1.97	-1.91	-4.12	-8.81	-9.75	-1.35	-1.80	-9.63
1QVO	QVPLRPMTYK	10	2	-7.1±0.3	-4.58	-15.13	4.59	5.43	0.79	5.18	-4.68	-9.02	-1.47	-3.60	-5.25
2X4S	AMDSNTLEL	9	-2	-5.8	-1.85	-15.83	3.15	3.54	-0.53	-2.59	-5.33	-6.21	-0.72	-0.40	-0.50
2X4U	ILKEPVHGV	9	0	-6.7±0.6	-3.98	-13.24	5.16	5.34	0.78	1.92	-5.15	-6.31	-0.80	-0.20	2.99
2GT9	EAAGIGILTV	10	-1	-6.6	-1.34	-11.34	6.76	6.07	3.84	1.24	-0.85	-2.07	1.00	2.80	11.38
1T22	SLYNTVATL	9	0	-7.2±0.1	1.92	-10.44	5.77	5.86	5.96	2.89	-2.46	-2.42	0.45	0.50	4.82
5W1W	VMAPRTLVL	9	1	-6.8	0.87	-10.44	7.35	6.93	4.21	5.40	1.43	-2.64	0.32	1.00	9.42
2X4O	KLTPLCVTL	9	1	-6.5	4.20	-9.56	8.11	6.76	4.14	5.24	-1.25	-2.85	-0.17	0.50	8.16
5EU3	YLEPGPVTA	9	-1	-7.0	-0.22	-9.19	5.25	5.73	2.54	-0.12	1.15	-6.09	-0.46	-0.20	0.51
3MRM	KLVALGINAV	10	1	-7.3	-1.04	-8.66	6.67	6.47	4.49	4.96	-2.80	-2.48	0.45	1.50	11.89
3UTQ	ALWGPDPAAA	10	-1	-7.9	2.42	-8.44	5.86	6.13	3.04	0.28	0.41	-6.49	-0.38	1.10	2.17
3GSN	NLVPMVATV	9	0	-6.6±0.1	1.69	-7.45	7.28	7.72	6.03	2.93	0.86	-2.43	0.32	2.00	10.45
3QEQ	AAGIGILTV	9	0	-7.01	1.60	-6.59	7.40	6.62	6.03	3.72	0.35	-0.11	1.51	3.50	13.93
3MRG	CINGVCWTV	9	0	-5.9±0.1	2.96	-5.68	9.23	5.97	7.57	3.81	-3.25	-1.56	0.44	3.60	9.05
2PYE	SLLMWITQC	9	0	-6.4	8.19	-2.24	10.22	8.33	8.01	5.05	-0.59	-1.53	0.51	2.30	7.73
1HHK	LLFGYPVYV	9	0	-8.4	7.24	-0.04	10.27	9.86	9.98	5.76	6.84	-1.68	1.11	1.90	10.36
2VLL	GILGFVFTL	9	0	-8.6±0.1	6.76	1.21	10.26	9.00	8.95	6.46	3.75	0.47	1.89	3.70	14.88
3OX8	FLPSDFFPSV	10	-1	-8.7±0.1	4.75	1.30	8.88	8.42	5.23	3.13	4.11	-5.71	0.09	1.20	5.91

[a] Estimated generally using cellular MHC/competitive/fluorescence half maximal inhibitory concentration (IC$_{50}$), and exceptionally from radiactive assays. When several data were available, the binding affinity is given as the mean value together with the standard deviation.

The results show that the ProtL scale works better than the SolvL scale (correlation coefficients of 0.58 and 0.42, respectively; Figure 36) when the whole set of 19 peptides is considered, yielding correlation coefficients that compare with Moon-Fleming and Eisenberg-McLachlan scales (r = 0.61 and 0.51, respectively; see Table 17). This correlation is remarkable keeping in mind the heterogeneity of the peptides, and the uncertainty arising from the combination of data taken from different studies and determined using distinct experimental approaches. Further, a significant improvement is observed upon exclusion of the two Cys-containing peptides (PDB codes 3MRG, and 2PYE), perhaps reflecting a quenching effect of cysteine in fluorescence assays.[224,225] Thus, upon exclusion the correlation coefficient of ProtL and SolvL scales increases up to 0.80 and 0.73, respectively, outperforming the results obtained with the experimental scales (r < 0.67; see Table 17).
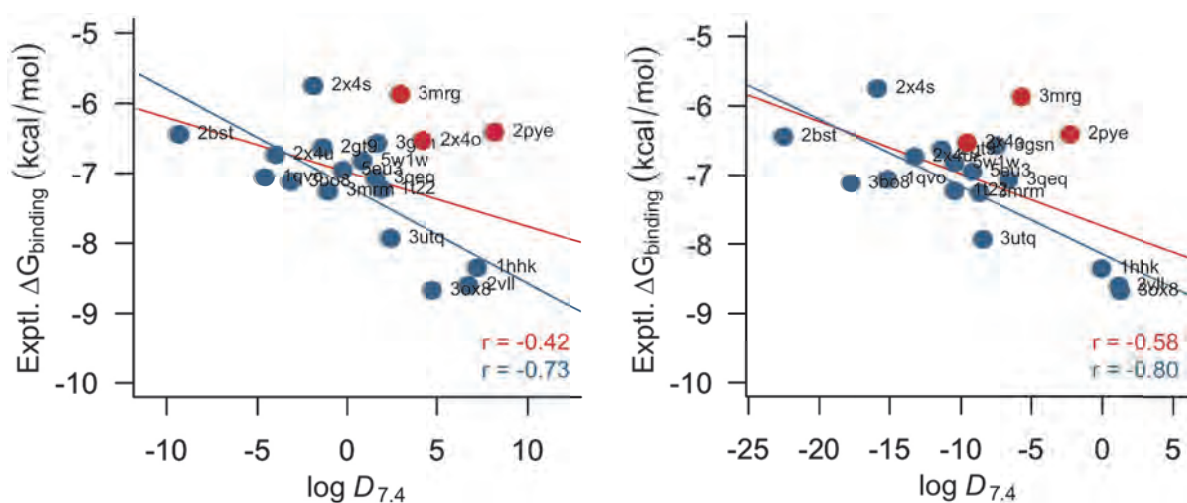
Figure 36. Relationship between the cumulative lipophilicities determined from (left) SolvL and (right) ProtL scales versus experimental binding affinities of MHC-bound peptides. Cys-containing peptides are indicated as red dots.

Table 17. Correlation Coefficient of Cumulative Lipophilicity Determined Using Different Lipophilicity Scales of MHC (HLA-A*02:01 allele)-Bound Peptides with Experimental Estimates of Binding Affinities.

Scale	Entire Set (n=19)	No Cys Set (n=16)
Fauchère-Pliska	0.34	0.67
Eisenberg-McLachlan	0.51	0.66
Hopp-Woods	0.36	0.62
Wimley et al.	0.18	0.31
Moon-Fleming	0.61	0.65
Hessa et al.	0.07	0.25
Koehler et al.	0.32	0.41
Janin et al.	0.18	0.39
Kyte-Doolittle	0.21	0.34
ProtL	0.58	0.80
SolvL	0.42	0.73

3.12 Relationship between Toxicity and Lipophilicity of Aβ$_{42}$ Peptides Involved in the Alzheimer`s Disease.

Formation of molecular agragates in the brain, is one of the main hallmark in the Alzheimer`s disease. These aggregates are form of repeated units of amyloid peptides. The amyloid peptides come from the transmembrane amyloid precursor protein (APP *e.g* presenilin 1 and presenilin 2) upon proteolysis by the combination of β- and γ-secretases[226,227] being the Aβ$_{42}$ peptides the more toxic ones.[228,229] Hydrophobicity has been recognized to take part actively in the Alzheirmer`s disease. For instance, mutants that compromise the stability of APP (presenilin 1) in the membrane, via a reduction of the global hydrophobicity of the protein, are correlated with a increase of toxic species (Aβ$_{42}$ peptides relative to Aβ$_{40}$ peptides) that could cause the disease.[230] Also, toxic species from amyloid peptides (*i.e* dimers and aggregates) have been reported to be highly dependent of its hydrophicity,

especially of the hydrophobic residues.[231,232] Lately, attention have been put in the wild type and mutants of monomeric species of Aβ42 peptides where computed properties as secondary structure, solvent accessible surface areas and radius of gyration have improved the understanding of experimental toxicity reported in these biomolecules. Aβ42 monomers with reduced helix propensity of the ensembles have produced the more toxic mutants.[233]

Thus, the understanding of changes of the conformational ensemble of Aβ42 monomers upon residue mutations could provide crucial information about the propensity to aggregation and also to the toxicity of specific proteins. These conformational ensembles of the monomers will determine ultimately the main features of the monomer like global and local lipophilicity.

The available experimental toxicity data[233] for wild type and mutated Aβ42 is presented in Table 18. The toxicity of the Aβ peptides in Alzheimer's disease is attributed to the oligomerization propensity. Somehow such propensity is encoded into the properties of the corresponding monomers, the basic building blocks of oligomers.

Table 18. Net Charge (Q), Toxicity (relative EC_{50} to WT peptide), and Average Cumulative Lipophilicity $\left(\left\langle \log D_{7.4} \right\rangle\right)$ in Aβ42 Peptides.

Peptide	Q	Toxicity (EC_{50})	$\left\langle \log D_{7.4} \right\rangle$
E22Q	-2	0.07	-6.9
E22G	-2	0.14	-8.6
E22K	-1	0.14	-11.2
D23N	-2	0.38	-5.1
D7N	-2	0.70	-6.1
A2V	-3	0.80	-4.2
H6R	-2	0.80	-7.2
D7H	-2	0.80	-4.9
WT	-3	1.0	-5.6
K16N	-4	1.0	-6.2
A21G	-3	1.7	-7.0

There is a general consensus that hydrophobic interactions take part actively to the oligomer formation and mutations that increase the lipophilicity promote aggregation. Our ProtL scale, however, despite of having some degree of correlation, did not show significant tendency between global lipophilicity and toxicity. Recently, it has been demonstrated that hydrophobic fragments in the $A\beta_{42}$ sequence are responsible to promote aggregation instead of the specific side chains of those residues or well the entire peptide.[231] Thus, segment rather than global features of the monomer $A\beta_{42}$ mutants could shed some light in the understanding of the toxicity and give a better explanation of this phenomenon. Here, we studied the classical segments reported in the literature for $A\beta_{42}$ peptides (Figure 37): N-terminal (NT; residues 1-16), Central hydrophobic core (CHC, residues 17-21), Loop (Loop, residues 22-30), C-terminal (CT, residues 31-42) and Body (residues 17-42).

The classical segments (see Figure 37) gave a significant correlation (r = 0.76) with the NT fragment, and some correlation (r = 0.52) with the CHC fragment and Loop (r = -0.44).
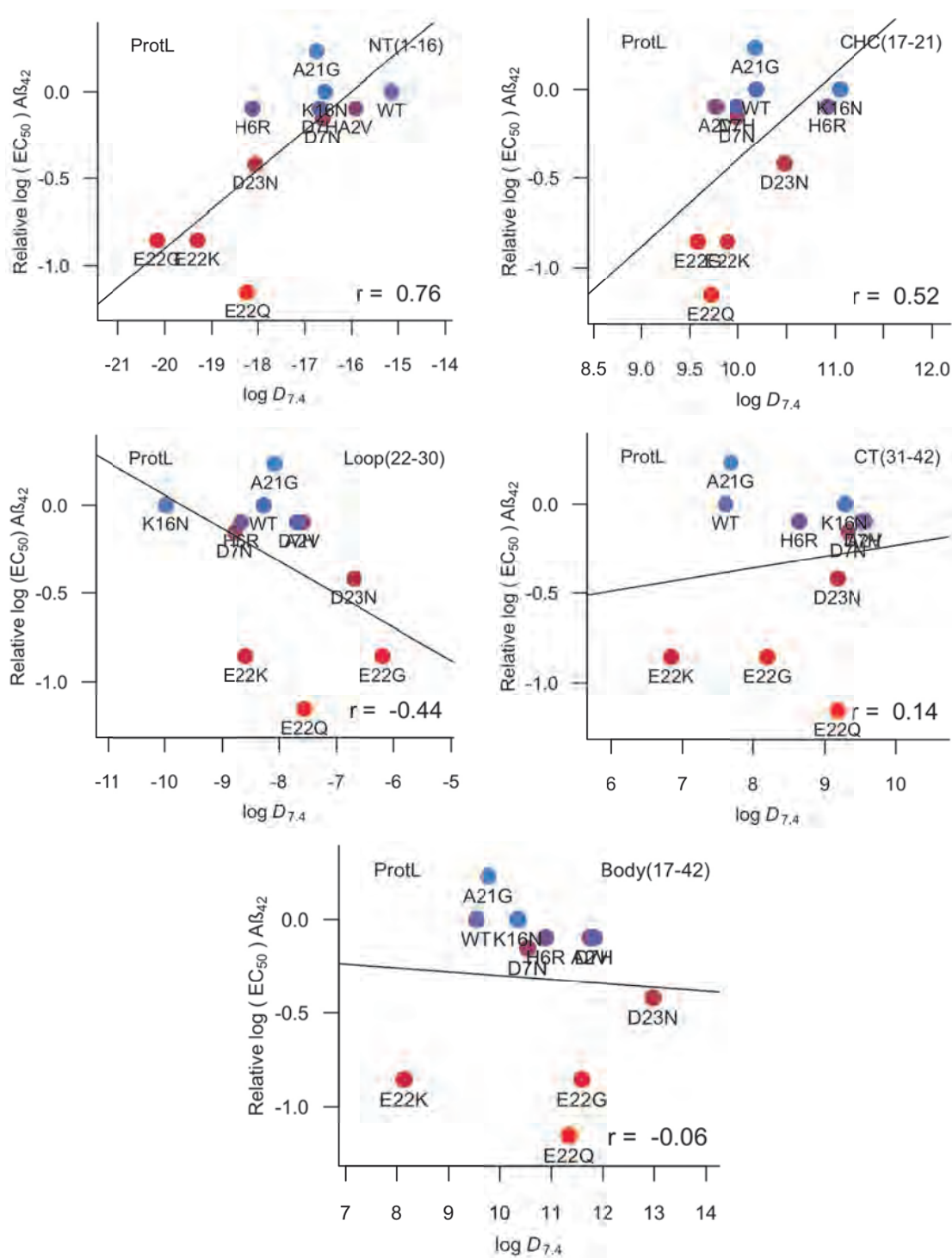
Figure 37. Correlation between experimental toxicity data and lipophilicity $\left(\left\langle \log D_{7.4} \right\rangle\right)$ of classical fragments in Aβ$_{42}$ peptides.

A systematic analysis for searching the non-classical fragments that better correlate with the experimental toxicity (see Table 19) showed, in agreement with the classical segments, two different trends. In one hand, the best positive correlation (r=0.91) for a non-classical segment is composed for a merge segment that includes residues of the NT and CHC stretches. On the other hand, the best negative correlation (r=-0.81) with another mixed segment form the Loop and CHC fragments. The first fragment suggest that toxic peptides tends to be more hydrophilic in this region and contrary, the second one suggests that the toxicity of the peptide increases with the hydrophobicity of this segment.

Table 19. Systematic search for the top 10 best positive and negative correlations (r) between non-classical fragment lipophilicity and experimental toxicity for 11 Aβ_{42} peptides. The length (L) and residues involved with its classification based on the classical segments are also given for each fragment.

Top 10 best positive non-classical fragments correlation				Top 10 best negative non-classical fragments correlation			
Fragment (residues)	Classical segments involved	L	r	Fragment (residues)	Classical segments involved	L	r
2-20	NT-CHC	19	0.91	21-24	CHC-Loop	4	-0.81
2-18	NT-CHC	17	0.90	21-23	CHC-Loop	3	-0.77
2-19	NT-CHC	18	0.90	21-25	CHC-Loop	5	-0.77
3-20	NT-CHC	18	0.90	20-22	CHC-Loop	3	-0.76
3-18	NT-CHC	16	0.89	21-26	CHC-Loop	6	-0.76
3-19	NT-CHC	17	0.88	22-23	Loop-Loop	2	-0.75
1-18	NT-CHC	18	0.88	22-24	Loop-Loop	3	-0.74
1-19	NT-CHC	19	0.87	21-22	CHC-Loop	2	-0.73
3-21	NT-CHC	19	0.87	20-24	CHC-Loop	5	-0.73
2-21	NT-CHC	20	0.87	20-25	CHC-Loop	6	-0.71
mean		*18*	*0.89*	*mean*		*4*	*-0.75*

These results pointed out two divergent tendencies between classical and non-classical (i.e functional fragments) that led us to the hypothesis that the amphipathicity, expressed as the difference in lipophilicity between the two segments, could explain better the toxicity of $A\beta_{42}$ peptides. In fact, it is well known that amphipathic peptides (e.g antimicrobial peptides) are more active but also more toxic when their amphipathicity increase.[234] However, despite of the $A\beta_{42}$ peptides having an amphipathic sequence (polar amino acids are found in the first 28 residues while the apolar ones in the last 14 residues)[235] there is no a reported relationship between toxicity and amphipathicity for these peptides. To this end, we have examined systematically all possible combinations between different segments with diverse lengths (l > 2) with an opposite correlation with toxicity in order to compute its amphipathicity and correlate it with the experimental toxicity (see Table 20 and Figure 38).

Table 20. Systematic search for the top 10 best correlations (r) between non-classical fragment amphipathicity and experimental toxicity for 11 $A\beta_{42}$ peptides. The residues involved with its classification based on the classical segments are also given for each fragment.

Residues	Classical segments involved	r
(35-39) vs (2-18)	CT vs NT-CHC	-0.97
(36-39) vs (2-18)	CT vs NT-CHC	-0.96
(36-39) vs (2-19)	CT vs NT-CHC	-0.96
(35-39) vs (2-19)	CT vs NT-CHC	-0.96
(34-39) vs (2-19)	CT vs NT-CHC	-0.96
(33-39) vs (2-19)	CT vs NT-CHC	-0.96
(33-41) vs (2-19)	CT vs NT-CHC	-0.96
(36-39) vs (2-20)	CT vs NT-CHC	-0.96
(21-22) vs (2-18)	CHC-Loop vs NT-CHC	-0.95
(21-23) vs (2-18)	CHC-Loop vs NT-CHC	-0.95
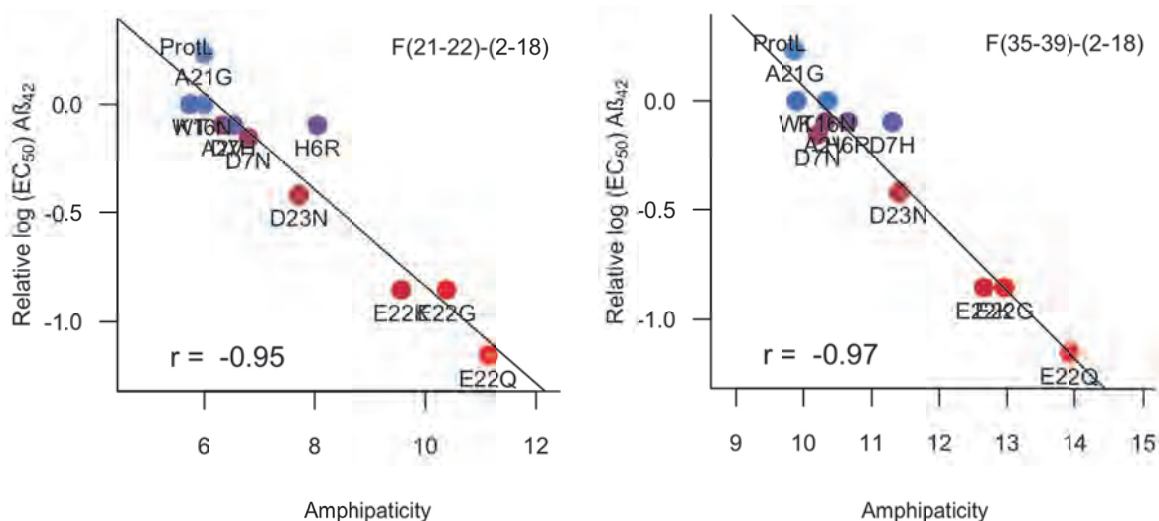
Figure 38. Correlation between experimental toxicity data and amphipathicity of functional hydrophilic fragment F(2-18) versus functional hydrophobic fragment F(21-22) (left) and F(35-39) (right).

Figure 38 shows that the correlation coefficients obtained are very significant ($r \geq |0.95|$) and presents another relevant hydrophobic fragment, F(35-39), which was not identified as relevant in terms of its correlation between lipophilicity and toxicity (see Table 19).

The results suggest the regulatory function of the NT and CHC segment on the aggregation propensity of the rest of the peptide. The aggregation of the peptide would be facilitated by a non-interfering NT and CHC, but the non-aggregation character would be achieved by a NT segment adopting helix conformations while interacting with the main body of the peptide thus interfering with its aggregation. Accordingly, the Loop and CT, especially dominated by the increase in the lipophilicity of residues 21-22 and 38 upon mutations, respectively could stabilize the formation of a hydrophobic body of the peptide that would favour the formation of aggregates.

Overall, the results are in line with the role of beta-hairpins in the aggregation[11] and the prominent role of the NT region in the toxicity profile.[12,13] Even more, overall, the present results point out to beta barrels as possible toxic oligomer structure, where the CHC-CT fragments would form the main body of the barrel and the NT segment would be exposed into solvent.[13]

Finally, our findings suggest that the amphipathicity of monomeric $A\beta_{42}$ peptides could drive its toxicity, similarly as has been widely reported in antimicrobial peptides and from there, the modulation of this physicochemical descriptor could be evaluated to control the toxic amyloid species.

4

Chapter

Methods

4. METHODS

4.1 Refinement of the MST Model for Solvation of Neutral and Ionic Compounds in *n*-Octanol.

For the purpose of this study, several sets of molecules were used to refine the MST model. First, a set of 19 small nitrogen-containing aromatic compounds was used to calibrate the parametrization of pyridine-type and pirrole-type nitrogen atoms. This data set includes 6-methyluracil, 9-methyladenine, adenine, albendazole, caffeine, clonidine, cytosine, diphenylamine, fluconazole, fluorouracil, guanine, imidazole, metronidazole, phenytoin, pyridine, pyridimidine, pyrrole, thymine, and uracil (see Figure 7).

Next, a set of 27 compounds was used to calibrate the suitability of the MST model to compute log P_I values, This set included 18 cations (2-(2-pyridyl)-ethylammonium, 2-phenylethylammonium, 3-carboxyanilinium, 4-carboxyanilinium, 4-methyl-N-ethylbenzylammonium, 4-methyl-N-heptylbenzylammonium, 4-methyl-N-pentylbenzylammonium, 4-phenylbutylamine, AceHisNH$_2$, AceLysNH$_2$, amitriptyline, desipramine, imipramine, lidocaine, tetrabutylammonium, tetraethylammonium, tetramethylammonium, tetrapentylammonium) and 9 anions (2-4-dichlorophenoxyacetate, 5-phenylvalerate, AceAspNH$_2$, AceGluNH$_2$, diclofenac, ibuprofen, indomethacin, naproxen, pentachlorophenolate; see Figure 10).

Finally, a set of 35 compounds was used to calibrate the behavior of the refined MST model for predicting the partition coefficient of the neutral compound (P_N), and the distribution coefficient at pH 7.4 ($D_{7.4}$), taking advantage of the available experimental data for pK_a, log P_N and log $D_{7.4}$. This data set includes 31 drugs (albendazole, amitriptyline, antipyrine, bumetanide, caffeine, clormipramine, clonidine, clozapine, cyclobenzaprine, desipramine, diazepam, diclofenac, diphenylamine, estradiol, fluconazole, flurbiprofen, ibuprofen, imipramine, indomethacin, lidocaine, loratadine, maleic acid, metoclopromaide, metronidazole,

 mezoridazine, naproxen, paracetamol, pentachlorophenol, pentoxifylline, phenytoin, and triflupromazine; see Figure 13) and 4 amino acids (aspartic acid, glutamic acid, histidine and lysine, capped with acetyl and amide groups at the N- and C-terminus).

On the other hand, the refined model was checked by computing the pH-dependent distribution profiles in the framework of the different partitioning schemes discussed above. To this end, several drugs (ibuprofen, imipramine, desipramine, pentachlorophenol, lidocaine, amitryptiline, and warfarin), and four amino acid analogues (N-acetyl-L-aspartic acid amide, N-acetyl-L-lysine amide, N-acetyl-L-glutamic acid amide, and N-acetyl-L-histidine amide) were considered taking advantage of the available experimental data for these compounds.

The molecular geometries of the compounds were fully optimized at the B3LYP/6-31G(d) level of theory in the corresponding solvent phase, water or n-octanol, using the IEFPCM version of MST model. Then, single point calculations in gas phase and in solution were performed in order to estimate the free energy of solvation in water and n-octanol. All calculations were performed in Gaussian 09.[239]

For the set of ionizable compounds, computations were performed for the minimum structure obtained after geometry optimization of an extended conformation of the molecule. This strategy was motivated by the generally low number of rotatable bonds present in these compounds (see Figure 14), as well as by the similar n-octanol/water transfer free energies obtained from a single-conformation approach and from conformational ensembles for drug-like compounds in a previous study.[57] Nevertheless, for the subset of N-acetyl-L-amino acid amides, calculations were performed taking into account all possible rotamers with a probability contribution higher than 5% to the total conformational space as given by the backbone-dependent conformational library reported by Dunbrack and Karplus.[58,59] Calculation of the log D was accomplished using eqs 5, 6, and 30. Data for the

partition of inorganic ions required for eq 30 were taken from the literature (

$\log P^{\circ}_{I,Cl^-} = -4.5$, $\log P^{\circ}_{I,Na^+} = -2.5$ and $\log P^{\circ}_{I,K^+} = -2.6$).[191,240,241]

4.2 Experimental Determination of Partition of Neutral and Ionic Species for Indomethacin and Chlorpromazine.

Partition of neutral and ionic species for indomethacin and chlorpromazine were performed using the potentiometric method. For this task was designed a multiset of the potentiometric pK_a determinations using different ratios of volumes of water and n-octanol how is shown in the Tables 21 and 22.

Table 21. Experimental Design for the Potentiometric Determination of log P_N and log P_I for Indomethacin.

id	Mass (mg)	Initial KCl in aqueous solution (mol/L)	Volume water (V_w, mL)	Volume n-octanol (V_o, mL)	Ratio (V_w/ V_o)	File
1	3.15		15	3	5	0502
2	3.95		10	1	10	0503
3	3.16	-	19	1	19	0702
4	3.44		20	0.5	40	1002
5	2.74		21	0.3	70	1003
1	4.12		15	1	5	0803
2	3.32		10	2	10	0804
3	3.93	1×10^{-5}	19	3	19	1202
4	3.26		20	4	40	1203
5	3.45		21	5	70	1204
1	3.10		19	1	19	2904
2	3.14		16	3	5	2905
3	3.01	1×10^{-3}	20	0.5	40	3002
4	3.33		10	1	10	3003
5	3.04		17	0.2	85	3102
1	3.13		15	3	5	3103
2	2.93		10	1	10	3104
3	3.59	0.15	19	1	19	0603
4	3.24		20	0.5	40	0604
5	3.15		21	0.3	70	0605

Table 22. Experimental Design for the Potentiometric Determination of log P_N and log P_I for Chlorpromazine.

id	Mass (mg)	Initial KCl in aqueous solution (mol/L)	Volume water (V_w, mL)	Volume n-octanol (V_o, mL)	Ratio (V_w/ V_o)	File
1	3.81		15	1	5	1103
2	3.98		10	2	10	1104
3	3.93	-	19	3	19	1105
4	3.07		20	4	40	1106
5	3.00		21	5	70	1107
1	3.13		15	1	5	1204
2	3.18		10	2	10	1205
3	3.59	1×10^{-5}	19	3	19	1206
4	3.32		20	4	40	1207
5	3.30		21	5	70	1702
1	3.56		19	1	19	1703
2	3.86		16	3	5	1704
3	3.64	1×10^{-3}	20	0.5	40	1705
4	3.50		10	1	10	1802
5	3.33		17	0.2	85	1803
1	3.20		15	3	5	1804
2	3.27		10	1	10	1805
3	3.74	0.15	19	1	19	1806
4	3.40		20	0.5	40	1807
5	3.10		21	0.3	70	-

4.3 Development of the Lipophilicity Scale for Amino Acid Residues.

Following a previous study on the hydration free energy of the natural amino acids,[155] the N-acetyl-L-amino acid amides (CH_3-CO-NH-CHR-$CONH_2$) were chosen as molecular models. Using the backbone-dependent conformational library reported by Dunbrack and coworkers,[190,242,243] a total of 572 rotamers (i.e., conformers with a probability contribution higher than 5% to the total conformational space of each residue) were compiled. These structures were then used to compute the n-octanol/water transfer free energies, which were performed with the B3LYP/6-31G(d) MST[136] version of the IEF-PCM[143,244] model. Computation of the distribution coefficients at a given pH (log D) was performed by combining the partition coefficient of neutral and ionic species (for ionizable residues) using eq 6.

The contribution of the conformational species in water and *n*-octanol was accounted for considering two weighting schemes, giving rise to the Solvent-like (SolvL) and Protein-like (ProtL) lipophilicities scales, respectively.

(i) In the SolvL scale, the contribution of each conformational state to the partition coefficient of the neutral/ionized species was determined using a Boltzmann weighting scheme, where the effective free energy was estimated by combining the internal energy of the conformer and its solvation free energy in water and *n*-octanol. To this end, the geometry of all rotamers was optimized at the B3LYP/6-31G(d) level of theory while keeping the backbone dihedrals fixed to the torsional values of the Dunbrack's library, and subsequently single-point calculations in the gas phase and in solution. The log D was then computed using eq 6, adopting the pK_a values reported for ionizable residues from experimental peptide models in aqueous solutions.[192,193]

(ii) In the ProtL scale, the contribution of each conformation to the partition between the two solvents was determined by using the weights reported in the Dunbrack`s library, which reflect the rotameric distribution in a protein environment. The pK_as of ionizable residues were taken from values in folded proteins.[194,195]

For the sake of comparison, we also computed both approaches with the SMD model using the B3LYP/6-31G(d) level of theory.[137]All calculations were performed using a locally modified version of Gaussian 09.[239]

4.4 Comparison of the Lipophilicity Scale for Amino Acid Residues with Experimental Scales.

Due to the diversity of experimental lipophilicity scales of amino acids, generally expressed in terms of transfer free energies, comparison was made by converting them to partition/distribution coefficients, which were subsequently normalized to Gly following eq 37.

$$\log P_N / D_{pH} = \frac{(-\Delta\Delta G_{transf,AA} - \Delta\Delta G_{transf,Gly})}{RT \ln 10} \tag{37}$$

where $\Delta\Delta G_{transf,AA}$ is the transfer free energy of a given amino acid from the aqueous phase to the organic/biological environment, and $\Delta\Delta G_{transf,Gly}$ is the free energy of transfer of Gly.

4.5 Determination of the Cumulative Lipophilicity in Peptides.

Most of the experimental scales present in the literature compute the lipophilicity of a given peptide as the sum of individual lipophilicity of the constituent amino acids relative to a reference residue, usually Gly or Ala. Since the MST solvation model gives atomic contributions to the transfer free energy,[245-247] we can separate the global lipophilicity in contributions corresponding to the backbone (*bb*), side-chain (*sc*), and the capping groups (*cg*). Combination of the *bb* and *sc* contributions yields the amino acid lipophilicity (reported in Table 1 in the manuscript), whereas the contribution of the capping groups has been estimated to be (N-terminus) CH_3CO- ($\log P_N = 0.20$), NH_3^+- ($\log D_{7.4} = -2.99$), and (C-terminus) NH_2- ($\log P_N = -1.08$), $NMe-$ ($\log P_N = 0.35$), $COO^-–$ ($\log D_{7.4} = -4.89$).

The cumulative lipophilicity of a peptide with N_{res} residues may be estimated by using eq 38.

$$\log(P_N / D_{pH})^{peptide} = \sum_{i=1}^{Nres} \log(P_N^i / D_{pH}^i)^{bb+sc} + \sum_{i=1}^{Ncg} \log(P_N^i / D_{pH}^i)^{cg} \tag{38}$$

where P_N^i / D_{pH}^i stands for the fragment (*bb+sc* or *cg*) partition/distribution coefficient, N_{res} and N_{cg} being the total number of residues and capping groups in the peptide.

For practical applications, this simple expression is convenient when there is no explicit knowledge about the 3D structure of peptides, as may occur in structureless peptides. For our purposes here, this is the expression adopted to evaluate the lipophilicity of small, flexible peptides in solution.

On the other hand, if the 3D structure of the peptide is known from experimental (X-ray, NMR) or computational (Molecular Dynamics) approaches, then the cumulative lipohilicity may be estimated taking into account the specific structural features of peptides/proteins, as noted in eq 39.

$$\log(P_N / D_{pH})^{peptide} = \sum_{i=1}^{Nres} \left(\lambda^i \cdot \log(P_N^i / D_{pH}^i)^{bb+sc} + \lambda^i \cdot \log(P_N^i / D_{pH}^i)^{cg} + \alpha^i + \beta^i \right) \quad (39)$$

In eq 39, λ^i stands for the fraction of solvent-exposed surface area (SASA) of the amino acid (*bb+sc*) or capping group (*cg*) according to the local structural environment of in a peptide/protein. For our purposes, the SASA was determined using NACCESS.[202]

In addition, two correction factors were also introduced. The parameter α^i introduces a correction to the hydrophobic contribution when the backbone participates in a hydrogen bond (HB). This contribution can be estimated to amount, on average, to 0.73 (log *P* units) per HB.[248] The occurrence of this kind of HBs in a given 3D structural model was determined with the DSSP program.[213] Finally, the β^i factor accounts for a correction due to the burial of the side chain of hydrophobic residues (Ala, Leu, Ile, Val, Pro, Phe, Trp, Met and Tyr) from water to a lipophilic environment. This contribution has been estimated to be 0.023 kcal mol^{-1} Å^{-2} according to the studies reported by Moon-Fleming for the transfer of nonpolar side

chains from water into a lipid bilayer.[167] Therefore, the β^i term has been estimated from the fraction of the buried side chain with respect to the fully buried side chain, as noted in eq 40.

$$\beta^i = H^i_{res} \cdot (1 - \lambda^i)^{sc} \tag{40}$$

where H^i_{res} stands for the hydrophobic contribution (in log P units) of a specific apolar residue, which was estimated as noted in eq 41

$$H^i_{res} = \frac{0.023 \cdot \text{SASA}^{sc}_{res}}{2.303RT} \tag{41}$$

where SASA^{sc}_{res} is the average SASA of a given residue type, R is the gas constant, and T is the temperature in the Kelvin scale.

The H^i_{res} values for nonpolar residues are given in Table 23.

Table 23. Average Solvent Accessible Surface Area (SASA) for the Side Chain of the Hydrophobic Residues and the Hydrophobic Effect Contribution Value when the Side Chain is Fully Buried.

Residue	Average SASA (Å²)	H^i_{res} (log P units)
Ala	69	1.2
Val	130	2.2
Leu	158	2.7
Ile	157	2.6
Met	166	2.8
Pro	115	1.9
Phe	188	3.2
Trp	232	3.9
Tyr	201	3.4

4.6 Analysis of Aβ₄₂ Peptides Involved in the Alzheimer`s Disease.

Understanding the effect of mutations on the segment hydrophobicity and amphipacity of a series of Aβ₄₂ mutant monomers, whose toxicity are known, could shed some light on the geometrical and property factors that confers toxicity to the peptides. To this end, the global and segment lipophilicity of the Aβ₄₂ mutant monomers (E22Q, E22G, E22K, D23N, D7N, A2V, D7H, H6R, WT, K16N, A21G) was computed from the corresponding conformational ensembles (8000 conformers for each mutant) generated from multi-seeded molecular dynamics simulations[233] and confronted with the corresponding experimental toxicity in search of the highest and significant correlation coefficients. Global and fragment lipophilicity was measured as an averague of $\log D_{7.4}$ according eq 39. The functional segments employed here (NT 1-16, CHC 17-21, Loop 22-30, CT 31-42, body of the peptide 17-42) are those already identified in literature according to the secondary structure. Averague amphipathicity, $\left\langle Amp_{7.4}^{peptide} \right\rangle$, for the Aβ₄₂ mutant monomers was computed from the difference between the averague lipophilicity of a hydrophobic fragment, $\left\langle \log D_{7.4}^{F,hpho} \right\rangle$, versus the averague lipophilicity of a hydrophilic fragment, $\left\langle \log D_{7.4}^{F,hphi} \right\rangle$ according to eq 42.

$$\left\langle Amp_{7.4}^{peptide} \right\rangle = \left\langle \log D_{7.4}^{F,hpho} \right\rangle - \left\langle \log D_{7.4}^{F,hphi} \right\rangle \tag{42}$$

where *Amp* refers to amphipathicity descriptor.

5

Chapter

Conclusions

5. CONCLUSIONS

a. The refinements made in the IEFPCM/MST method for the treatment of solutes in *n*-octanol improve the ability of the model for predicting both partitioning of neutral and ionic compounds. Although the balance between accuracy and computational cost is more favourable for empirical methods, the availability of refined versions of QM-based continuum solvation methods opens the way to the analysis of factors implicated in the partition of (bio)organic molecules in complex chemical systems.

b. For the set of 35 compounds used in the determination of log $D_{7.4}$, there are little differences between the calculated values determined at physiological pH with the three formalisms examined here. However, eq 6 is the minimal scheme required to rationalize the pH-dependent distribution profiles of ionizable compounds. The role of the Galvani potential difference (eq 30) between the two phases, however, may be relevant at higher concentrations of the background salt.

c. The refined lipophilicity models (Schemes 6 and 7), in conjunction with the general model of lipophilicity developed (Scheme 8) in this thesis, can be useful to explain the effect of the background salt used in experiments. However, these must be tested experimentally (work underway).

d. Accounting for pH conditions and structural preferences are necessary for improving the quantitative description of the lipophilicity of amino acids. The computation of log D_{pH} using the Boltzmann`s weighting scheme (*Solvent-like*) led to a close agreement not only with Fauchère-Pliska scale, but also to high correlation with bulk-solvent scales. Furthermore, it also lead to significant correlations with the retention time of a wide variety of short peptides.

e. ProtL scale works in an efficient way to describe the local-context dependet lipophilicity, which was supported through of the differentiation of lipophilic profiles between the same peptides whose structure was reported using different structural methods in variate chemical environments. Also, this approach was able to describe the lipophilic complementarity in protein-protein complexes represented by MHC complex.

f. The amphipathicity, derived from the lipohilic descriptors developed in this work, between the NT and CHC region and the Loop of the peptide in $A\beta_{42}$ peptides emerges as a key factor for the aggregation propensity and its associated toxicity. Mutations that reinforce the hydrophilic character of the NT and CHC regions and the hydrophobic character of the main body of the peptide will cause an increment of aggregation propensity and toxicity.

g. Present results pave the way to explore the application of this methodology to the calculation of hydrophobic parameters for other species of native residues (proline *cis*) or non-proteogenic amino acids, as well as to other fragments relevant to proteins. On the other hand, the applicability possibilities of the present versatile scale are vast, such as the development of scoring functions for peptide-protein or protein-protein docking protocols, among others.

Bibliography

Bibliography

1. Mannhold, R. & Dross, K. Calculation Procedures for Molecular Lipophilicity: A Comparative Study. *Quant. Struct. Relationships* **15,** 403–409 (**1996**).

2. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **102,** 3762–3772 (**1998**).

3. Mannhold, R. & Van De Waterbeemd, H. Substructure and Whole Molecule Approaches for Calculating log *P. J. Comput. Aided. Mol. Des.* **15,** 337–354 (**2001**).

4. Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log *P* Methods on More Than 96,000 Compounds. *J. Pharm. Sci.* **98,** 861–893 (**2009**).

5. Leo, A., Hansch, C. & Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **71,** 525 (**1971**).

6. Sangster, J. Octanol-Water Partition Coeffients of Simple Organic Compounds. *J. Phys. Chem. Ref. Data* **18,** 1111–1227 (**1989**).

7. Leo, A. J. Calculating log *P*oct from Structures. *Chem. Rev.* **93,** 1281–1306 (**1993**).

8. Sangster, J. *Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry.* ; Wiley: Chichester, (**1997**).

9. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **23,** 3–25 (**1997**).

10. Eugene Kellogg, G. & Abraham, D. J. Hydrophobicity: Is Log *P*(o/w) More than the Sum of Its Parts? *Eur. J. Med. Chem.* **35,** 651–661 (**2000**).

11. Bhal, S. K., Kassam, K., Peirson, I. G. & Pearl, G. M. The Rule of Five Revisited: Applying log *D* in place of log *P* in Drug-Likeness Filters. *Mol. Pharm.* **4,** 556–560 (**2007**).

12. Takaács-Novák, K. In Physico-Chemical Methods in Drug Discovery and Development; Mandic, Z., Ed.; IAPC Publishing: Zagreb. pp 1−52 (**2012**)

13. Muñoz-Muriedas, J. In Physico-Chemical and Computational Approaches to Drug Discovery; RSC Drug Discovery no. 23; Luque, F. J., Barril,X., Eds.; The Royal Society of Chemistry: Cambridge. pp 104−127 (**2012**)

14. Berthelot, M. P. & Jungfleisch, E. C. Sur les Lois Qui Président Au Partage D'un Corps Entre Deux Dissolvants (Expériences). *Ann. Chim. Phys.* 4th series, 396–407 (**1872**).

15. Nernst, W. Verteilung eines Stoffes zwischen zwei Lösungsmitteln und zwischen Lösungsmittel und Dampfraum. *Z Phys Chem* **8,** 110–139 (**1891**).

16. E. Meyer, H. Zur Theorie der Alkoholnarkose. *Arch. Exptl. Pathol. Pharmakol.* 42, 109–118. (**1899**).

17. Overton, E. *Studien uber die Narkose*; Fischer: Jena, Germany. (**1901**)

18. Baum, F. Naunyn-Schmiedebergs *Arch. Exp. Pharmakol.* 42, 119-137 (**1899**)

19. Meyer, H. H. Die Narkose und ihre allgemeine Theorie. *Handb. Norm. Pathol. Physiol.* **1,** 531–549 (**1927**).

20. Lifnick, R. L. Hans Horst Meyer and the Lipoid Theory of Narcosis. *Trends Pharmacol. Sci.* **10,** 265–269 (**1989**).

21. Baumann, E. and Kast, A. *Z. Physiol. Chem*. 14, 52-74 (**1890**).

22. Richet, C. *CR Soc. Biol*. (Paris). 54, 775-776 (**1893**)

23. Missner, A. & Pohl, P. 110 Years of the Meyer-Overton Rule: Predicting Membrane Permeability of Gases and Other Small Compounds. *ChemPhysChem* **10,** 1405–1414 (**2009**).

24. Pliška, V., Testa, B. & van de Waterbeemd, H. *Lipophilicity in Drug Action and Toxicology. Lipophilicity in Drug Action and Toxicology* **4,** (2008).

25. *Lipophilicity. IUPAC Compendium of Chemical Terminology* **1077,** (2014).

26. *Hydrophobicity. IUPAC Compendium of Chemical Terminology* **1137,** (2014).

27. Johnson, T. W., Gallego, R. A. & Edwards, M. P. Lipophilic Efficiency as an Important Metric in Drug Design. *J. Med. Chem.* **61,** 6401–6420 (**2018**).

28. *Hydrophilicity. IUPAC Compendium of Chemical Terminology* **1137,** (2014).

29. Gao, L. & Mccarthy, T. J. Teflon is Hydrophilic. Comments on Definitions of Hydrophobic, Shear versus Tensile Hydrophobicity, and Wettability Characterization. *Langmuir* **24,** 546–550 (**2008**).

30. Law, K.-Y. Definitions for Hydrophilicity, Hydrophobicity, and Superhydrophobicity: Getting the Basics Right. *J. Phys. Chem. Lett.* **5,** 686–688 (**2014**).

31. van Oss, C. *Interfacial Forces in Aqueous Media*. Marcel Dekker, Inc. (**1994**).

32. Willard, A. P. Illuminating the Interactions between Small Solutes in Liquid Water. *J. Phys. Chem. Lett.* **6,** 1616–1617 (**2015**).

33. Harris, R. C. & Pettitt, B. M. Reconciling the Understanding of 'Hydrophobicity' with Physics-Based Models of Proteins. *J. Phys. Condens. Matter* **28,** 083003 (**2016**).

34. Ben-Amotz, D. Water-Mediated Hydrophobic Interactions. *Annu. Rev. Phys. Chem.* **67,** 617–638 (**2016**).

35. van der Vegt, N. F. A. & Nayar, D. The Hydrophobic Effect and the Role of Cosolvents. *J. Phys. Chem. B* **121,** 9986-9998 (**2017**).

36. Lum, K., Chandler, D. & Weeks, J. D. Hydrophobicity at Small and Large Length Scales. *J. Phys. Chem. B* **103,** 4570–4577 (**1999**).

37. Chandler, D. Interfaces and the Driving Force of Hydrophobic Assembly. *Nature* **437,** 640–647 (**2005**).

38. Garde, S. & Patel, A. J. Unraveling the Hydrophobic Effect, One Molecule at a Time. *Proc. Natl. Acad. Sci.* **108,** 16491–16492 (**2011**).

39. Stone, A. *The Theory of Intermolecular Forces*. Oxford. (2013).

40. Freiría-Gándara, J., Losada-Barreiro, S., Paiva-Martins, F. & Bravo-Díaz, C. Differential Partitioning of Bioantioxidants in Edible Oil–Water and Octanol–Water Systems: Linear Free Energy Relationships. *J. Chem. Eng. Data* **63,** 2999–3007 (**2018**).

41. Toulmin, A., Wood, J. M. & Kenny, P. W. Toward Prediction of Alkane/Water Partition Coefficients. *J. Med. Chem.* **51,** 3720–3730 (**2008**).

42. Bannan, C. C., Calabro, G., Kyu, D. Y. & Mobley, D. L. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *J. Chem. Theory Comput.* **12,** 4015–4024 (**2016**).

43. Fujita, T., Iwasa, J. & Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **86,** 5175–5180 (**1964**).

44. Avdeef, A., Box, K. J., Comer, J. E. A., Hibbert, C. & Tam, K. Y. pH-Metric log *P* 10. Determination of Liposomal Membrane-Water Partition Coefficients of Ionizable Drugs. *Pharmaceutical Research* **15,** 209–215 (**1998**).

45. Balon, K., Riebesehl, B. U. & Müller, B. W. Drug Liposome Partitioning as a Tool for the Prediction of Human Passive Intestinal Absorption. *Pharm. Res.* **16,** 882–888 (**1999**).

46. Balon, K., Riebesehl, B. U. & Müller, B. W. Determination of Liposome Partitioning of Ionizable Drugs by Titration. *J. Pharm. Sci.* **88,** 802–806 (**1999**).

47. Loidl-Stahlhofen, A. et al. Multilamellar Liposomes and Solid-Supported Lipid Membranes (TRANSIL): Screening of Lipid-Water Partitioning toward a High-Throughput Scale. *Pharm. Res.* **18,** 1782–1788 (**2001**).

48. Loureiro, D. R. P. *et al.* Accessing Lipophilicity of Drugs with Biomimetic Models: A Comparative Study Using Liposomes and Micelles. *Eur. J. Pharm. Sci.* **115,** 369–380 (**2018**).

49. Tshepelevitsh, S., Hernits, K. & Leito, I. Prediction of Partition and Distribution Coefficients in Various Solvent Pairs with COSMO-RS. *J. Comput. Aided. Mol. Des.* **32,** 711–722 (**2018**).

50. Manallack, D. T. The pKa Distribution of Drugs: Application to Drug Discovery. *Perspect. Medicin. Chem.* **1,** 25–38 (**2007**).

51. Laborda, E. *et al.* Practical Application of Ligand Efficiency Metrics in Lead Optimisation. *Drug Discov. Today Technol.* **27,** 3447–3453 (**2018**).

52. Zhang, Y. *et al.* Physicochemical Property Guidelines for Modern Agrochemicals. *Pest Manag. Sci.* **74,** 1979–1991 (**2018**).

53. Avdeef, a. Physicochemical Profiling (Solubility, Permeability and Charge State). *Curr. Top. Med. Chem.* **1,** 277–351 (**2001**).

54. Scherrer, R. a & Howard, S. M. Use of Distribution Coefficients in Quantitative Structure-Activity Relationships. *J. Med. Chem.* **20,** 53–58 (**1977**).

55. Kostal, J., Voutchkova-Kostal, A., Anastas, P. T. & Zimmerman, J. B. Identifying and Designing Chemicals with Minimal Acute Aquatic Toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 6289–94 (**2015**).

56. Shappell, N. W. *et al.* Distribution of Animal Drugs among Curd, Whey, and Milk Protein Fractions in Spiked Skim Milk and Whey. *J. Agric. Food Chem.* **65,** 938–949 (**2017**).

57. Lupton, S. J., Shappell, N. W., Shelver, W. L. & Hakk, H. Distribution of Spiked Drugs between Milk Fat, Skim Milk, Whey, Curd, and Milk Protein Fractions: Expansion of Partitioning Models. *J. Agric. Food Chem.* **66,** 306–314 (**2018**).

58. Shelver, W. L., Lupton, S. J., Shappell, N. W., Smith, D. J. & Hakk, H. Distribution of Chemical Residues among Fat, Skim, Curd, Whey, and Protein Fractions in Fortified, Pasteurized Milk. *ACS Omega* **3,** 8697–8708 (**2018**).

59. Sugano, K. et al. Coexistence of Passive and Carrier-Mediated Processes in Drug Transport. *Nat. Rev. Drug Discov.* **9,** 597–614 (**2010**).

60. Andrés, A. *et al.* Setup and Validation of Shake-Flask Procedures for the Determination of Partition Coefficients (log *D*) from Low Drug Amounts. *Eur. J. Pharm. Sci.* **76,** 181–191 (**2015**).

61. Freeman-Cook, K. D., Hoffman, R. L. & Johnson, T. W. Lipophilic Efficiency: The Most Important Efficiency Metric in Medicinal Chemistry. *Future Med. Chem.* **5,** 113–115 (**2013**).

62. Ribeiro, M. M. B., Melo, M. N., Serrano, I. D., Santos, N. C. & Castanho, M. A. R. B. Drug-lipid Interaction Evaluation: Why a 19th Century Solution? *Trends Pharmacol. Sci.* **31,** 449–454 (**2010**).

63. Leeson, P. D. & Springthorpe, B. The Influence of Drug-Like Concepts on Decision-Making in Medicinal Chemistry. *Nat. Rev. Drug Discov.* **6,** 881–890 (**2007**).

64. Hopkins, A. L., Keserü, G. M., Leeson, P. D., Rees, D. C. & Reynolds, C. H. The Role of Ligand Efficiency Metrics in Drug Discovery. *Nat. Rev. Drug Discov.* **13,** 105–121 (**2014**).

65. Murray, C. W. et al. Validity of Ligand Efficiency Metrics. *ACS Med. Chem. Lett.* **5,** 616–618 (**2014**).

66. Scott, J. S. & Waring, M. J. Practical Application of Ligand Efficiency Metrics in Lead Optimisation. *Bioorganic Med. Chem.* **26,** 3006–3015 (**2018**).

67. Wenlock, M. C., Potter, T., Barton, P. & Austin, R. P. A Method for Measuring the Lipophilicity of Compounds in Mixtures of 10. *J. Biomol. Screen.* **16,** 348–55 (**2011**).

68. Shultz, M. D. Improving the Plausibility of Success with Inefficient Metrics. *ACS Med. Chem. Lett.* **5,** 2–5 (**2014**).

69. Jacobs, M. H. Some Aspects of Cell Permeability to Weak Electrolytes. *Cold Spring Harb. Symp. Quant. Biol.* **8,** 30–39 (**1940**).

70. Schanker, L. S., Shore, P. A., Brodie, B. B. & Hogben, A. M. Absorption of Drugs from Stomach I. The Rat. *J. Pharmacol. Exp. Ther* **120,** 528–539 (**1957**).

71. M. Hogben, C. A., Tocco, D. J., Schanker, L. S. & Brodie, B. B. On the Mechanism of Intestinal Absortion of Drugs. *J. Pharmacol. Exp. Ther.* **125,** 275–282 (**1958**).

72. Schanker, L. S. On the Mechanism of Absorption Of Drugs from the Gastrointestinal Tract. *J. Med. Pharm. Chem.* **2,** 343–59 (**1960**).

73. Lewis S. Schanker. Mechanisms of Drug Absorption and Distribution. *Annu. Rev. Pharmacol.* **1,** 29–45 (**1961**).

74. Divatia, B. G. J. & Biles, J. A. Physical Chemical Study of the Distribution of Some Amine Salts BetGeen Immiscible Solvents. *J. Pharm. Sci.* **50,** 916–922 (**1961**).

75. Murthy, K. S. & Zografi, G. Oil-Water Partitioning of Chlorpromazine and Other Phenothiazine Derivatives Using Dodecane and *n*-Octanol. *J. Pharm. Sci.* **59,** 1281–1285 (**1970**).

76. Kaufman, J. J., Semo, N. M. & Koski, W. S. Microelectrometric Titration Measurement of the pKa's and Partition and Drug Distribution Coefficients of Narcotics and Narcotic Antagonists and Their pH and Temperature Dependence. *J. Med. Chem.* **18,** 647–655 (**1975**).

77. Ronald D. Schoenwald, H.-S. H. Corneal Penetration Behavior of β-Blocking Agents I: Physicochemical Factors. *J. Pharm. Sci.* **72,** 1266–1272 (**1983**).

78. Schellenberg, K., Leuenberger, C. & Schwarzenbach, R. P. Sorption of Chlorinated Phenols by Natural Sediments and Aquifer Materials. *Environ. Sci. Technol.* **18,** 652–657 (**1984**).

79. Clarke, F. H. Ionization Constants by Curve Fitting: Application to the Determination of Partition Coefficients. *J. Pharm. Sci.* **73,** 226–230 (**1984**).

80. Westall, J. C., Leuenberger, C. & Schwarzenbach, R. P. Influence of pH and Ionic Strength on the Aqueous-Nonaqueous Distribution of Chlorinated Phenols. *Environ. Sci. Technol.* **19,** 193–198 (**1985**).

81. Maitani, Y., Nakagaki, M. & Nagai, T. Determination of the Acid Dissociation Constants in Ethanol-Water Mixtures and Partition Coefficients for Diclofenac. *Int. J. Pharm.* **74,** 105–116 (**1991**).

82. Berthod, A., Carda-Broch, S. & Garcia-Alvarez-Coque, M. C. Hydrophobicity of Ionizable Compounds. A Theoretical Study and Measurements of Diuretic Octanol-Water Partition Coefficients by Countercurrent Chromatography. *Anal. Chem.* **71,** 879–888 (**1999**).

83. Herbig, M. E. & Evers, D. H. Correlation of Hydrotropic Solubilization by Urea with log *D* of Drug Molecules and Utilization of This Effect for Topical Formulations. *Eur. J. Pharm. Biopharm.* **85,** 158–160 (**2013**).

84. Freiría-Gándara, J., Losada-Barreiro, S., Paiva-Martins, F. & Bravo-Díaz, C. Differential Partitioning of Bioantioxidants in Edible Oil-Water and Octanol-Water Systems: Linear Free Energy Relationships. *J. Chem. Eng. Data* **63,** 2999–3007 (**2018**).

85. Auerbach, M. E. Germicidal Quaternary Ammonium Salts in Dilute Solution A Colorimetric Assay Method. *Ind. Eng. Chem. Anal. Ed.* **15,** 492–493 (**1943**).

86. S. R Epton. A Rapid Method of Analysis for Certain Surface-Active Agents. *Nature* **160,** 909 (**1947**).

87. Mukerjee, P. Use of Ionic Dyes in the Analysis of Ionic Surfactants and other Ionic Organic Compounds. *Anal. Chem.* **28,** 870–873 (**1956**).

88. Amine, S. *et al.* Physical Chemical Study of the Distribution of Some Amine Salts Between Inmiscible Solvents II. Complexation in the Organic Phase. *J. Pharm. Sci.* **53,** 869–872 (**1963**).

89. Biles, J. A., Plakogiannis, F. M., Wong, B. J. & Biles, P. M. Distribution of Quaternary Ammonium Salts Between Chloroform and Water. *J. Pharm. Sci.* **55,** 909–913 (**1966**).

90. Levine, R. M., Blair, M. R. & Clark, B. B. Factors Influencing the Intestinal Absorption of Certain Monoquaternary Anticholinergic Compounds with Special Reference to Benzomethamine [*N*-diethylaminoethyl-*N'*-methylbenzilamide methobromide (MC-3199)]. *J. Pharmacol. Exp. Ther.* **114,** 78–86 (**1955**).

91. Högerle, M. L. & Winne, D. Drug Absorption by the Rat Jejunum Perfused in situ. *Naunyn. Schmiedebergs. Arch. Pharmacol.* **322,** 249–255 (**1983**).

92. Schanker, L. S. Mechanism of Drug Absortion and Distribution. *Annu. Rev. Pharmacol.* **1,** 29–45 (**1961**).

93. Tsuji, A., Kubo, O., Miyamoto, E., Yamana, T. Physicochemical Properties of β-Lactam Antibiotics: Oil-Water Distribution. *J. Pharm. Sci.* **66,** 1675–16795 (**1977**).

94. Clarke, F. & Cahoon, N. Ionization Constants by Curve Fitting: Determination of Partition and Distribution Coefficients of Acids and Bases and Their Ions. *J. Pharm. Sci.* **76,** 1–10 (**1987**).

95. Austin, R. P., Davis, a. M. & Manners, C. N. Partitioning of Ionizing Molecules between Aqueous Buffers and Phospholipid Vesicles. *J. Pharm. Sci.* **84,** 1180–1183 (**1995**).

96. Csizmadia, F., Tsantili-Kakoulidou, A., Panderi, I. & Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **86,** 865–871 (**1997**).

97. Pagliara, A., Carrupt, P.-A., Caron, G., Gaillard, P. & Testa, B. Lipophilicity Profiles of Ampholytes. *Chem. Rev.* **97,** 3385–3400 (**1997**).

98. Comer, J. & Tam, K. Lipophilicity Profiles: Theory and Measurement. in *Pharmacokinetic Optimization in Drug Research.* Verlag Helvetica Chimica Acta: Zürich and Wiley VCH:Weinheim. (**2001**).

99. Avdeef, A. *Absorption and Drug Development: Solubility, Permeability and Charge State.* John Wiley & Sons, Inc. (**2003**)

100. Barzanti, C. *et al.* Potentiometric Determination of Octanol-Water and Liposome-Water Partition Coefficients (log *P*) of Ionizable Organic Compounds. *Tetrahedron Lett.* **48,** 3337–3341 (**2007**).

101. Kah, M. & Brown, C. D. Log *D*: Lipophilicity for Ionisable Compounds. *Chemosphere* **72,** 1401–1408 (**2008**).

102. Fu, W., Franco, A. & Trapp, S. Methods for Estimating the Bioconcentration Factor of Ionizable Organic Chemicals. *Env. Toxicol Chem* **28,** 1372–1379 (**2009**).

103. Liu, X., Testa, B. & Fahr, A. Lipophilicity and Its Relationship with Passive Drug Permeation. *Pharm. Res.* **28,** 962–977 (**2011**).

104. Pieńko, T., Grudzień, M., Taciak, P. P. & Mazurek, A. P. Cytisine Basicity, Solvation, log *P*, and log *D* Theoretical Determination as Tool for Bioavailability Prediction. *J. Mol. Graph. Model.* **63,** 15–21 (**2016**).

105. Zhu, S. *et al.* Glycine Substitution Reduces Antimicrobial Activity and Helical Stretch of diPGLa-H in Lipid Micelles. *Biochim. Biophys. Acta Biomembr.* **113,** 4817–4822 (**2017**).

106. Li, H., Cao, Y., Zhang, D. & Pan, B. pH-Dependent K_{OW} Provides New Insights in Understanding the Adsorption Mechanism of Ionizable Organic Chemicals on Carbonaceous Materials. *Sci. Total Environ.* **618,** 269–275 (**2017**).

107. OECD/OCDE. OECD Guideline for the Testing of Chemicals: Partition Coefficient (n-Octanol/Water): Shake Flask Method. *Oecd Guidel. Test. Chem.* **107,** 1–4 (1995).

108. Port, A. et al. Critical Comparison of Shake-Flask, Potentiometric and Chromatographic Methods for Lipophilicity Evaluation (log $P_{o/w}$) Of Neutral, Acidic, Basic, Amphoteric, And Zwitterionic Drugs. *Eur. J. Pharm. Sci.* **122,** 331–340 (**2018**).

109. Wolfenden, R. Waterlogged Molecules. *Science.* **222,** 1087–1093 (1983).

110. Tissandier, M. D. et al. The Proton's Absolute Aqueous Enthalpy and Gibbs Free Energy of Solvation from Cluster-Ion Solvation Data. *J. Phys. Chem. A* **102,** 7787–7794 (**1998**).

111. Pliego, J. R. & Riveros, J. M. Gibbs Energy of Solvation of Organic Ions in Aqueous and Dimethyl Sulfoxide Solutions. *Phys. Chem. Chem. Phys.* **4,** 1622–1627 (**2002**).

112. Pliego, J. R. & Miguel, E. L. M. Absolute Single-Ion Solvation Free Energy Scale in Methanol Determined by the Lithium Cluster-Continuum Approach. *J. Phys. Chem. B* **117,** 5129–5135 (**2013**).

113. Pearson, R. G. Ionization Potentials and Electron Affinities in Aqueous Solution. *J. Am. Chem. Soc.* **108,** 6109–6114 (**1986**).

114. Winget, P., Cramer, C. J. & Truhlar, D. G. Computation of Equilibrium Oxidation and Reduction Potentials for Reversible and Dissociative Electron-Transfer Reactions in Solution. *Theor. Chem. Acc.* **112,** 217–227 (**2004**).

115. Colaizzi, J. L. & Klink, P. R. pH-Partition Behavior of Tetracyclines. *J. Pharm. Sci.* **58,** 1184–1189 (**1969**).

116. Marcus, Y. & Hefter, G. Ion-Pairing. *Chem. Rev.* **106,** 4585–4621 (**2006**).

117. Inagi, T., Muramatsu, T., Nagai, H. & Terada, H. Mechanism of Indomethacin Partition between n-Octanol and Water. *Chem. Pharm. Bull.* **29,** 2330–2337 (**1981**).

118. Ingram, T., Richter, U., Mehling, T. & Smirnova, I. Modelling of pH Dependent n-Octanol/Water Partition Coefficients of Ionizable Pharmaceuticals. *Fluid Phase Equilib.* **305,** 197–203 (**2011**).

119. Jafvert, C. T., Westall, J. C., Grieder, E. & Schwarzenbach, R. P. Distribution of Hydrophobic Ionogenic Organic Compounds Between Octanol and Water: Organic Acids. *Environ. Sci. Technol.* **24,** 1795–1803 (**1990**).

120. Takács-Novák, K. & Szász, G. Ion-Pair Partition of Quaternary Ammonium Drugs: The Influence of Counter Ions of Different Lipophilicity, Size, and Flexibility. *Pharm. Res.* **16,** 1633–1638 (**1999**).

121. Fini, A. et al. Formation of Ion-Pairs in Aqueous Solutions of Diclofenac Salts. *Int. J. Pharm.* **187,** 163–173 (**1999**).

122. Sarveiya, V., Templeton, J. F. & Benson, H. a E. Ion-Pairs of Ibuprofen: Increased Membrane Diffusion. *J. Pharm. Pharmacol.* **56,** 717–724 (**2004**).

123. Scherrer, R. A. & Donovan, S. F. Automated Potentiometric Titrations in KCl/Water-Saturated Octanol: Method for Quantifying Factors Influencing Ion-Pair Partitioning. *Anal. Chem.* **81,** 2768–2778 (**2009**).

124. Fini, A., Bassini, G., Monastero, A. & Cavallari, C. Diclofenac Salts, VIII. Effect of the Counterions on the Permeation through Porcine Membrane from Aqueous Saturated Solutions. *Pharmaceutics* **4,** 413–429 (**2012**).

125. Chen, C. S. & Lin, S. T. Prediction of pH Effect on the Octanol-Water Partition Coefficient of Ionizable Pharmaceuticals. *Ind. Eng. Chem. Res.* **55,** 9284–9294 (**2016**).

126. Rutkowska, E., Pająk, K. & Jóźwiak, K. Lipophilicity-Methods of Determination and Its Role in Medicinal Chemistry. *Acta Pol. Pharm. Drug Res.* **70,** 3–18 (**2013**).

127. Giaginis, C. & Tsantili-Kakoulidou, A. Alternative Measures of Lipophilicity: From Octanol–Water Partitioning to IAM Retention. *J. Pharm. Sci.* **97,** 2984–3004 (**2008**).

128. Valkó, K. Application of High-Performance Liquid Chromatography Based Measurements of Lipophilicity to Model Biological Distribution. *J. Chromatogr. A* **1037,** 299–310 (**2004**).

129. Wiczling, P., Waszczuk-Jankowska, M., Markuszewski, M. J. & Kaliszan, R. The Application of Gradient Reversed-Phase High-Performance Liquid Chromatography to the pK_a and log k_w Determination of Polyprotic Analytes. *J. Chromatogr. A* **1214,** 109–114 (**2008**).

130. Lapins, M. *et al.* A Confidence Predictor for log D Using Conformal Regression and A Support-Vector Machine. *J. Cheminform.* **10,** 1–10 (**2018**).

131. ACD/I-Lab. Advanced Chemistry Development, Inc.: Toronto, ON, Canada; http://www.acdlabs.com.

132. ChemAxon, Budapest, Hungary, http://www.chemaxon.com.

133. Bachs, M., Luque, F. J. & Orozco, M. Optimization of Solute Cavities and van der Waals Parameters in Ab Initio MST - SCRF Calculations of Neutral Molecules. *J. Comput. Chem.* **15,** 446–454 (**1994**).

134. Luque, F. J., Bachs, M. & Orozco, M. An Optimized AM1/MST Method for the MST-SCRF Representation of Solvated Systems. *J. Comput. Chem.* **15,** 847–857 (**1994**).

135. Curutchet, C., Orozco, M. & Luque, F. J. Solvation in Octanol: Parametrization of the Continuum MST Model. *J. Comput. Chem.* **22,** 1180–1193 (**2001**).

136. Soteras, I., Curutchet, C., Bidon-Chanal, A., Orozco, M. & Luque, F. J. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct. Theochem* **727,** 29–40 (**2005**).

137. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal Solvation Model Based On Solute Electron Density and On A Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **113,** 6378–6396 (**2009**).

138. Ghose, A. K. & Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **27,** 21–35 (**1987**).

139. Vellarkad N.V., Arup K.G., Ganapathi R.R., Roland K.R. An Estimation of the Atomic Contribution to Octanol-Water Partition Coefficient and Molar Refractivity from Fundamental Atomic and Structural Properties: Its Uses in Computer-Aided Drug Design. *Math. Comput. Model* **14,** 505–510 (**1990**).

140. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39,** 868–873 (**1999**).

141. Wang, R., Fu, Y. & Lai, L. A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **37,** 615–621 (**1997**).

142. Miertus, S., Scrocco, E. & Tomasi, J. A New Integral Equation Formalism for the Polarizable. *Chem. Phys.* **55,** 117–129 (**1981**).

143. Cancès, E., Mennucci, B. & Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **107,** 3032 (**1997**).

144. Luque, F. J., Bofill, J. M. & Orozco, M. New Strategies to Incorporate the Solvent Polarization in Self-Consistent Reaction Field and Free-Energy Perturbation Simulations. *J. Chem. Phys.* **103,** 10183–10191 (**1995**).

145. Paneth, A. et al. Lipophilicity Studies on Thiosemicarbazide Derivatives. *Molecules* **22,** (**2017**).

146. Levy, Y., Peter G. Wolynes, A. & Onuchic, J. N. Protein Topology Determines Binding Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 511–516 (**2004**).

147. Jamadagni, S. N., Godawat, R. & Garde, S. Hydrophobicity of Proteins and Interfaces: Insights from Density Fluctuations. *Annu. Rev. Chem. Biomol. Eng.* **2,** 147–71 (**2011**).

148. Fosgerau, K. & Hoffmann, T. Peptide therapeutics: Current Status and Future Directions. *Drug Discovery Today* **20,** 122–128 (**2015**).

149. Jamadagni, S. N., Godawat, R. & Garde, S. Hydrophobicity of Proteins and Interfaces: Insights from Density Fluctuations. *Annu. Rev. Chem. Biomol. Eng.* **2,** 147–171 (**2011**).

150. Patel, A. J. & Garde, S. Efficient Method to Characterize the Context-Dependent Hydrophobicity of Proteins. *J. Phys. Chem. B* **118,** 1564–1573 (**2014**).

151. Loo, J. A., Loo, R. R. O., Udseth, H. R., Edmonds, C. G. & Smith, R. D. Solvent-Induced Conformational Changes of Polypeptides Probed by Electrospray-Ionization Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **5,** 101–105 (**1991**).

152. Nozaki, Y. & Tanford, C. The Solubility of Amino Acids and Two Glycine Peptides in Aqueous Ethanol and Dioxane Solutions: Establishment of a Hydrophobicity Scale. *J. Biol. Chem.* **246,** 2211–2217 (**1971**).

153. Chothia, C. Hydrophobic Bonding and Accessible Surface Area In Proteins. *Nature* **248,** 338–339 (**1974**).

154. Reynolds, J. A., Gilbert, D. B. & Tanford, C. Empirical Correlation Between Hydrophobic Free Energy and Aqueous Cavity Surface Area. *Proc. Natl. Acad. Sci.* **71,** 2925–2927 (**1974**).

155. Fauchère, J. L. & Pliška, V. Hydrophobic Parameters π of Amino Acid Side Chains from the Partitioning of *N*-Acetyl-*L*-Amino Acid Amides. *Eur. J. Med. Chem.* **18,** 369–375 (**1983**).

156. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. Hydrophobicity of Amino Acid Residues In Globular Proteins. *Science* **229,** 834 (**1985**).

157. Eisenberg, D. & McLachlan, A. D. Solvation Energy in Protein Folding and Binding. *Nature* **319,** 199–203 (**1986**).

158. Wimley, W. C., Gawrisch, K., Creamer, T. P. & White, S. H. Direct Measurement of Salt-Bridge Solvation Energies Using a Peptide Model System: Implications for Protein Stability. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 2985–2990 (**1996**).

159. Moon, C. P. & Fleming, K. G. Side-Chain Hydrophobicity Scale Derived from Transmembrane Protein Folding into Lipid Bilayers. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 10174–7 (**2011**).

160. Simm, S., Einloft, J., Mirus, O. & Schleiff, E. 50 Years of Amino Acid Hydrophobicity Scales: Revisiting the Capacity for Peptide Classification. *Biol. Res.* **49,** 31 (**2016**).

161. Peters, C. & Elofsson, A. Why Is the Biological Hydrophobicity Scale More Accurate than Earlier Experimental Hydrophobicity Scales? *Proteins* **82,** 2190–8 (**2014**).

162. MacCallum, J. L. & Tieleman, D. P. Hydrophobicity Scales: A Thermodynamic Looking Glass into Lipid-Protein Interactions. *Trends Biochem Sci* **36,** 653–662 (**2011**).

163. Wall, D., Eisenberg, E., Schwarz, M. & Komaromy, R. Analysis of Membrane and Surface Protein Sequences with the Hydrophobic Moment Plot. *J. Mol. Biol.* **179,** 125–142 (**1984**).

164. Hopp, T. P. & Woods, K. R. Prediction of Protein Antigenic Determinants from Amino Acid Sequences. *Proc. Natl. Acad. Sci.* **78,** 3824–3828 (**1981**).

165. Wimley, W. C., Creamer, T. P. & White, S. H. Solvation Energies of Amino Acid Side Chains and Backbone in a Family of Host-Guest Pentapeptides. *Biochemistry* **35,** 5109–5124 (**1996**).

166. Kyte, J. & Doolittle, R. F. A Simple Method for Displaying The Hydropathic Character of a Protein. *J. Mol. Biol.* **157,** 105–132 (**1982**).

167. Moon, C. P. & Fleming, K. G. Side-Chain Hydrophobicity Scale Derived from Transmembrane Protein Folding into Lipid Bilayers. *Proc. Natl. Acad. Sci.* **108,** 10174–10177 (**2011**).

168. Hessa, T. et al. Recognition of Transmembrane Helices by the Endoplsmic Reticulum Translocon. *Nature* **433,** 377–381 (**2005**).

169. Koehler, J., Woetzel, N., Staritzbichler, R., Sanders, C. R. & Meiler, J. A Unified Hydrophobicity Scale for Multispan Membrane Proteins. *Proteins Struct. Funct. Bioinforma.* **76,** 13–29 (**2009**).

170. Janin, J. Surface and Inside Volumne in Globular Proteins. *Nature* **277,** 491–492 (**1979**).

171. Pierotti, R. A. A Scaled Particle Theory of Aqueous and Nonaqueous Solutions. *Chem. Rev.* **76,** 717–726 (**1976**).

172. Claverie, P. et al. Studies Of Solvent Effects. 1. Discrete, Continuum, and Discrete-Continuum Models and Their Comparison for Some Simple Cases: Ammonium (1+) Ion, Methanol, and Substituted Ammonium (1+) Ion. *J. Phys. Chem.* **82,** 405–418 (**1978**).

173. Luque, F. J., Barril, X. & Orozco, M. Fractional Description of Free Energies of Solvation. *J. Comput. Aided. Mol. Des.* **13,** 139–152 (**1999**).

174. Hansch, C., Leo, A. & Hoekman, D. Exploring QSAR: Hydrophobic, Electronic, and Steric Constants. In *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*. American Chemical Society. (**1995**).

175. Curutehet, C., Bidon-Chanal, A., Soteras, I., Orozco, M. & Luque, F. J. MST Continuum Study of the Hydration Free Energies of Monovalent Ionic Species. *J. Phys. Chem. B* **109,** 3565–3574 (**2005**).

176. Reymond, F. *et al.* Ionic Partition Diagrams Of Ionisable Drugs: pH-Lipophilicity Profiles, Transfer Mechanisms and Charge Effects On Solvation. *J. Electroanal. Chem.* **462,** 235–250 (**1999**).

177. Luque, F. J., Orozco, M., Bhadane, P. K. & Gadre, S. R. Effect of Solvation on the Shapes, Sizes, and Anisotropies of Polyatomic Anions Via Molecular Electrostatic Potential Topography: An Ab Initio Self-Consistent Reaction Field Approach. *J. Chem. Phys.* **100,** 6718–6728 (**1994**).

178. Luque, F. J., Gadre, S. R., Bhadane, P. K. & Orozco, M. The Effect of Hydration on the Molecular Charge Distribution of Cations. An Ab Initio SCRF Study. *Chem. Phys. Lett.* **232,** 509–517 (**1995**).

179. Mestres, J., Solà, M., Carbó, R., Luque, F. J. & Orozco, M. Effect of Solvation on the Charge Distribution of a Series of Anionic, Neutral, and Cationic Species. A Quantum Molecular Similarity Study. *J. Phys. Chem.* **100,** 606–610 (**1996**).

180. Oliveira, P. R. de, Tasic, L., Rocco, S. A. & Rittner, R. Stereoelectronic and Inductive Effects on 1H and 13C NMR Chemical Shifts of Some *cis*-1,3-Disubstituted Cyclohexanes. *Magn. Reson. Chem.* **44,** 790–796 (**2006**).

181. Hung, L. Q. Electrochemical Properties of the Interface between Two Immiscible Electrolyte Solutions. *J. Electroanal. Chem.* **115,** 159–174 (**1980**).

182. Kakiuchi, T. Limiting Behavior in Equilibrium Partitioning of Ionic Components in Liquid−Liquid Two-Phase Systems. *Anal. Chem.* **68,** 3658 (**1996**).

183. Bouchard, G., Carrupt, P. A., Testa, B., Gobry, V. & Girault, H. H. The Apparent Lipophilicity of Quaternary Ammonium Ions Is Influenced by Galvani Potential Difference, Not Ion-Pairing: A Cyclic Voltammetry Study. *Pharm. Res.* **18,** 702–708 (**2001**).

184. Peljo, P. & Girault, H. H. Liquid/Liquid Interfaces, Electrochemistry at *Encycl. Anal. Chem.*, © 2000, John Wiley & Sons, Ltd. (2012).

185. Petrauskas, A. A. & Kolovanov, E. A. ACD/Log *P* Method Description. *Perspect. Drug Discov. Des.* **19,** 99–116 (**2000**).

186. Mazák, K. & Noszál, B. Drug Delivery: A Process Governed by Species-Specific Lipophilicities. *Eur. J. Pharm. Sci.* **62,** 96–104 (**2014**).

187. Klamt, A., Huniar, U., Spycher, S. & Keldenich, J. COSMOmic: A Mechanistic Approach to the Calculation of Membrane-Water Partition Coefficients and Internal Distributions within Membranes and Micelles. *J. Phys. Chem. B* **112,** 12148–12157 (**2008**).9

188. Juárez-Jiménez, J., Barril, X., Orozco, M., Pouplana, R. & Luque, F. J. Assessing the Suitability of the Multilevel Strategy for the Conformational Analysis of Small Ligands. *J. Phys. Chem. B* **119,** 1164–1172 (**2015**).

189. Yordanova, D. et al. Solute Partitioning in Micelles: Combining Molecular Dynamics Simulations, COSMOmic, and Experiments. *J. Phys. Chem. B* **121,** 5794–5809 (**2017**).

190. Dunbrack, R. L. & Karplus, M. Conformational Analysis of the Backbone-Dependent Rotamer Preferences of Protein Side Chains. *Nat. Struct. Biol.* **1,** 334–40 (**1994**).

191. Bouchard, G. *et al.* Standard Partition Coefficients of Anionic Drugs in the *n*-Octanol/Water System Determined by Voltammetry at Three-Phase Electrodes. *Phys. Chem. Chem. Phys.* **5,** 3748–3751 (**2003**).

192. Arnold, M. R., Kremer, W., Lüdemann, H. D. & Kalbitzer, H. R. 1H-NMR Parameters of Common Amino Acid Residues Measured In Aqueous Solutions of The Linear Tetrapeptides Gly-Gly-X-Ala at Pressures Between 0.1 And 200 MPa. *Biophys. Chem.* **96,** 129–140 (**2002**).

193. Kortemme, T. & Creighton, T. E. Ionisation of Cysteine Residues at the Termini of Model α-Helical Peptides. Relevance to Unusual Thiol pK_a Values in Proteins of the Thioredoxin Family. *J. Mol. Biol.* **253,** 799–812 (**1995**).

194. Harms, M. J., Schlessman, J. L., Sue, G. R. & Garcia-Moreno E., B. Arginine Residues at Internal Positions in a Protein Are Always Charged. *Proc. Natl. Acad. Sci.* **108,** 18954–18959 (**2011**).

195. Grimsley, G. R., Scholtz, J. M. & Pace, C. N. A Summary of the Measured p*K* Values of the Ionizable Groups in Folded Proteins. *Protein Sci.* **18,** 247–251 (**2009**).

196. Musafia, B., Buchner, V. & Arad, D. Complex Salt Bridges in Proteins: Statistical Analysis of Structure and Function. *J. Mol. Biol.* **254,** 761–770 (**1995**).

197. Tomlinson, J. H., Ullah, S., Hansen, P. E. & Williamson, M. P. Characterization of Salt Bridges to Lysines in the Protein G B1 Domain. *J. Am. Chem. Soc.* **131,** 4674–4684 (**2009**).

198. Isom, D. G., Castaneda, C. A., Cannon, B. R. & Garcia-Moreno E., B. Large Shifts in pK_a Values of Lysine Residues Buried Inside a Protein. *Proc. Natl. Acad. Sci.* **108,** 5260–5265 (**2011**).

199. André, I., Linse, S. & Mulder, F. A. A. Residue-Specific pK_a Determination of Lysine and Arginine Side Chains by Indirect ^{15}N And ^{13}C NMR Spectroscopy: Application to Apo Calmodulin. *J. Am. Chem. Soc.* **129,** 15805–15813 (**2007**).

200. MacCallum, J. L., Bennett, W. F. D. & Tieleman, D. P. Partitioning Of Amino Acid Side Chains Into Lipid Bilayers: Results from Computer Simulations and Comparison to Experiment. *J. Gen. Physiol.* **129,** 371–7 (**2007**).

201. Kapcha, L. H. & Rossky, P. J. A Simple Atomic-Level Hydrophobicity Scale Reveals Protein Interfacial Structure. *J. Mol. Biol.* **426,** 484–498 (**2014**).

202. Hubbard, S. & Thornton, J. *NACCESS* Department of Biochemistry and Molecular Biology (1993).

203. Wilce, M. C. J., Aguilar, M. I. & Hearn, M. T. W. Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Anal. Chem.* **67,** 1210–1219 (**1995**).

204. Biswas, K. M., DeVido, D. R. & Dorsey, J. G. Evaluation Of Methods For Measuring Amino Acid Hydrophobicities And Interactions. *Journal of Chromatography A* **1000,** 637–655 (**2003**).

205. König, G., Bruckner, S. & Boresch, S. Absolute Hydration Free Energies of Blocked Amino Acids: Implications for Protein Solvation and Stability. *Biophys. J.* **104,** 453–62 (**2013**).

206. Amrhein, S., Oelmeier, S. A., Dismer, F. & Hubbuch, J. Molecular Dynamics Simulations Approach for the Characterization of Peptides with Respect to Hydrophobicity. *J. Phys. Chem. B* **118,** 1707–14 (**2014**).

207. Houghten, R. A. et al. Effect of Positional Environmental Domains on the Variation of High-Performance Liquid Chromatographic Peptide Retention Coefficients. *J. Chromatogr.* **386,** 223–228 (**1987**).

208. Reimer, J., Spicer, V. & Krokhin, O. V. Application of Modern Reversed-Phase Peptide Retention Prediction Algorithms to the Houghten and Degraw Dataset: Peptide Helicity and Its Effect On Prediction Accuracy. *J. Chromatogr. A* **1256,** 160–8 (**2012**).

209. Buchwald, P. & Bodor, N. Octanol-Water Partition of NonZwitterionic Peptides: Predictive Power of a Molecular Size-Based Model. *Proteins Struct. Funct. Genet.* **30,** 86–99 (**1998**).

210. Darr, J. P. *et al.* The Hydropathy Scale as a Gauge of Hygroscopicity in Sub-Micron Sodium Chloride-Amino Acid Aerosols. *J. Phys. Chem. A* **122,** (**2018**).

211. Kister, A. E. & Phillips, J. C. A Stringent Test for Hydrophobicity Scales: Two Proteins with 88% Sequence Identity but Different Structure and Function. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 9233–9237 (**2008**).

212. Xi, E. et al. Hydrophobicity of Proteins and Nanostructured Solutes is Governed by Topographical and Chemical Context. *Proc. Natl. Acad. Sci.* **114,** 13345–13350 (**2017**).

213. Kabsch, W. & Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features. *Biopolymers* **22,** 2577–2637 (**1983**).

214. Sharma, M. *et al.* Insight into the Mechanism of the Influenza A Proton Channel from a Structure in a Lipid Bilayer. *Science* **330,** 509–12 (**2010**).

215. Mortenson, D. E. *et al.* High-Resolution Structures of a Heterochiral Coiled Coil. *Proc. Natl. Acad. Sci.* **112,** 13144–13149 (**2015**).

216. Dougherty, D. A. The Cation-π Interaction. *Acc. Chem. Res.* **46,** 885–893 (**2013**).

217. Pinheiro, S. et al. Structural and energetic study of cation–π–cation interactions in proteins. *Phys. Chem. Chem. Phys.* **19,** 9849–9861 (**2017**).

218. Chen, J., Wang, J., Ulrich, J., Yin, Q. & Xue, L. Effect of Solvent on the Crystal Structure and Habit of Hydrocortisone. *Cryst. Growth Des.* **8,** 1490–1494 (**2008**).

219. Srinivasan, K., Sankaranarayanan, K., Thangavelu, S. & Ramasamy, P. Influence of Organic Solvents on the Habit of NMBA (4-nitro-4′-methyl benzylidene aniline) Crystals. *J. Cryst. Growth* **212,** 246–254 (**2000**).

220. Eyal, E., Gerzon, S., Potapov, V., Edelman, M. & Sobolev, V. The Limit of Accuracy of Protein Modeling: Influence of Crystal Packing on Protein Structure. *J. Mol. Biol.* **351,** 431–442 (**2005**).

221. Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. Two Crystal Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein G and Comparison with NMR. *Biochemistry* **33,** 4721–4729 (**1994**).

222. Rose, P. W. *et al.* The RCSB Protein Data Bank: Integrative View of Protein, Gene and 3D Structural Information. *Nucleic Acids Res.* **45,** D271–D281 (**2017**).

223. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 Update. *Nucleic Acids Res.* **47,** 339–343 (**2018**).

224. Chen, Y. & Barkley, M. D. Toward Understanding Tryptophan Fluorescence in Proteins. *Biochemistry* **37,** 9976–9982 (**1998**).

225. D'Auria, S., Staiano, M., Kuznetsova, I. M. & Turoverov, K. K. The Combined Use of Fluorescence Spectroscopy and X-Ray Crystallography Greatly Contributes to Elucidating Structure and Dynamics of Proteins. in *Reviews in Fluorescence* 25–61 (**2005**).

226. De Strooper, B. *et al.* Deficiency of Presenilin-1 Inhibits the Normal Cleavage of Amyloid Precursor Protein. *Nature* **391,** 387–390 (**1998**).

227. Haass, C. & Selkoe, D. J. A Technical KO of Amyloid-β Peptide. *Nature* **391,** 339–340 (**1998**).

228. Eisenberg, D. & Jucker, M. The Amyloid State of Proteins in Human Diseases. *Cell* **148,** 1188–203 (**2012**).

229. Tiwari, M. K. & Kepp, K. P. Modeling the Aggregation Propensity and Toxicity of Amyloid-β Variants. *J. Alzheimer's Dis.* **47,** 215–229 (**2015**).

230. Somavarapu, A. K. & Kepp, K. P. Loss of Stability and Hydrophobicity of Presenilin 1 Mutations Causing Alzheimer's Disease. *J. Neurochem.* **137,** 101–111 (**2016**).

231. Kim, W. & Hecht, M. H. Generic Hydrophobic Residues Are Sufficient to Promote Aggregation of the Alzheimer's $A\beta_{42}$ Peptide. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 15824–9 (**2006**).

232. Pouplana, R. & Campanera, J. M. Energetic Contributions of Residues to the Formation of Early Amyloid-β Oligomers. *Phys. Chem. Chem. Phys.* **17,** 2823–2837 (**2015**).

233. Somavarapu, A. K. & Kepp, K. P. Direct Correlation of Cell Toxicity to Conformational Ensembles of Genetic Aβ Variants. *ACS Chem. Neurosci.* **6,** 1990–1996 (**2015**).

234. Edwards, I. A. *et al.* Contribution of amphipathicity and hydrophobicity to the antimicrobial activity and cytotoxicity of β-hairpin peptides. *ACS Infect. Dis.* **2,** 442–450 (2016).

235. Sipe, J. D. *Amyloid Proteins. Proteins* (2005).

236. Nguyen, P. H., Sterpone, F., Pouplana, R., Derreumaux, P. & Campanera, J. M. Dimerization Mechanism of Alzheimer $A\beta_{40}$ Peptides: The High Content of Intrapeptide-Stabilized Conformations in A2V and A2T Heterozygous Dimers Retards Amyloid Fibril Formation. *J. Phys. Chem. B* **120,** 12111–12126 (**2016**).

237. Murray, B., Sharma, B., Belfort, G. & Isermann, H. P. N-Terminal Hypothesis for Alzheimer's Disease. *ACS Chem. Neurosci.* **8,** 432–434 (**2017**).

238. Serra-Batiste, M. et al. $A\beta_{42}$ Assembles into Specific β-barrel Pore-Forming Oligomers in Membrane-Mimicking Environments. *Proc. Natl. Acad. Sci.* **113,** 10866–10871 (**2016**).

239. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. Gaussian 09, revision D.01; Gaussian, Inc.: Wallingford CT, 2009. 2009 (2009).

240. Abraham, M. H. & Acree, W. E. The Transfer of Neutral Molecules, Ions and Ionic Species from Water to Wet Octanol. *Phys. Chem. Chem. Phys.* **12,** 13182–13188 (**2010**).

241. Quentel, F., Mirčeski, V. & L'Her, M. Electrochemical Study of the Thermodynamics and Kinetics of Hydrophilic Ion Transfers across Water | n-Octanol Interface. *J. Solid State Electrochem.* **12,** 31–39 (**2008**).

242. Dunbrack, R. L. & Karplus, M. Backbone-Dependent Rotamer Library for Proteins: Application to Side-Chain Prediction. *J. Mol. Biol.* **230,** 543–574 (**1993**).

243. Shapovalov, M. V. & Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **19,** 844–858 (**2011**).

244. Zamora, W. J., Curutchet, C., Campanera, J. M. & Luque, F. J. Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus-Scrocco-Tomasi Continuum Solvation Calculations. *J. Phys. Chem. B* **121,** 9868–9880 (**2017**).

245. Luque, F. J. et al. Continuum Solvation Models: Dissecting the Free Energy of Solvation. *Phys. Chem. Chem. Phys.* **5,** 3827–3836 (**2003**).

246. Ginex, T. et al. Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Models in 3D-QSAR. *J. Comput. Chem.* **37,** 1147–1162 (**2016**).

247. Vázquez, J. et al. Development and Validation of Molecular Overlays Derived from Three-Dimensional Hydrophobic Similarity with PharmScreen. *J. Chem. Inf. Model.* **58,** 1596–1609 (**2018**).

248. Pace, C. N. et al. Contribution of Hydrogen Bonds to Protein Stability. *Protein Sci.* **23,** 652–661 (**2014**).

Appendices

Appendix I. Article I:

Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus−Scrocco−Tomasi Continuum Solvation Calculations.

W. J. Zamora, C. Curutchet, J. M. Campanera and F. J. Luque, *Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus−Scrocco−Tomasi Continuum Solvation Calculations*, J. Phys. Chem. B **2017**, 121(42), 9868–9880.

Toward Refined Theoretical Models for the Description of Lipophilicity in Biomolecules

Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus−Scrocco−Tomasi Continuum Solvation Calculations

*Published as part of The Journal of Physical Chemistry virtual special issue "Manuel Yáñez and Otilia Mó Festschrift".*

William J. Zamora,[†,‡] Carles Curutchet,[‡] Josep M. Campanera,*,[‡] and F. Javier Luque*,[†]

[†]Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, and Institut de Biomedicina (IBUB), Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain

[‡]Departament de Farmàcia i Tecnologia Farmacèutica i Fisicoquímica, and Institut de Biomedicina (IBUB), Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Avgda. Diagonal 643, 08028 Barcelona, Spain

S *Supporting Information*

ABSTRACT: Hydrophobicity is a key physicochemical descriptor used to understand the biological profile of (bio)-organic compounds as well as a broad variety of biochemical, pharmacological, and toxicological processes. This property is estimated from the partition coefficient between aqueous and nonaqueous environments for neutral compounds (P_N) and corrected for the pH-dependence of ionizable compounds as the distribution coefficient (D). Here, we have extended the parametrization of the Miertus−Scrocco−Tomasi continuum solvation model in *n*-octanol to nitrogen-containing hetero-cyclic compounds, as they are present in many biologically

relevant molecules (e.g., purines and pyrimidines bases, amino acids, and drugs), to obtain accurate log P_N values for these molecules. This refinement also includes solvation calculations for ionic species in *n*-octanol with the aim of reproducing the experimental partition of ionic compounds (P_I). Finally, the suitability of different formalisms to estimate the distribution coefficient for a wide range of pH values has been examined for a set of small acidic and basic compounds. The results indicate that in general the simple pH-dependence model of the ionizable compound in water suffices to predict the partitioning at or around physiological pH. However, at extreme pH values, where ionic species are predominant, more elaborate models provide a better prediction of the *n*-octanol/water distribution coefficient, especially for amino acid analogues. Finally, the results also show that these formalisms are better suited to reproduce the experimental pH-dependent distribution curves of log D for both acidic and basic compounds as well as for amino acid analogues.
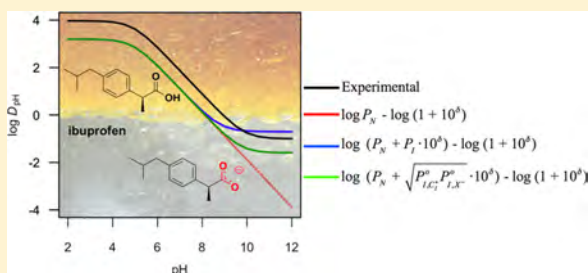
■ INTRODUCTION

The differential solubility of solutes in aqueous and nonaqueous (organic) environments is a fundamental physicochemical property for understanding a wide range of biochemical, pharmacological,

Scheme 1. Thermodynamic Cycle Used to Determine the Transfer Free Energy of a Compound (X) between Two Immiscible Solvents

$$\log P_N = \frac{-\Delta G^{o/w}}{2.303RT} = \frac{\Delta G^w_{sol} - \Delta G^o_{sol}}{2.303RT}$$
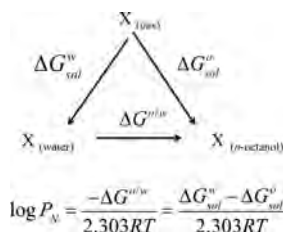
Scheme 2. Mechanism of *n*-Octanol/Water Partition for an Ionizable Neutral Compound (HX)
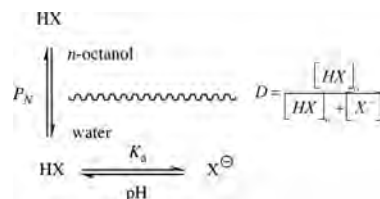
and toxicological processes of bioactive compounds.[1−8] These studies have primarily relied on molecular hydrophobicity, a property that can be quantified by the partition coefficient (P_N)

of a neutral molecule (X) between water and an organic phase, typically n-octanol (eq 1).

$$P_N = \frac{[X]_o}{[X]_w} \tag{1}$$

The transfer free energy of the solute between water and n-octanol ($\Delta G^{o/w}$) can be related to the difference in the solvation free energy upon transfer from the gas phase to the two solvents (ΔG_{sol}^w and ΔG_{sol}^o; Scheme 1). From a computational point of view, quantum mechanical (QM) self-consistent continuum solvation methods have proved to be a cost-effective approach for the calculation of solvation free energies.[9−13] Indeed, these methods have been carefully parametrized to predict the solvation free energies of neutral compounds in a wide variety of solvents, typically with an uncertainty less than 1 kcal/mol.[14,15]

For an ionizable compound (HX), it is generally assumed that only the neutral species can partition between water and n-octanol, whereas both neutral and ionized species may exist at a given pH in aqueous solution (Scheme 2). Under these circumstances, the total partitioning of the compound between aqueous and organic phases is better described by the distribution coefficient (D), which depends on the pH of the aqueous solution (eq 2).[7,16]

$$\log D = \log P_N - \log(1 + 10^\delta) \tag{2}$$

where $\delta = \text{pH} - pK_a$ for acids and $\delta = pK_a - \text{pH}$ for bases.

Most druglike compounds included in chemical libraries contain ionizable groups;[16−20] therefore, distinct neutral and ionized species may exist at a given pH. For an acidic compound, the distribution coefficient will be close to the partition coefficient at low pH, but the hydrophobicity profile will be affected by the partitioning of the ionized compound at high pH.[21] At this point, it is well-known that the absorption of bioactive compounds is influenced by the pH changes along the human gastrointestinal tract, with a maximal absorption of weak acids in the jejunum (pH $\approx$ 4.5) and weak bases in the ileum (pH $\approx$ 8.0).[22] Furthermore, it has been noticed that the pH-dependent hydrophobicity profile may be influenced not only by the partition of distinct neutral and ionic species but also by the potential contribution due to the formation of ion pairs formed with counterions.[23−25]

Predicting the hydrophobicity profile of ionizable compounds challenges the suitability of continuum solvation models for estimating the differential solvation of ionic species with chemical accuracy. This can be justified by the larger experimental uncertainties associated with experimental values of the solvation free energies of charged species compared to neutral ones.[26−28] Thus, the solvation free energy of neutral solutes is generally determined from partition coefficients between the gas phase and aqueous solutions, and the experimental uncertainty increases with the solvation free energy, limiting the applicability of this technique to solutes with solvation free energy (in absolute terms) less than ~12 kcal/mol.[29] Accordingly, ionic compounds require the use of indirect approaches based on the use of suitable thermodynamic cycles.[26−28,30,31] On the other hand, the strong solute−solvent interactions existing between ionic species and polar solvent molecules in the first solvation shells may hardly be captured from the crude representation of a polarizable continuum, which is better suited to account for bulk solvent electrostatic effects, making it necessary to carry out a careful adjustment of electrostatic and nonelectrostatic contributions to the solvation free energy.

The aim of this study is twofold. First, we report the refinement of the Miertus−Scrocco−Tomasi (MST) continuum solvation model,[10,32] which relies on the integral equation formalism of the polarizable continuum model (IEFPCM),[33] to account for the solvation free energy of nitrogen-containing heterocyclic molecules, as well as ionic compounds, in n-octanol. This is accomplished within the framework of the B3LYP/6-31G(d) version of the integral IEFPCM/MST model,[34] taking advantage of the experimental data compiled for a variety of neutral and ionic species in this solvent. Second, the MST model is used to determine the pH-dependent hydrophobicity profile taking into account different physicochemical models for the partition of ionizable compounds. In particular, attention will be paid to the experimental distribution curves of $\log D$ of acidic compounds that have already been reported in previous studies,[16,35,36] but also for basic compounds and for amino acid analogues, which have been scarcely examined in the literature.[37,38]

■ THEORY AND COMPUTATIONAL DETAILS

Physicochemical Models of Hydrophobicity Profile. Scheme 2 shows the simplest and most widely used model to account for the pH dependence on the partition of ionizable compounds. Nevertheless, more elaborate models have been proposed to refine the distribution model of these compounds. The most straightforward correction comes from the assumption that a certain amount of the ionic species may also partition between water and n-octanol (Scheme 3). In this context, for a

Scheme 3. Mechanism of n-Octanol/Water Partition for Both Neutral (HX) and Ionic (X⁻) Species
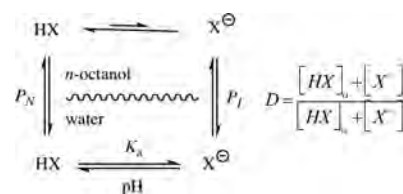
monoprotic acid (HX) the total partition of the solute can be expressed in terms of the partition constant of the neutral compound (P_N; eq 1) and of the ionic species (P_I; eq 3), as noted in eq 4.[39]

$$P_I = \frac{[X^-]_o}{[X^-]_w} \tag{3}$$

$$\log D = \log(P_N + P_I \cdot 10^\delta) - \log(1 + 10^\delta) \tag{4}$$

More elaborate models take into account the fact that the distribution of an ionizable compound may be also influenced by the electric potential created at the water−organic interphase ($\Delta_o^w \phi$), which would affect the partition of the ionic species (X⁻) and other counterions (C_I^+) present in solution (Scheme 4). At equilibrium, the ion distribution is determined by the equality of the electrochemical potential between two immiscible electrolyte solutions, and the apparent partition coefficient of X⁻ ($\log P_{I,X^-}$) can be determined from eq 5.[40,41]

$$\log P_{I,X^-} = \log P_{I,X^-}^\circ + \frac{z_X F}{RT \ln 10} \Delta_o^w \phi \tag{5}$$

where $\log P_{I,X^-}^\circ$ represents the partition coefficient for a nonpolarized interface and depends only on the chemical structure

Scheme 4. Mechanism of *n*-Octanol/Water Partition for Neutral (HX) and Ionic Species (X^- and C_I^+) Influenced by the Electric Potential at the Interphase ($\Delta_o^w \phi$)
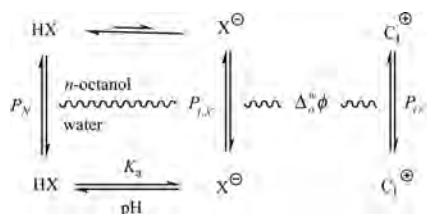
of X^-; $\Delta_o^w \phi$ is the Galvani potential difference between the two phases, R the gas constant, T the absolute temperature, and F the Faraday constant; z_X stands for the formal charge of X^-.

Assuming that the two immiscible electrolyte solutions are dilute, and that all ionic species are fully dissociated in both phases (i.e., no ion pair formation), it has been shown that for a generic electrolyte (C^+A^-) the Galvani potential difference can be rewritten as[42,43]

$$\Delta_o^w \phi = \left(\frac{\Delta_o^w \phi_{C^+}^o + \Delta_o^w \phi_{A^-}^o}{2} \right) \tag{6}$$

Because the standard partition coefficient of a given ionic species i is given by[40]

$$\log P_{I,i}^o = -\frac{z_i F}{RT \ln 10} \Delta_o^w \phi_i^o \tag{7}$$

it can be deduced that the apparent partition of X^- can be expressed as

$$P_X^{app} = \sqrt{P_X^o \cdot P_{C_I^+}^o} \tag{8}$$

where C_I^+ denotes the corresponding counterion of species X^-.

The distribution coefficient is given by

$$\log D = \log(P_N + \sqrt{P_{I,C_I^+}^o P_{I,X^-}^o} \cdot 10^\delta) - \log(1 + 10^\delta) \tag{9}$$

which explains why the distribution coefficient increases in the presence of a more hydrophobic counterion.[41]

As a final remark, let us note that partitioning of a cationic species, X^+, would give rise to distribution coefficients formally analogous to eqs 4 and 9, which are omitted here for the sake of brevity. It is also worth noting that the preceding formalisms limit the distribution coefficient to the partition of both neutral and ionic species of an ionizable compound. However, it is conceivable that partitioning may also involve other chemical entities, especially for nondilute solutions, such as ionic pairs with counterions present in solution[30,44,45] and formation of molecular aggregates.[46,47] This represents an additional level of complexity to the partitioning scheme, making it necessary to account for thermodynamic data regarding association equilibria and partitioning of the ion pair and other aggregated species, which is beyond the scope of this study.

MST Model. In the Miertus−Scrocco−Tomasi model, the solvation free energy (ΔG_{sol}) is calculated by adding nonelectrostatic (ΔG_{n-ele}) and electrostatic (ΔG_{ele}) contributions, which are calculated using a double molecule-shaped cavity for the solute embedded in the polarizable continuum medium.[10] The nonelectrostatic component is determined by combining cavitation (ΔG_{cav}) and van der Waals (ΔG_{vW}) contributions; ΔG_{cav} is determined by following Pierotti's scaled particle theory[48] adapted to molecule-shaped cavities using the procedure

proposed by Claverie (eq 10).[49] In turn, ΔG_{vW} is computed using a linear relationship to the solvent-exposed surface of each atom (eq 11). Both cavitation and van der Waals terms are estimated by using the van der Waals surface of the solute.

$$\Delta G_{cav} = \sum_{i=1}^{N} \Delta G_{cav,i} = \sum_{i=1}^{N} \frac{S_i}{S_T} \Delta G_{P,i} \tag{10}$$

where $\Delta G_{P,i}$ is the cavitation free energy of atom i determined using Pierotti's formalism, whose contribution is weighted by the contribution of the solvent-exposed surface (S_i) of atom i to the total surface (S_T).

$$\Delta G_{vW} = \sum_{i=1}^{N} \Delta G_{vW,i} = \sum_{i=1}^{N} \xi_i S_i \tag{11}$$

where ξ_i denotes the atomic surface tension of atom i, which is determined by fitting the experimental free energy of solvation.

The electrostatic term (ΔG_{ele}) measures the work needed to build up the solute charge distribution in the solvent. To this end, a solvent-excluded surface is obtained by scaling the atomic radii by a factor (λ) of 1.25 for solvation in water and 1.50 for solvation in *n*-octanol.[32,34] These scaling factors were derived from a systematic analysis that included the comparison between the electrostatic component obtained from MST calculations and the work required to annihilate the solute charge in solution as determined from classical free-energy calculations. However, while this strategy is valuable for describing the solvation of neutral solutes, accounting for the strong electrostatic response of the solvent induced by ionic species, and the structural perturbation of the solvent molecules in the first hydration shell relative to the bulk solvent,[50−52] was treated by reducing the solvent-excluded surface in the IEFPCM/MST model.[53] Thus, the optimum cavity for the hydration of charged compounds was defined by scaling the atomic radii of the groups bearing the formal charge by a factor of ∼1.13, which implies a reduction of ca. 10% relative to neutral solutes.

Besides retaining the simplicity of the original MST formalism, this strategy introduces a minimum number of parameters to describe the hydration of ionic species. However, it is unclear to what extent the reduction in the solvent-excluded surface is well-suited to the calculation of the electrostatic component of the solvation free energy in nonaqueous solvents. In particular, one of our aims here is to check the suitability of this strategy for describing the partition of ionic species between water and *n*-octanol, taking advantage of the availability of P_I values for a representative number of druglike compounds.

Computational Details. For the purpose of this study, several sets of molecules were used to refine the MST model. First, a set of 19 small nitrogen-containing aromatic compounds was used to calibrate the parametrization of pyridine-type and pyrrole-type nitrogen atoms. This data set includes 6-methyluracil, 9-methyladenine, adenine, albendazole, caffeine, clonidine, cytosine, diphenylamine, fluconazole, fluorouracil, guanine, imidazole, metronidazole, phenytoin, pyridine, pyridimidine, pyrrole, thymine, and uracil (Figure S1).

Next, a set of 27 compounds was used to calibrate the suitability of the MST model to compute log P_I values. This set included 18 cations [2-(2-pyridyl)-ethylammonium, 2-phenylethylammonium, 3-carboxyanilinium, 4-carboxyanilinium, 4-methyl-N-ethylbenzylammonium, 4-methyl-N-heptylbenzylammonium, 4-methyl-N-pentylbenzylammonium, 4-phenylbutylammonium, *N*-acetyl-*L*-histidine amide, *N*-acetyl-*L*-lysine amide, amitriptyline, desipramine,

Table 1. Calculated and Experimental n-Octanol/Water Partition Coefficient (log P_N) for the Series of Neutral Nitrogen-Containing Aromatic Compounds Used in the Refinement of ξ_N and ξ_{NH} Atomic Surface Tensions for n-Octanol

compound[a]	computed log P_N (original)	computed log P_N (refined)	exptl[b]
6-methyluracil (**1**)	−1.7	−0.5	−1.2
9-methyladenine (**2**)	−2.9	−0.3	0.0
adenine (**3**)	−4.1	−1.1	−0.1
albendazole (**4**)	2.0	3.7	2.7
caffeine (**5**)	−0.2	0.9	−0.1
clonidine (**6**)	1.1	2.8	1.6
cytosine (**7**)	−4.3	−2.2	−1.7
diphenylamine (**8**)	3.1	3.7	3.5
fluconazole (**9**)	−1.2	1.1	0.4
fluorouracil (**10**)	−2.2	−0.9	−0.9
guanine (**11**)	−5.9	−2.7	−0.9
imidazole (**12**)	−2.2	−1.1	−0.1
metronidazole (**13**)	−0.9	0.0	0.0
phenytoin (**14**)	2.0	3.2	2.5
pyridine (**15**)	0.4	0.9	0.7
pyrimidine (**16**)	−0.8	0.1	−0.4
pyrrole (**17**)	−0.2	0.5	0.8
thymine (**18**)	−1.8	−0.5	−0.6
uracil (**19**)	−2.3	−1.1	−1.1
mse[c]	1.4	−0.1	
mue[c]	1.4	0.6	
rmsd[c]	1.9	0.8	

[a]See Figure S1. [b]Ref 62. [c]Mean signed error (mse), mean unsigned error (mue), and root-mean square deviation (rmsd) calculated relative to the experimental values are given in log P units.
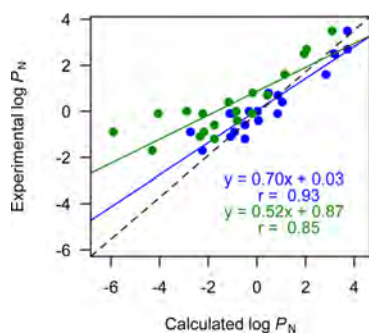

Figure 1. Comparison between experimental and calculated n-octanol/water log P_N for the series of neutral nitrogen-containing aromatic compounds. Calculated values determined from IEFPCM/MST calculations using the original parametrization of the IEF-MST method (green) and the refined atomic surface tension for N- and NH atom types (blue).

imipramine, lidocaine, tetrabutylammonium, tetraethylammonium, tetramethylammonium, tetrapentylammonium] and 9 anions (2-4-dichlorophenoxyacetate, 5-phenylvalerate, N-acetyl-L-aspartic acid amide, N-acetyl-L-glutamic acid amide, diclofenac, ibuprofen, indomethacin, naproxen, and pentachlorophenolate; see Figure S2).

Finally, a set of 35 compounds was used to calibrate the behavior of the refined MST model for predicting the partition coefficient of the neutral compound (P_N), and the distribution coefficient at pH 7.4 ($D_{7.4}$), taking advantage of the available experimental data for pK_a, log P_N, and log $D_{7.4}$. This data set

Table 2. Experimental and Computed log P_I Values for the Set of 27 Ionic Compounds Used in the Refinement of the MST Method

compound[a]	computed log P_I (original)	computed log P_I (refined)	exptl
2-2-pyridyl-ethylammonium	−1.7	−0.4	−2.3[b]
2-phenethylammonium	−7.1	0.4	−1.6[b]
2-4-dichlorophenoxyacetate	−5.5	−0.7	−0.9[c]
3-carboxyanilinium	−9.2	−1.6	−0.9[c]
4-carboxyanilinium	−9.4	−1.6	−0.4[c]
4-methyl-N-ethylbenzylammonium	−6.4	−1.9	−0.8[d]
4-methyl-N-heptylbenzylammonium	−3.6	0.1	2.1[d]
4-methyl-N-pentylbenzylammonium	−4.7	−0.9	0.8[d]
4-phenylbutylammonium	−6.1	1.5	0.7[e]
5-phenylvalerate	−6.1	−0.4	−1.0[c]
N-acetyl-L-aspartic amide[f]	−6.8	−2.8	−2.6[g]
N-acetyl-L-glutamic amide[f]	−6.2	−3.4	−2.5[g]
N-acetyl-L-histidine amide[h]	−7.9	−1.7	−3.4[g]
N-acetyl-L-lysine amide[f]	−7.7	−1.8	−2.8[g]
amitriptyline	−1.1	1.3	0.2[c]
desipramine	−2.7	0.0	0.3[c]
diclofenac	−4.3	0.7	0.7[c]
ibuprofen	−5.9	−0.7	−0.2[c]
imipramine	−0.6	1.8	0.5[c]
indomethacin	−2.7	2.1	0.6[c]
lidocaine	−2.6	−0.7	−0.5[c]
naproxen	−5.5	−0.6	−0.2[c]
pentachlorophenol	−1.9	1.8	1.3[i]
tetrabutylammonium	1.2	3.3	2.3[j]
tetraethylammonium	−2.8	−0.8	−0.9[j]
tetramethylammonium	−5.6	−2.7	−2.0[j]
tetrapentylammonium	3.0	5.1	3.8[j]
mse[k]	4.1	−0.2	
mue[k]	4.1	0.9	
rmsd[k]	4.7	1.1	

[a]See Figure S2. [b]Ref 63. [c]Ref 39. [d]Ref 25. [e]Ref 64. [f]Values derived from log $D_{7.4}$ data reported in ref66, assuming full ionization of the compounds at physiological pH. [g]Refs 65. and66. [h]Estimated from additive scheme (see Supporting Information). [i]Ref 24. [j]Ref 67. [k]Mean signed error (mse), mean unsigned error (mue), and root-mean square deviation (rmsd) calculated relative to the experimental values are given in log P units.

includes 31 drugs (albendazole, amitriptyline, antipyrine, bumetanide, caffeine, clormipramine, clonidine, clozapine, cyclobenzaprine, desipramine, diazepam, diclofenac, diphenylamine, estradiol, fluconazole, flurbiprofen, ibuprofen, imipramine, indomethacin, lidocaine, loratadine, maleic acid, metoclopromaide, metronidazole, mezoridazine, naproxen, paracetamol, pentachlorophenol, pentoxifylline, phenytoin, and triflupromazine; see Figure S3) and 4 amino acids (aspartic acid, glutamic acid, histidine, and lysine, capped with acetyl and amide groups at the N- and C-terminus). Table S1 reports the experimental values for pK_a, log P_N, and log $D_{7.4}$ for all these molecules.[39,54,55]

Finally, the refined model was checked by computing the pH-dependent distribution profiles in the framework of the different partitioning schemes discussed above. To this end, several drugs (ibuprofen, imipramine, desipramine, pentachlorophenol,
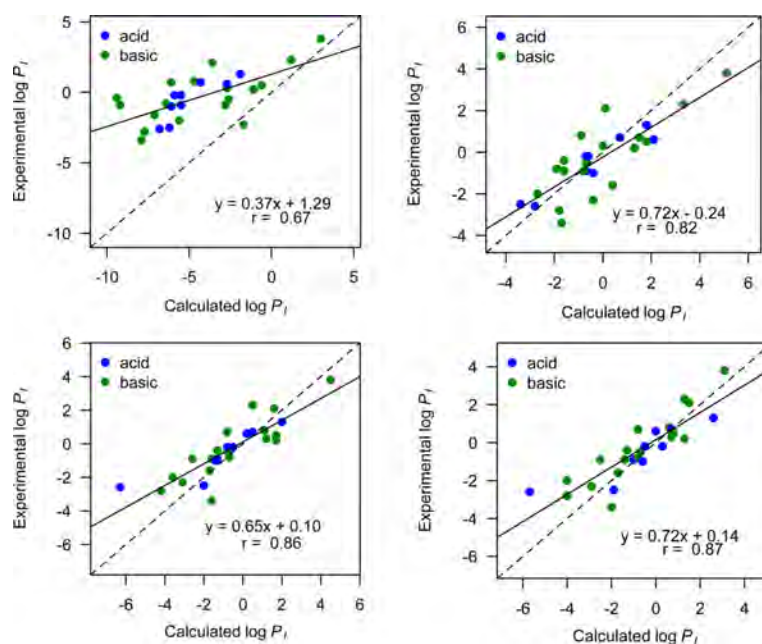
170

Figure 2. Comparison between experimental and calculated log P_I values for the series of 27 ionic compounds. Acidic and basic compounds are shown in blue and green, respectively. Calculated values were determined from IEFPCM/MST computations using the original (top left) and refined (top right) parameters, as well as ACD/I-Lab (bottom left) and ChemAxon (bottom right).

lidocaine, amitriptyline, and warfarin) and four amino acid analogues (*N*-acetyl-*L*-aspartic acid amide, *N*-acetyl-*L*-lysine amide, *N*-acetyl-*L*-glutamic acid amide, and *N*-acetyl-*L*-histidine amide) were considered taking advantage of the available experimental data for these compounds.

The molecular geometries of the compounds were fully optimized at the B3LYP/6-31G(d) level of theory in the corresponding solvent phase, water or *n*-octanol, using the IEFPCM version of the MST model. Then, single-point calculations in the gas phase and in solution were performed in order to estimate the free energy of solvation in water and *n*-octanol. All calculations were performed in Gaussian 09.[56] For the set of ionizable compounds, computations were performed for the minimum structure obtained after geometry optimization of an extended conformation of the molecule. This strategy was motivated by the generally low number of rotatable bonds present in these compounds (see Figures S2 and S3), as well as by the similar *n*-octanol/water transfer free energies obtained from a single-conformation approach and from conformational ensembles for druglike compounds in a previous study.[57] Nevertheless, for the subset of *N*-acetyl-*L*-amino acid amides, calculations were performed taking into account all possible rotamers with a probability contribution higher than 5% to the total conformational space as given by the backbone-dependent conformational library reported by Dunbrack and Karplus.[58,59]

Calculation of log D was accomplished using eqs 2, 4, and 9 using experimental pK_a values (Table S1). Application of eq 9 was performed paying particular attention to the counterion used for computation of the distribution coefficient, maintaining consistency with the experimental procedure reported in the original works (see below). Specifically, data for the partition of inorganic ions (log $P_{I,Cl^-}^{\circ} = -4.5$, log $P_{I,Na^+}^{\circ} = -2.5$ and log $P_{I,K^+}^{\circ} = -2.6$) required for the application of eq 9 to acidic and basic compounds were taken from the literature.[25,60,61]

RESULTS

Refinement of the MST Model for Solvation in *n*-Octanol. One of the initial aims of this study was to refine the parametrization of the MST model for nitrogen-containing aromatic compounds (Figure S1), as they are key structural elements in many biologically relevant molecules and drugs but were poorly represented in the data set of compounds considered in the original B3LYP/6-31G(d) parametrization of the IEFPCM/MST model. Indeed, preliminary calculations performed for a subset of 12 heterocyclic organic compounds (2−7, 9, 11−13, 15, and 16; see Table 1 and Figure S1) revealed the need to adjust the surface tension of the pyridine-like nitrogen atom for solvation in *n*-octanol. Thus, the original atomic surface tension assigned to the N-type atom ($\xi_N = -0.115$ kcal mol^{-1} Å^{-2}) was found to underestimate the solvation free energy in *n*-octanol, and a better agreement with experimental data was achieved upon adjustment to a surface tension of -0.161 kcal mol^{-1} Å^{-2}, which was therefore adopted in the refined version. Additional analyses were performed to check the surface tension for the pyrrole-like nitrogen atom (NH-type), even though in this case adjustment of the original surface tension ($\xi_{NH} = -0.234$ kcal mol^{-1} Å^{-2}) to -0.295 kcal mol^{-1} Å^{-2} was found to have a lower effect on the solvation free energy in *n*-octanol.

The effect of these refinements is shown in Table 1, which reports the solvation free energies determined with the original and refined parameters, as well as the experimental data[62] for the set of compounds. The adjustment of the surface tension of these two atoms types sufficed to improve significantly the ability of the IEFPCM/MST model for predicting the log P_N values of these compounds. This is noted in the reduction of the root-mean square deviation (rmsd) from 1.9 (log P units) in the original parametrization to 0.8 for the refined version (Table 1), as well as in the comparison between experimental and calculated log P_N values, as the refined surface tensions (ξ_N and ξ_{NH}) improve the

171

Table 3. Calculated and Experimental n-Octanol/Water Partition Coefficient ($\log P_N$) for the Set of 35 Small Molecules

compound	computed	exptl[a]
albendazole	3.7	2.7
amitriptyline	6.5	4.9
antipyrine	2.0	0.4
bumetanide	2.6	2.6
caffeine	0.9	−0.1
clomipramine	6.7	5.2
clonidine	2.8	1.6
clozapine	5.5	3.2
cyclobenzaprine	6.3	5.2
desipramine	5.7	4.9
diazepam	4.5	2.8
diclofenac	5.6	4.5
diphenylamine	3.7	3.5
estradiol	4.2	3.7
fluconazole	1.1	0.4
fulbipronen	4.2	4.2
ibuprofen	3.2	4.0
imipramine	5.9	4.8
indomethacin	4.9	4.3
lidocaine	2.8	2.4
loratadine	7.4	5.2
maleic acid	−1.5	−0.5
metoclopramide	2.2	2.6
metronidazole	0.0	0.0
mezoridazine	6.5	3.9
N-acetyl-L-aspartic amide[b]	−2.3	−2.0
N-acetyl-L-glutamic amide[b]	−1.5	−1.9
N-acetyl-L-histidine amide[b]	−0.9	−1.9
N-acetyl-L-lysine amide[b]	−0.4	−0.8
naproxen	2.7	3.2
paracetamol	−0.1	0.5
pentachlorophenol	3.8	5.0
pentoxifylline	1.6	0.3
phenytoin	2.0	2.5
triflupromazine	6.6	5.5
mse[c]	−0.6	
mue[c]	0.9	
rmsd[c]	1.1	

[a]See Table S1. [b]Estimated from additive scheme (see the Supporting Information). [c]Mean signed error (mse), mean unsigned error (mue), and root-mean square deviation (rmsd) calculated relative to the experimental values are given in $\log P$ units.

regression correlation with the experimental values from 0.85 to 0.93 (see Figure 1).

Calibration of the MST Model for Ionic Compounds in n-Octanol. In the MST model the electrostatic contribution to the hydration free energy of charged species is determined by reducing the solvent-exposed cavity of the charged atoms by a factor close to 10%.[53] While this strategy proved to be valuable for calculating the solvation of univalent ionic species in water, its suitability in other solvents has never been checked. Therefore, for our purposes here, it is necessary to calibrate the suitability of this strategy for the solvation of ionic compounds in n-octanol. To this end, calculations were performed for a set of 27 compounds (see Table 2 and Figure S2), including 9 anions and 18 cations, taking advantage of the availability of partition coefficients for these charged species.[23,24,38,63−67]
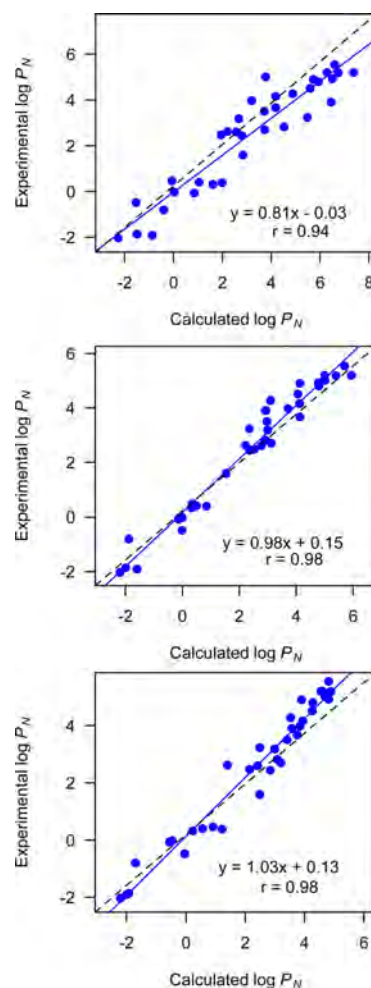

Figure 3. Comparison between experimental and calculated $\log P_N$ for a set of 35 small molecules. Computed values were determined by using the refined IEFPCM/MST calculations (top) and by using ACD/I-Lab $\log D$ (middle) and ChemAxon (bottom).

Comparison of the calculated and experimental $\log P_I$ values determined for these compounds suggested that the optimal scaling factor, λ, for solvation in n-octanol must be reduced by around 19%, which implies that the scaling factor used for neutral compounds ($\lambda = 1.50$) must be close to 1.20 for charged chemical groups. This adjustment enhances the contribution of the electrostatic component to the solvation free energy for charged compounds, following the trends reported for the hydration of monovalent ions,[53] an effect interpreted from the balance between the gain in solvent−solute stabilization energy triggered by the solute's electron density redistribution upon solvation and the energy cost associated with distortion of the electron density by the solvent reaction field.[50−52] Because of the formal simplicity of this correction, the suitability of the atomic surface tension was further checked. In the case of cations with a localized charge on the sp^3 nitrogen atom, it was necessary to enlarge the surface tension of the nitrogen atom (NH atom type) by a factor of 17% ($\xi_{NH} = -0.274$ kcal mol^{-1} Å^{-2}). This enlargement was also extended to the methylene/methyl groups bound to the protonated nitrogen atom ($\xi_{CHx} = -0.227$ kcal mol^{-1} Å^{-2}), which may be related to the inductive effect noted in the increased

Table 4. Experimental and Calculated Distribution Coefficients (log $D_{7.4}$) Determined for the Set of 35 Ionizable Compounds Used to Calibrate the IEFPCM/MST Model[a]

compound	eq 2	eq 4	eq 9	exptl[b]
albendazole	3.7	3.7	3.7	3.3
amitriptyline	4.5	4.5	4.5	2.8
antipyrine	2.0	2.0	2.0	0.2
bumetanide	−1.2	−0.2	−1.0	−0.1
caffeine	0.8	0.8	0.8	0.0
clomipramine	4.7	4.7	4.7	3.3
clonidine	2.1	2.1	2.1	0.6
clozapine	5.1	5.1	5.1	3.0
cyclobenzaprine	5.2	5.2	5.2	2.9
desipramine	2.7	2.7	2.7	1.4
diazepam	4.5	4.5	4.5	2.7
diclofenac	2.3	2.4	2.3	1.1
diphenylamine	3.7	3.7	3.7	3.4
estradiol	4.2	4.2	4.2	4.0
fluconazole	1.1	1.1	1.1	0.5
fulbipronen	1.0	2.2	1.0	0.9
ibuprofen	0.7	0.7	0.7	1.3
imipramine	3.9	3.9	3.9	2.5
indomethacin	2.0	2.4	2.0	0.8
lidocaine	2.1	2.1	2.1	1.6
loratadine	7.4	7.4	7.4	4.4
maleic acid	−7.1	−4.5	−3.5	−5.0
metoclopramide	0.3	0.3	0.3	0.5
metronidazole	0.0	0.0	0.0	−0.1
mezoridazine	5.0	5.0	5.0	1.8
N-acetyl-L-aspartic amide	−5.8	−2.8	−2.6	−2.6
N-acetyl-L-glutamic amide	−4.7	−3.4	−2.9	−2.5
N-acetyl-L-histidine amide	−1.0	−0.9	−1.0	−3.5
N-acetyl-L-lysine amide	−1.8	−1.8	−3.1	−2.8
naproxen	−0.6	−0.3	−0.5	0.3
paracetamol	−0.1	−0.1	−0.1	0.3
pentachlorophenol	1.2	1.9	1.2	2.5
pentoxifylline	1.6	1.6	1.6	0.3
phenytoin	1.9	1.9	1.9	2.2
triflupromazine	4.6	4.6	4.6	3.4
mse[c]	−0.6	−0.9	−0.8	
mue[c]	1.3	1.1	1.1	
rmsd[c]	1.6	1.4	1.4	

[a]Calculated values were obtained by combining log P_N and log P_I values using eqs 2, 4, and 9. [b]See Table S1. [c]Mean signed error (mse), mean unsigned error (mue), and root-mean square deviation (rmsd) calculated relative to the experimental values are given in log P units.

chemical shift observed in ^{1}H NMR studies (see Table S2).[68] This effect is known to be less important for the carbon atoms bound to the groups with delocalized charges (i.e., carboxylate anions; see Table S2), where no further adjustment was needed.

The log P_I values calculated for the whole set of ionic organic compounds using the new parameters are presented in Table 2, which also collects the experimental data. The mean signed error was reduced from 4.1 to −0.2 (log P_I units), and the rmsd was decreased from 4.7 to 1.1 (log P_I units) after implementation of the preceding adjustments in the MST model. The difference between calculated and experimental values may, at least in part, reflect the variance in the experimental data, as noted for diclofenac, because the experimental data may vary between 0.45 and 0.8 depending on the experimental conditions used to estimate the partition coefficient.[36] Furthermore, Figure 2 shows the improved correlation between the refined log P_I values and
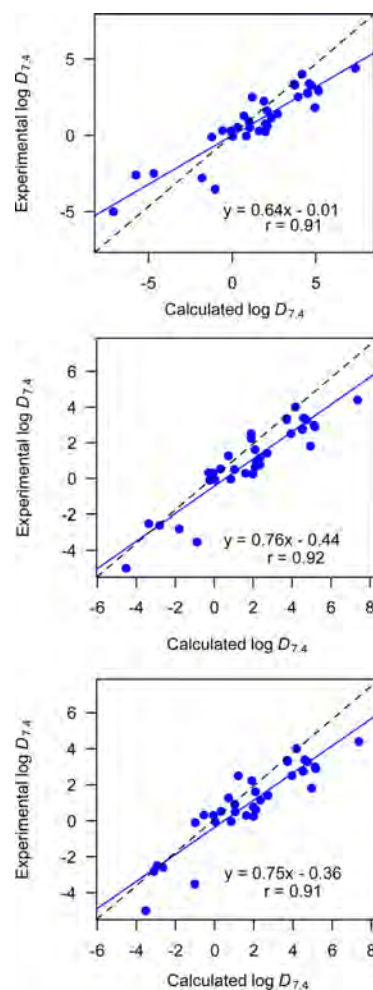

Figure 4. Comparison between experimental (determined by the shake-flask method) and calculated log $D_{7.4}$ for 35 ionizable small molecules. Computed values were derived from IEFPCM/MST (blue) calculations using eqs 2 (top), 4 (middle), and 9 (bottom).

Table 5. Statistical Parameters of the Comparison between Experimental and Calculated log $D_{7.4}$ Values for the Series of 35 Small Molecules

method	mse	mue	rmsd	r
ACD/I-Lab	0.0	0.5	0.8	0.95
ChemAxon	0.2	0.5	0.8	0.95
IEFPCM/MST, eq 2	−0.6	1.3	1.6	0.91
IEFPCM/MST, eq 2 (exptl log P_N)	0.1	0.5	0.8	0.96
IEFPCM/MST, eq 4	−0.9	1.1	1.4	0.92
IEFPCM/MST, eq 4 (exptl log P_N)	−0.3	0.4	0.6	0.96
IEFPCM/MST, eq 9	−0.8	1.1	1.4	0.91
IEFPCM/MST, eq 9 (exptl log P_N)	−0.2	0.4	0.5	0.97

the experimental ones, which corrected the systematic tendency to overestimate the hydrophilicity of the charged compounds in the original parametrization of the IEFPCM/MST method. For the sake of comparison, it is worth noting that the refined log P_I values are in agreement with the behavior observed for the values estimated by using empirical methods, such as ACD/I-Lab[69] and ChemAxon[70] methods (see Figure 2).
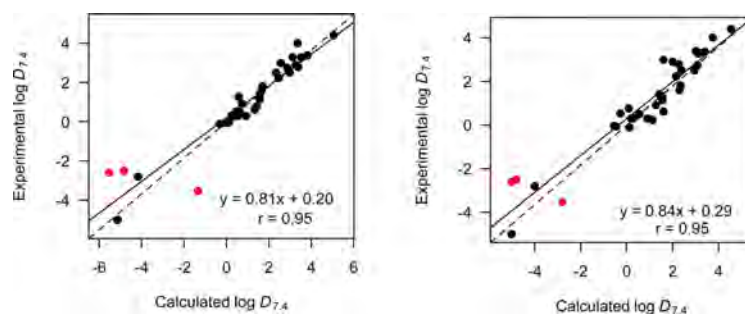
Figure 5. Comparison between experimental and calculated $\log D_{7.4}$ for the set of 35 small molecules. Values were obtained using (left) ACD/I-Lab and (right) ChemAxon. *N*-Acetyl-*L*-aspartic acid amide, *N*-acetyl-*L*-glutamic acid amide, and *N*-acetyl-*L*-histidine amide are shown as red dots.

Estimation of log $D_{7.4}$. The $\log D_{7.4}$ values compiled for a set of 35 ionizable small molecules (see Figure S3 and Table S1) were used as a test set to calibrate the suitability of the adjustments introduced in the refined IEFPCM/MST model. These compounds encompass a broad range of chemical diversity in selected physicochemical properties (see Figure S4), such as molecular weight (up to 400 Da), number of rotatable bonds (up to 8), number of aromatic rings (up to 3), and number of hydrogen bond donors (up to 5) and acceptors (up to 7). In order to reproduce the experimental distribution coefficients within the framework of the partition formalisms represented by eqs 2, 4, and 9, $\log P_N$ and $\log P_I$ values were estimated from MST calculations, whereas the pK_a of these compounds and the partition coefficient of the counterion were taken from experimental data (see Theory and Computational Details and Table S1).

We first evaluated the capacity of the refined MST model for predicting the experimental $\log P_N$ of these compounds. This comparison is shown in Table 3 and Figure 3. The rmsd between experimental and calculated values is 1.1 ($\log P_N$ units), and the calculated values exhibit a good correlation with the experimental ones ($r = 0.94$). Furthermore, these trends compare well with the values predicted by using empirical methods (ACD/I-lab or ChemAxon; Figure 3), although they exhibit a slightly better correlation with the experimental data, which likely reflects the most extensive parametrization of fragmental contributions that lies behind these methods.[71,72]

Because the distribution coefficient takes into account the partition of both neutral and ionic species of ionizable compounds, it provides an indirect approach to test the reliability of the calculated P_I values. This is more challenging, because the measured $\log D$ may be affected by the experimental conditions, such as the nature of the background salt and the concentration of the solution, which would influence the potential difference between the two phases.[73] Moreover, different formalisms have been proposed to combine $\log P_N$, $\log P_I$, and pK_a to estimate $\log D$. Accordingly, $\log D_{7.4}$ was determined using the $\log P_N$ and $\log P_I$ values determined from IEFPCM/MST computations and was combined with experimental pK_a values reported for the set of compounds (see Table S1) following Schemes 2 (eq 2), 3 (eq 4), and 4 (eq 9).

Table 4 reports the $\log D_{7.4}$ values obtained from IEFPCM/ MST calculations. In general, there is a slight tendency to overestimate the hydrophobicity of the compounds, as noted in the mean signed error (mse) of ~-0.7 ($\log D$ units) found for eqs 2, 4, and 9, while the rmsd amounts to ~1.5 ($\log D$ units). The performance of the IEFPCM/MST model is similar for the three formalisms examined in this study, with a slightly larger rmsd when eq 2 is used. Similar regression equations between

calculated and experimental $\log D_{7.4}$ values are also found ($r \approx 0.92$), although Figure 4 shows a slightly larger deviation from the perfect linear regression for the values obtained with eq 2. For this latter formalism the largest errors (given in $\log D$ units in parentheses) are found for the subset of amino acid analogues *N*-acetyl-*L*-aspartic acid amide ($+3.2$), *N*-acetyl-*L*-glutamic acid amide ($+2.2$), and *N*-acetyl-*L*-histidine amide ($+2.5$), as well as for mesoridazine (-3.1), loratadine (-3.0), cyclobenzaprine (-2.3), clozapine (-2.1), and maleic acid ($+2.1$). Nevertheless, when the partition of the ionic species is taken into account (eqs 4 and 9), the deviation found for aspartic and glutamic analogues and for maleic acid is largely reduced, whereas the value predicted for acetyl-*L*-histidine amide remains unaffected.

To further check the reliability of the $\log P_I$ values, the calculated $\log P_N$ values were replaced by the experimental ones, and $\log D_{7.4}$ was determined using eqs 2, 4, and 9. The distribution coefficients obtained by limiting the IEFPCM/MST calculation to the partition coefficient of the ionized species ($\log P_I$) reduces the rmsd between predicted and experimental data to ~-0.8 ($\log D$ units), and the correlation coefficient is increased to 0.96 (Table 5). Note that these statistical parameters compare well with the values estimated using empirical methods (ACD/I-Lab, ChemAxon), as noted in Table 5 and Figure 5. Overall, these results give confidence to the partition values of ionic species determined with the refined IEFPCM/MST method, especially taking into account the limited extension of the model refinement, and the single-conformation approach adopted in present calculations. Nevertheless, the use of representative conformational ensembles may be required to obtain more accurate estimates of $\log D_{7.4}$ in flexible molecules able to form distinct patterns of intramolecular interactions.[74−76]

pH-Dependent Hydrophobicity Profiles. While the preceding results support the refined IEFPCM/MST method, there is generally little difference between the distinct formalisms (eqs 2, 4, and 9) followed for calculation of $\log D_{7.4}$. This may reflect the fact that all molecules are approved drugs with high $\log P_N$ values and that $\log D$ was calculated at physiological pH, while the contribution of ionic species may be expected to be more relevant at extreme pH values. Hence, we decided to determine the lipophilicity profile of seven drugs and four amino acid analogues between pH 2 and 12, taking advantage of the experimental data about the pH-dependent partitioning of these compounds.[23,24,39,65,66]

For ibuprofen, warfarin, and pentachlorophenol, the three formalisms give similar $\log P_N$ values at low pH, where the neutral species predominates (Figure 6). However, the profiles diverge at intermediate pH values (between 6 and 8), following the increased population of the anionic species. Equation 2, which
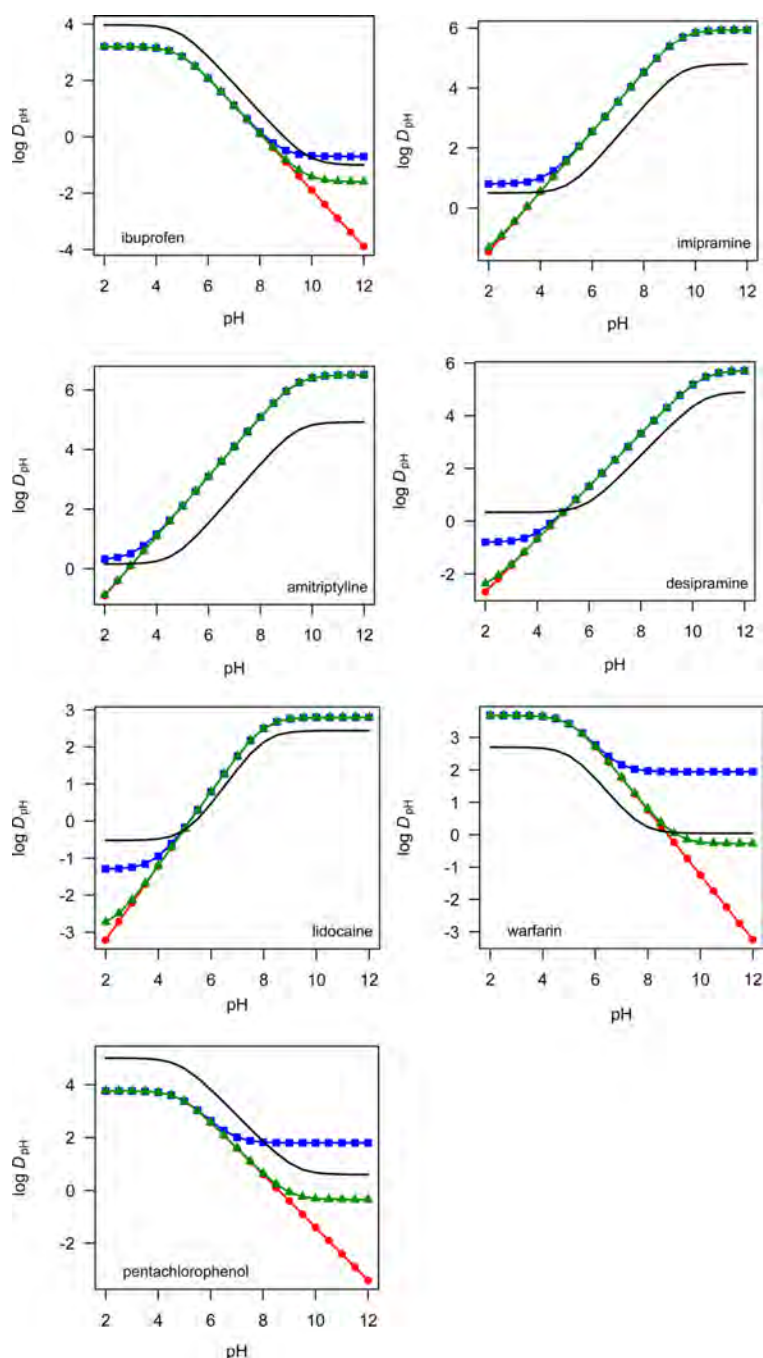
Figure 6. Comparison of *n*-octanol/water distribution coefficient profiles of selected drugs using eqs 2 (red), 4 (blue), and 9 (green). The experimental data are shown in black.

does not take into account the partition of the ionic species, gives rise to a profile that decreases steadily with increasing pH. In contrast, eqs 4 and 9 show an asymptotic behavior at basic pH. For imipramine, amitriptyline, desipramine, and lidocaine all the methods exhibit the same log D at pH $\sim$ 10, which arises from the partition of the neutral species. The pH-dependence of the profiles is similar up to acidic solutions (pH < 5), where the contribution of the cationic species is more important. Again, eq 2 shows a continuous decrease in log D with decreasing

pH, whereas the profiles obtained from eqs 4 and 9 show the appearance of an asymptotic behavior at low pH. Note, however, that the appearance of this asymptotic value occurs at lower pH values for eq 9, leading generally to a larger deviation with regard to the experimental profile compared with the results obtained from eq 4, which reproduces well the general trends of the experimental sigmoidal profile. On the other hand, it is worth noting that the IEFPC/MST profiles obtained with eq 4 compare well with the pH-dependent profiles obtained
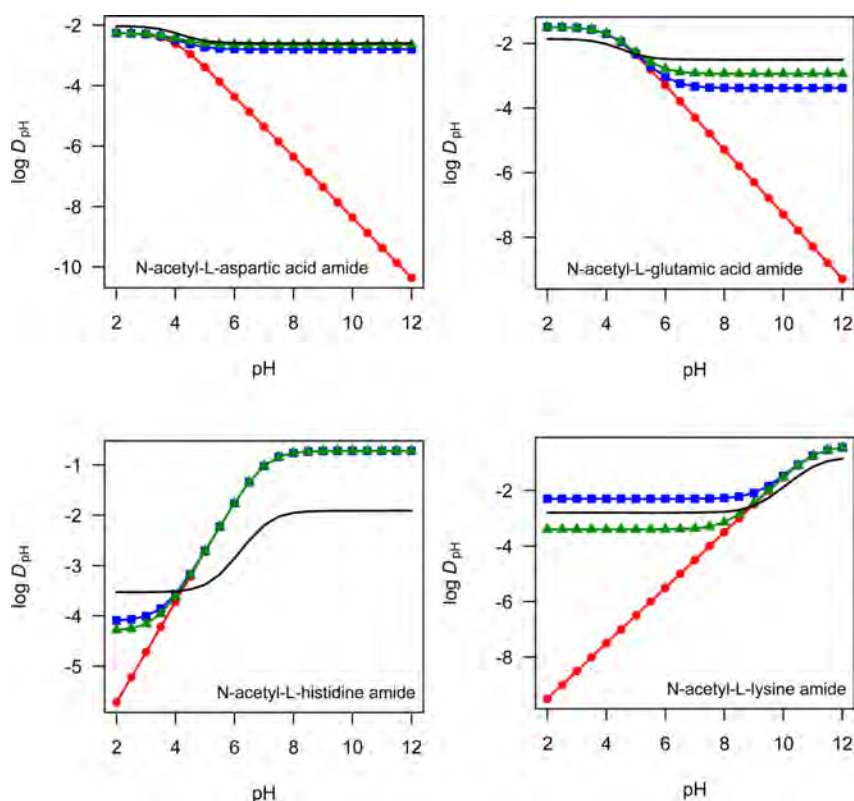
175

Figure 7. Comparison of *n*-octanol/water distribution coefficient profiles of selected amino acid analogues using eqs 2 (red), 4 (blue), and 9 (green). The experimental data are shown in black.

from empirical methods (ACD/I-Lab and ChemAxon; see Figure S5).

The pH-dependent hydrophobicity profile determined for the set of amino acid analogues is shown in Figure 7. As noted above, eq 2 works worse at extreme pH, as expected because of the neglect of the contribution arising from the partition of ionic species. On the other hand, eqs 4 and 9 give similar profiles that reproduce the experimental values for the whole range of pH values. Furthermore, for the analogues of aspartic and glutamic acids, and to less extent for lysine, the IEFPCM/MST results derived by using eqs 4 and 9 improve the pH-dependent profiles obtained from empirical methods, which predict a much higher hydrophilic behavior for pH values larger than 5 and lower than 9 for aspartic/glutamic acid and lysine, respectively (compare Figures 7 and S6).

Overall, the results support the suitability of eq 4 for estimating the pH-dependence of the distribution profiles of ionizable compounds. The limited success found for eq 9 is surprising, especially when one takes into account the results obtained for amino acid analogues, but it may reflect the marked influence of inorganic ions on the experimental measurements of the distribution coefficient of ionized compounds.[36,44,59,61,77] In general, shake-flask experiments are performed in wet *n*-octanol/water systems using 0.15 M KCl or NaCl, and reliable values for the inorganic standard partition coefficients have been reported.[25,60,61] However, it is also known that the distribution coefficient can be expected to increase substantially when a more hydrophobic cation is added in excess to the system, as has been reported for both ibuprofen and pentachlorophenol.[23,24,39]

Finally, the potential contribution due to the formation of ion pairs may also have a significant effect on the distribution coefficient of ionized compounds, especially when the salt concentration is large enough relative to the ionized compound.[35,36]

■ CONCLUSION

Predicting the pH dependence of the partition of organic compounds between *n*-octanol and water is extremely important for gaining insight into the behavior of bioactive compounds. A fundamental property to attain this goal is the distribution coefficient *D*, which encompasses the differential partition of both neutral and ionic species present in the two solvents. This makes it necessary to resort to physicochemical formalisms that take into account species-specific lipophilicities.[73] In this context, this study has examined the refinement of the IEFPCM/MST model, with special emphasis on the adjustment of specific parameters required for the solvation of ionic compounds in *n*-octanol, following the previous implementation reported for aqueous solution.[32,34,53]

The results point out that reduction of the solvent-excluded cavity used for the electrostatic term affords the major correction to the calculation of the solvation free energy in *n*-octanol. In conjunction with adjustments in specific atomic surface tensions, the refined IEFPCM/MST permits us to quantitate the contribution of the neutral and ionic forms of ionizable compounds and to estimate distribution coefficients that compare well with experimental values. For the set of compounds used in this study, the results also show that there are little differences between the distribution coefficients at physiological pH (log $D_{7.4}$) determined

with the three formalisms examined in this work. Nevertheless, inspection of the pH-dependent hydrophobicity profiles reveals the need to take into account the contribution played by the partition of the ionic species in order to reproduce the experimental data, as these formalisms exhibit the correct asymptotic behavior at extreme pH values, where the ionized species should predominate. Finally, present results suggest that eq 4 is the minimal scheme required to rationalize the pH-dependent distribution profile of ionizable compounds, including acidic and basic compounds as well as amino acid analogues. The role of the Galvani potential difference between the two phases, however, may be relevant at higher concentrations of the background salt.

Although the balance between accuracy and computational cost is more favorable for empirical methods, the availability of refined versions of QM-based continuum solvation methods opens the way to the analysis of factors implicated in the partition of (bio)organic molecules in complex chemical systems. In particular, future studies will address the influence played by conformational flexibility and the formation of ionic pairs with the aim to calibrate the reliability of more elaborate formalisms of drug partitioning.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: campanera@ub.edu.
*E-mail: fjluque@ub.edu.

ORCID ⓘ

Carles Curutchet: 0000-0002-0070-1208
Josep M. Campanera: 0000-0002-6698-874X
F. Javier Luque: 0000-0002-8049-3567

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Leo, A. J. Calculating log P_{oct} from Structures. *Chem. Rev.* **1993**, 93, 1281−1306.

(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, 23, 3−25.

(3) Sangster, J. *Octanol−Water Partition Coefficients: Fundamentals and Physical Chemistry*; Wiley: Chichester, 1997.

(4) Eugene Kellog, G.; Abraham, D. J. Hydrophobicity: Is LogP$_{o/w}$ More than the Sum of its Parts? *Eur. J. Med. Chem.* **2000**, 35, 651−661.

(5) Bhal, S. K.; Kassam, K.; Peirson, I. G.; Pearl, G. M. The Rule of Five Revisited: Applying log D in Place of log P in Drug-Likeness Filters. *Mol. Pharmaceutics* **2007**, 4, 556−560.

(6) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of LogP Methods on More Than 96,000 Compounds. *J. Pharm. Sci.* **2009**, 98, 861−893.

(7) Takács-Novák, K. In *Physico-Chemical Methods in Drug Discovery and Development*; Mandic, Z., Ed.; IAPC Publishing: Zagreb, 2012; pp 1−52.

(8) Muñoz-Muriedas, J. In *Physico-Chemical and Computational Approaches to Drug Discovery*; RSC Drug Discovery no. 23; Luque, F. J., Barril, X., Eds.; The Royal Society of Chemistry: Cambridge, 2012; pp 104−127.

(9) Cramer, C. J.; Truhlar, D. G. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* **1999**, 99, 2161−2200.

(10) Javier Luque, F.; Curutchet, C.; Muñoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpi, J. L.; Orozco, M. Continuum Solvation Models: Dissecting the Free Energy of Solvation. *Phys. Chem. Chem. Phys.* **2003**, 5, 3827−3836.

(11) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, 105, 2999−3094.

(12) Klamt, A. The COSMO and COSMO-RS Solvation Models. *WIRES Comput. Mol. Sci.* **2011**, 1 (5), 699−709.

(13) Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, 17, 6174−6191.

(14) Cramer, C. J.; Truhlar, D. G. A Universal Approach to Solvation Modeling. *Acc. Chem. Res.* **2008**, 41, 760−768.

(15) Klamt, A.; Mennucci, B.; Tomasi, J.; Barone, V.; Curutchet, C.; Orozco, M.; Luque, F. J. On the Performance of Continuum Solvation Methods. A Comment on "Universal Approaches to Solvation Modeling". *Acc. Chem. Res.* **2009**, 42, 489−492.

(16) Kah, M.; Brown, C. D. Log D: Lipophilicity for Ionizable Compounds. *Chemosphere* **2008**, 72, 1401−1408.

(17) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 868−873.

(18) Walters, W. P.; Murcko, M. A. Prediction of "Drug-Likeness". *Adv. Drug Delivery Rev.* **2002**, 54, 255−271.

(19) Manallack, D. The pK$_a$ Distribution of Drugs: Application to Drug Discovery. *Perspect. Medicin. Chem.* **2007**, 1, 25−38.

(20) Ursu, O.; Rayan, A.; Goldblum, A.; Oprea, T. I. Understanding Drug-Likeness. *WIRES Comput. Mol. Sci.* **2011**, 1, 760−781.

(21) Comer, J.; Tam, K. In *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; Verlag Helvetica Chimica Acta: Zurich, 2001; pp 275−304.

(22) Avdeef, A. Physicochemical Profiling (Solubility, Permeability and Charge State). *Curr. Top. Med. Chem.* **2001**, 1, 277−351.

(23) Westall, J. C.; Leuenberger, C.; Schwarzenbach, R. P. Influence of pH and Ionic Strength on the Aqueous-Nonaqueous Distribution of Chlorinated Phenols. *Environ. Sci. Technol.* **1985**, 19, 193−198.

(24) Jafvert, C. T.; Westall, J. C.; Grieder, E.; Schwarzenbach, R. P. Distribution of Hydrophobic Ionogenic Organic Compounds Between Octanol and Water: Organic Acids. *Environ. Sci. Technol.* **1990**, 24, 1795−1803.

(25) Abraham, M. H.; Acree, W. E., Jr. The Transfer of Neutral Molecules, Ions and Ionic Species from Water to Wet Octanol. *Phys. Chem. Chem. Phys.* **2010**, 12, 13182−13188.

(26) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R., Jr. The Proton's Absolute Aqueous Enthalpy and Gibbs Free Energy of Solvation from Cluster-Ion Solvation Data. *J. Phys. Chem. A* **1998**, 102, 7787−7794.

(27) Pliego, J. R., Jr.; Riveros, J. M. Gibbs Energy of Solvation of Organic Ions in Aqueous and Sulfoxide Solutions. *Phys. Chem. Chem. Phys.* **2002**, *4*, 1622−1627.

(28) Pliego, J. R., Jr.; Miguel, E. L. M. Absolute Single-Ion Solvation Free Energy Scale in Methanol Determined by the Lithium Cluster-Continuum Approach. *J. Phys. Chem. B* **2013**, *117*, 5129−5135.

(29) Wolfenden, R. Waterlogged Molecules. *Science* **1983**, *222*, 1087−1093.

(30) Pearson, R. G. Ionization Potentials and Electron Affinities in Aqueous Solution. *J. Am. Chem. Soc.* **1986**, *108*, 6109−6114.

(31) Winget, P.; Cramer, C. J.; Truhlar, D. G. Computation of Equilibrium Oxidation and Reduction Potentials for Reversible and Dissociative Electron-Transfer Reactions in Solution. *Theor. Chem. Acc.* **2004**, *112*, 217−227.

(32) Curutchet, C.; Orozco, M.; Luque, F. J. Solvation in Octanol: Parametrization of the Continuum MST Model. *J. Comput. Chem.* **2001**, *22*, 1180−1193.

(33) Cancès, E.; Mennucci, B.; Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032.

(34) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct.: THEOCHEM* **2005**, *727*, 29−40.

(35) Chen, C.-S.; Lin, S.-T. Prediction of pH Effect on the Octanol-Water Partition Coefficient of Ionizable Pharmaceuticals. *Ind. Eng. Chem. Res.* **2016**, *55*, 9284−9294.

(36) Ingram, T.; Richter, U.; Mehling, T.; Smirnova, I. Modelling of pH Dependent *n*-Octanol/Water Partition Coefficients of Ionizable Pharmaceuticals. *Fluid Phase Equilib.* **2011**, *305*, 197−203.

(37) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **1997**, *86*, 865−871.

(38) Pieńko, T.; Grudzień, M.; Taciak, P. P.; Mazurek, A. P. Cytisine Basicity, Solvation, Log P, and Log D Theoretical Determination as Tool for Bioavailability Prediction. *J. Mol. Graphics Modell.* **2016**, *63*, 15−21.

(39) Avdeef, A. *Absorption and Drug Development: Solubility, Permeability, and Charge State*; Wiley: Hoboken, NJ, 2003.

(40) Peljo, P.; Girault, H. H. Liquid/Liquid Interfaces, Electrochemistry at. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons: 2012.

(41) Bouchard, G.; Carrupt, P. A.; Testa, B.; Gobry, V.; Girault, H. H. The Apparent Lipophilicity of Quaternary Ammonium Ions is Influenced by Galvani Potential Difference, not Ion-Pairing: A Cyclic Voltammetry Study. *Pharm. Res.* **2001**, *18*, 702−708.

(42) Kakiuchi, T. In *Liquid−Liquid Interfaces, Theory and Methods*; Volkov, A. G., Deamer, D. W., Eds.; CRC Press: Boca Raton, FL, 1996; pp 1−18.

(43) Kakiuchi, T. Limiting Behaviour in Equilibrium Partitioning of Ionic Components in Liquid-Liquid Two-Phase Systems. *Anal. Chem.* **1996**, *68*, 3658−3664.

(44) Fini, A.; Fazio, G.; Gonzaléz-Rodriguez, M.; Cavallari, C.; Passerini, N.; Rodríguez, L. Formation of Ion-Pairs in Aqueous Solutions of Diclofenac Salts. *Int. J. Pharm.* **1999**, *187*, 163−173.

(45) Marcus, Y.; Hefter, G. Ion Pairing. *Chem. Rev.* **2006**, *106*, 4585−4621.

(46) Alam, M. S.; Mandal, A.; Mandal, A. B. Effect of KCl on the Micellization and Clouding Phenomenon of the Amphiphilic Phenothiazine Drug Prometazine Hydrochloride: Some Thermodynamic Properties. *J. Chem. Eng. Data* **2011**, *56*, 1540−1546.

(47) Zuriaga, E.; Lomba, L.; Royo, F. M.; Lafuente, C.; Giner, B. Aggregation Behaviour of Betablocker Drugs in Aqueous Solution. *New J. Chem.* **2014**, *38*, 4141−4148.

(48) Pierotti, R. A. A Scaled Particle Theory of Aqueous and Nonaqueous Solutions. *Chem. Rev.* **1976**, *76*, 717−726.

(49) Claverie, P. In *Intermolecular Interactions: From Diatomics to Biopolymers*; Pullman, B., Ed.; Wiley: New York, 1978; Vol. *1*, pp 69−305.

(50) Luque, F. J.; Orozco, M.; Bhadane, P. K.; Gadre, S. K. Effect of Solvation on the Shapes, Sizes and Anisotropies of Polyatomic Anions via MESP Topology: An ab initio SCRF Approach. *J. Chem. Phys.* **1994**, *100*, 6718−6726.

(51) Luque, F. J.; Gadre, S. R.; Bhadane, P. K.; Orozco, M. Effect of Hydration on the Molecular Charge Distribution of Cations. An ab initio SCRF Study. *Chem. Phys. Lett.* **1995**, *232*, 509−517.

(52) Mestres, J.; Solà, M.; Carbó, R.; Luque, F. J.; Orozco, M. Effect of Solvation on the Charge Distribution of a Series of Anionic, Neutral, and Cationic Species. *J. Phys. Chem.* **1996**, *100*, 606−610.

(53) Curutchet, C.; Bidon-Chanal, A.; Soteras, I.; Orozco, M.; Luque, F. J. MST Continuum Study of the Hydration Free Energies of Monovalent Ionic Species. *J. Phys. Chem. B* **2005**, *109*, 3565−3574.

(54) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−1097.

(55) Arnold, M. R.; Kremer, W.; Lüdemann, H.; Kalbitzer, H. R. ^{1}H-NMR Parameters of Common Amino Acid Residues Measured in Aqueous Solutions of the Linear Tetrapeptides Gly-Gly-X-Ala at Pressures Between 0.1 and 200 MPa. *Biophys. Chem.* **2002**, *96*, 129−140.

(56) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. *Gaussian 09*, revision D.01; Gaussian, Inc.: Wallingford CT, 2009.

(57) Kolár, M.; Fanfrlík, J.; Lepsík, M.; Forti, F.; Luque, F. J.; Hobza, P. Assessing the Accuracy and Performance of Implicit Solvent Models for Drug Molecules: Conformational Ensemble Approaches. *J. Phys. Chem. B* **2013**, *117*, 5950−5962.

(58) Dunbrack, J. R. L.; Karplus, M. Backbone-Dependent Rotamer Library for Proteins Application to Side-Chain Prediction. *J. Mol. Biol.* **1993**, *230*, 543−574.

(59) Dunbrack, J. R. L.; Karplus, M. Conformational Analysis of the Backbone-Dependent Rotamer Preferences of Protein Sidechains. *Nat. Struct. Mol. Biol.* **1994**, *1*, 334−340.

(60) Quentel, F.; Mirčeski, V.; L'Her, M. Electrochemical Study of the Thermodynamics and Kinetics of Hydrophilic Ion Transfers Across Water | n-Octanol Interface. *J. Solid State Electrochem.* **2008**, *12*, 31−39.

(61) Bouchard, G.; Galland, A.; Carrupt, P. A.; Gulaboski, R.; Mirčeski, V.; Scholz, F.; Girault, H. H. Standard Partition Coefficients of Anionic Drugs in the *n*-Octanol/Water System Determined by Voltammetry at Three-Phase Electrodes. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3748−3751.

(62) Hansch, C.; Leo, A.; Hoekman, D. H. *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*; American Chemical Society: Washington, DC, 1995.

(63) Scherrer, R. A.; Donovan, S. F. Automated Potentiometric Titrations in KCl/ Water-Saturated Octanol: Method for Quantifying Factors Influencing Ion-Pair Partitioning. *Anal. Chem.* **2009**, *81*, 2768−2778.

(64) Austin, R. P.; Davis, A. M.; Manners, C. N. Partitioning of Ionizing Molecules Between Aqueous Buffers and Phospholipid Vesicles. *J. Pharm. Sci.* **1995**, *84*, 1180−1183.

(65) Wimley, W. C.; Creamer, T. P.; White, S. H. Solvation Energies of Amino Acid Side Chains and Backbone in a Family of Host-Guest Pentapeptides. *Biochemistry* **1996**, *35*, 5109−5124.

(66) Fauchère, J. L.; Pliska, V. Hydrophobic Parameters π of Amino Acid Side Chains from the Partitioning of N-Acetyl-Amino-Acid Amides. *Eur. J. Med. Chem.* **1983**, *18*, 369−375.

(67) Reymond, F.; Chopineaux-Courtois, V.; Steyaert, G.; Bouchard, G.; Carrupt, P.-A.; Testa, B.; Girault, H. H. Ionic Partition Diagrams of Ionisable Drugs: pH-Lipophilicity Profiles, Transfer Mechanisms and Charge Effects on Solvation. *J. Electroanal. Chem.* **1999**, *462*, 235−250.

(68) de Oliveira, P. R.; Tasic, L.; Rocco, S. A.; Rittner, R. Stereoelectronic and Inductive Effects on 1H and 13C NMR Chemical Shifts of Some cis-1,3-Disubstituted Cyclohexanes. *Magn. Reson. Chem.* **2006**, *44*, 790−796.

(69) ACD/I-Lab. Advanced Chemistry Development, Inc.: Toronto, ON, Canada; http://www.acdlabs.com.

(70) ChemAxon, Budapest, Hungary. http://www.chemaxon.com.

178

(71) Petrauskas, A. A.; Kolovanov, E. A. ACD/Log P Method Description. *Perspect. Drug Discovery Des.* **2000**, *19*, 99−116.

(72) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Model.* **1989**, *29*, 163−172.

(73) Mazák, K.; Noszál, B. Drug Delivery: A Process Governed by Species-Specific Lipophilicities. *Eur. J. Pharm. Sci.* **2014**, *62*, 96−104.

(74) Klamt, A.; Huniar, U.; Spycher, S.; Keldenich, J. COSMOmic: A Mechanistic Approach to the Calculation of Membrane-Water Partition Coefficients and Internal Distributions within Membranes and Micelles. *J. Phys. Chem. B* **2008**, *112*, 12148−12157.

(75) Juárez-Jiménez, J.; Barril, X.; Orozco, M.; Pouplana, R.; Luque, F. J. Assessing the Suitability of the Multilevel Strategy for the Conformational Analysis of Small Ligands. *J. Phys. Chem. B* **2015**, *119*, 1164−1172.

(76) Yordanova, D.; Ritter, E.; Gerlach, T.; Jensen, J. H.; Smirnova, I.; Jakobtorweihen, S. Solute Partitioning in Micelles: Combining Molecular Dynamics Simulations, COSMOmic, and Experiments. *J. Phys. Chem. B* **2017**, *121*, 5794−5809.

(77) Fini, A.; Bassini, G.; Monastero, A.; Cavallari, C. Diclofenac Salts, VIII. Effect of the Counterions on the Permeation through Porcine Membrane from Aqueous Saturated Solutions. *Pharmaceutics* **2012**, *4*, 413−429.

Supporting Information

Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from MST Continuum Solvation Calculations

William J. Zamora,[1,2] Carles Curutchet,[2] Josep M. Campanera[2,*] and F. Javier Luque[1,*]

[1] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, and Institut de Biomedicina (IBUB), Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain

[2] Departament de Farmàcia i Tecnologia Farmacèutica i Fisicoquímica, and Institut de Biomedicina (IBUB), Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Avgda. Diagonal 643, 08028 Barcelona, Spain

Figure S1. Data set of 19 neutral nitrogen-containing aromatic compounds used to refine the MST model for solvation in *n*-octanol. Nitrogen atoms subjected to reparametrization are shown in blue.
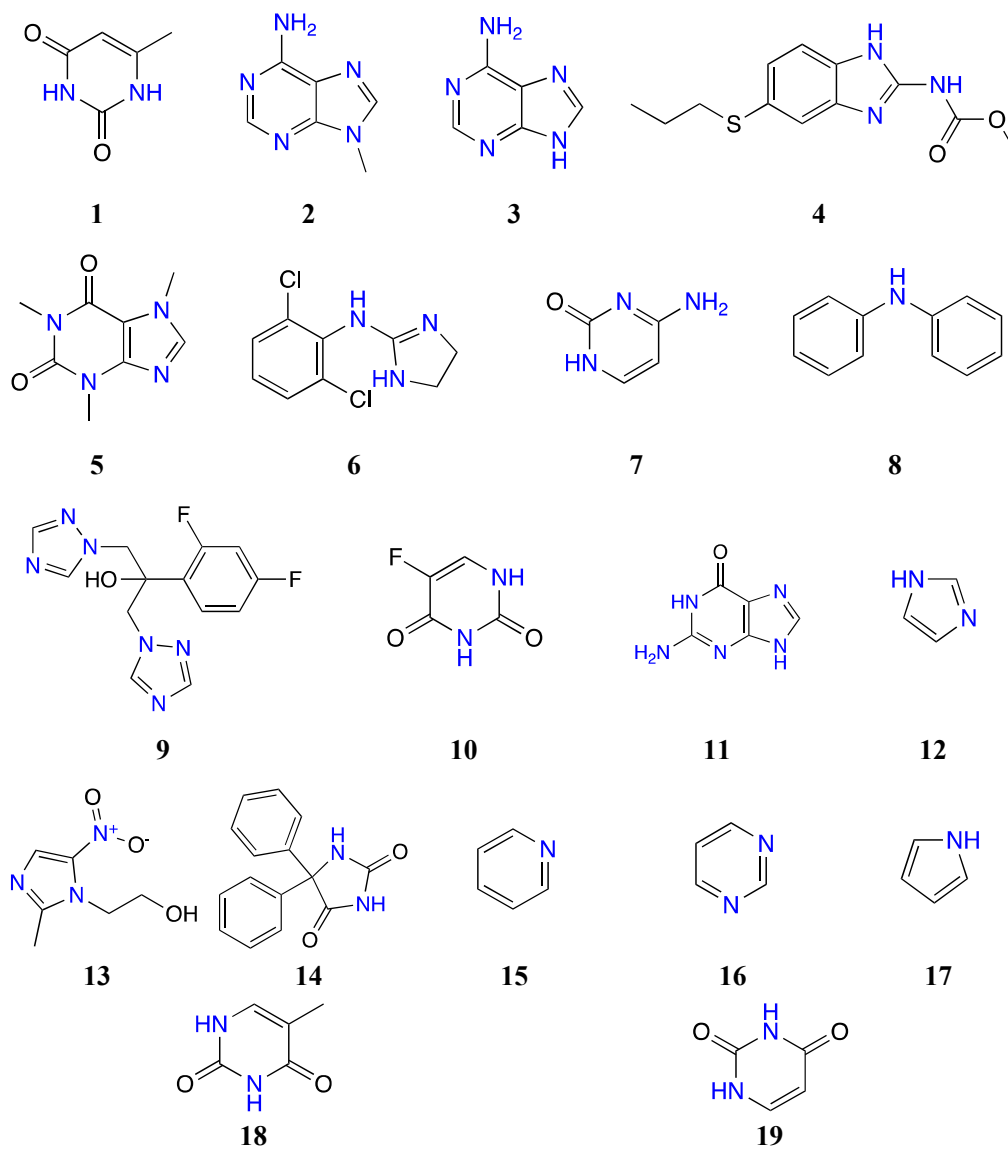
Figure S2. Data set of 27 ionizable compounds used to refine the MST solvation model for solvation in *n*-octanol. Atoms subjected to reparametrization are shown in blue and red for cations and anions, respectively.
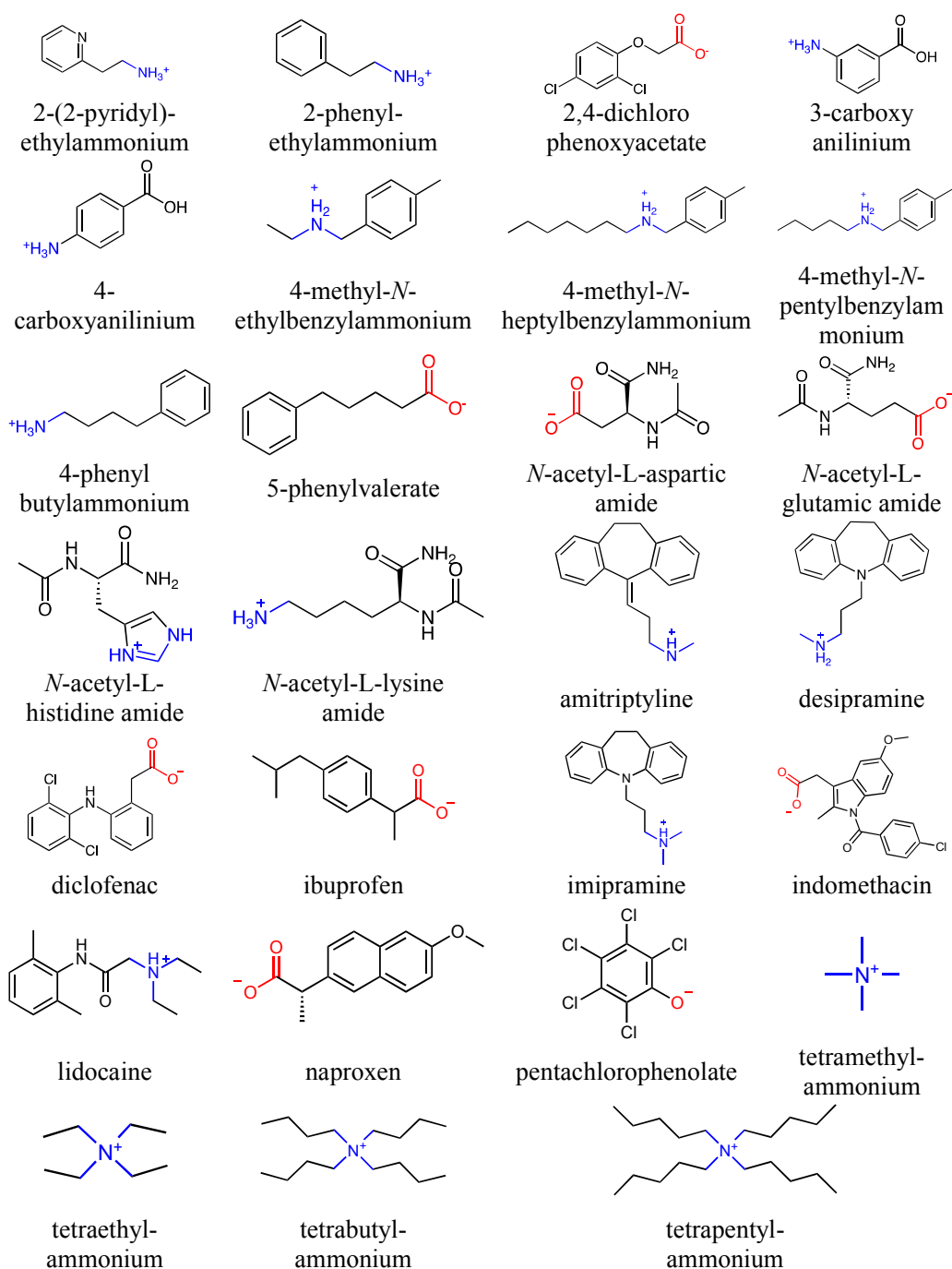

2-(2-pyridyl)-
ethylammonium

2-phenyl-
ethylammonium

2,4-dichloro
phenoxyacetate

3-carboxy
anilinium

4-
carboxyanilinium

4-methyl-*N*-
ethylbenzylammonium

4-methyl-*N*-
heptylbenzylammonium

4-methyl-*N*-
pentylbenzylam
monium

4-phenyl
butylammonium

5-phenylvalerate

N-acetyl-L-aspartic
amide

N-acetyl-L-
glutamic amide

N-acetyl-L-
histidine amide

N-acetyl-L-lysine
amide

amitriptyline

desipramine

diclofenac

ibuprofen

imipramine

indomethacin

lidocaine

naproxen

pentachlorophenolate

tetramethyl-
ammonium

tetraethyl-
ammonium

tetrabutyl-
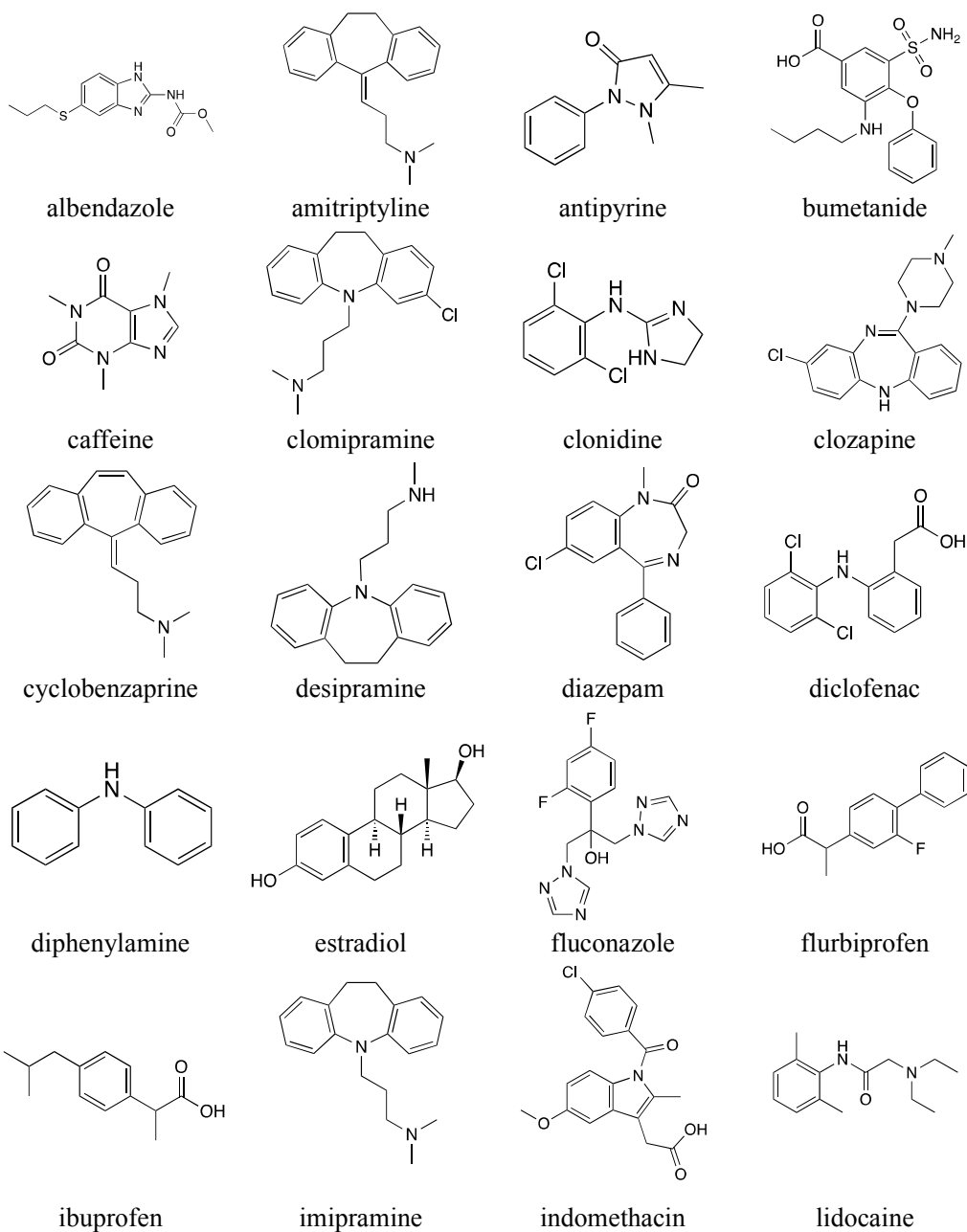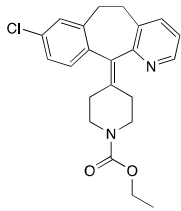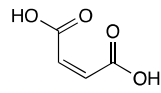ammonium

tetrapentyl-
ammonium

Figure S3. Data set of 35 ionizable compounds used to calibrate the MST solvation model for estimating the distribution coefficient.

| albendazole | amitriptyline | antipyrine | bumetanide |

| caffeine | clomipramine | clonidine | clozapine |

| cyclobenzaprine | desipramine | diazepam | diclofenac |

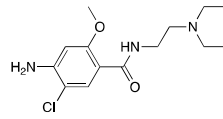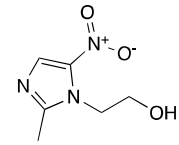| diphenylamine | estradiol | fluconazole | flurbiprofen |

| ibuprofen | imipramine | indomethacin | lidocaine |

loratadine

maleic acid

metoclopramide

metronidazole

mesoridazine
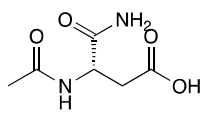
N-acetyl-L-aspartic amide
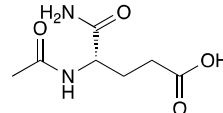
N-acetyl-L-glutamic amide
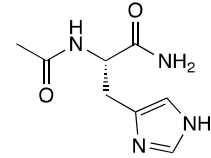
N-acetyl-L-histidine amide
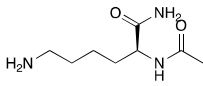
N-acetyl-L-lysine amide
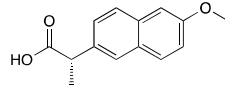
naproxen
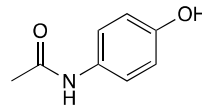
paracetamol
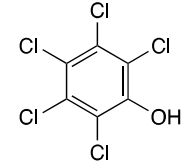
pentachlorophenol
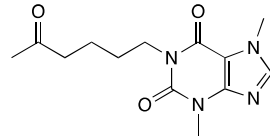
pentoxifylline
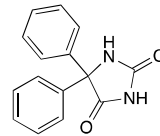
phenytoin
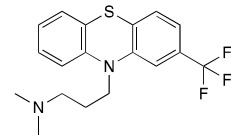
triflupromazine

Figure S4. Histograms of molecular properties (molecular weight, number of rotatable bonds, number of aromatic rings and hydrogen bond acceptors and donors) for the set of 35 small molecules.
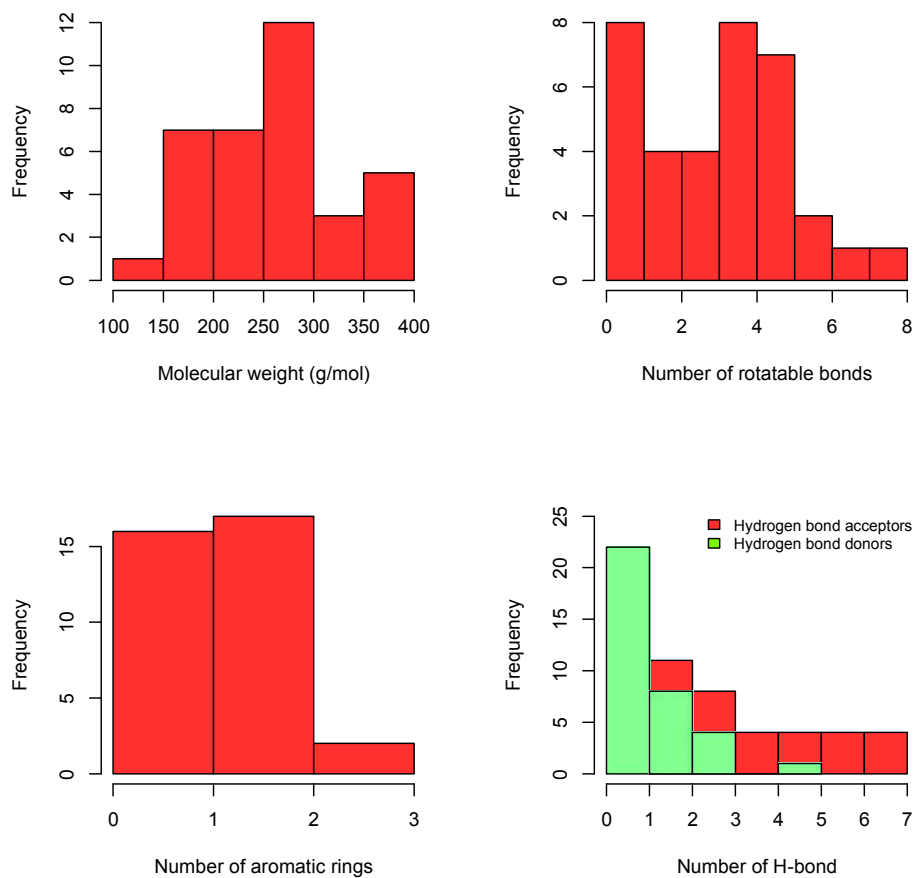
Figure S5. Comparison of *n*-octanol/water distribution coefficient profiles of selected drugs using ACD/I-Lab (cyan), ChemAxon (purple) and experimental data (black).
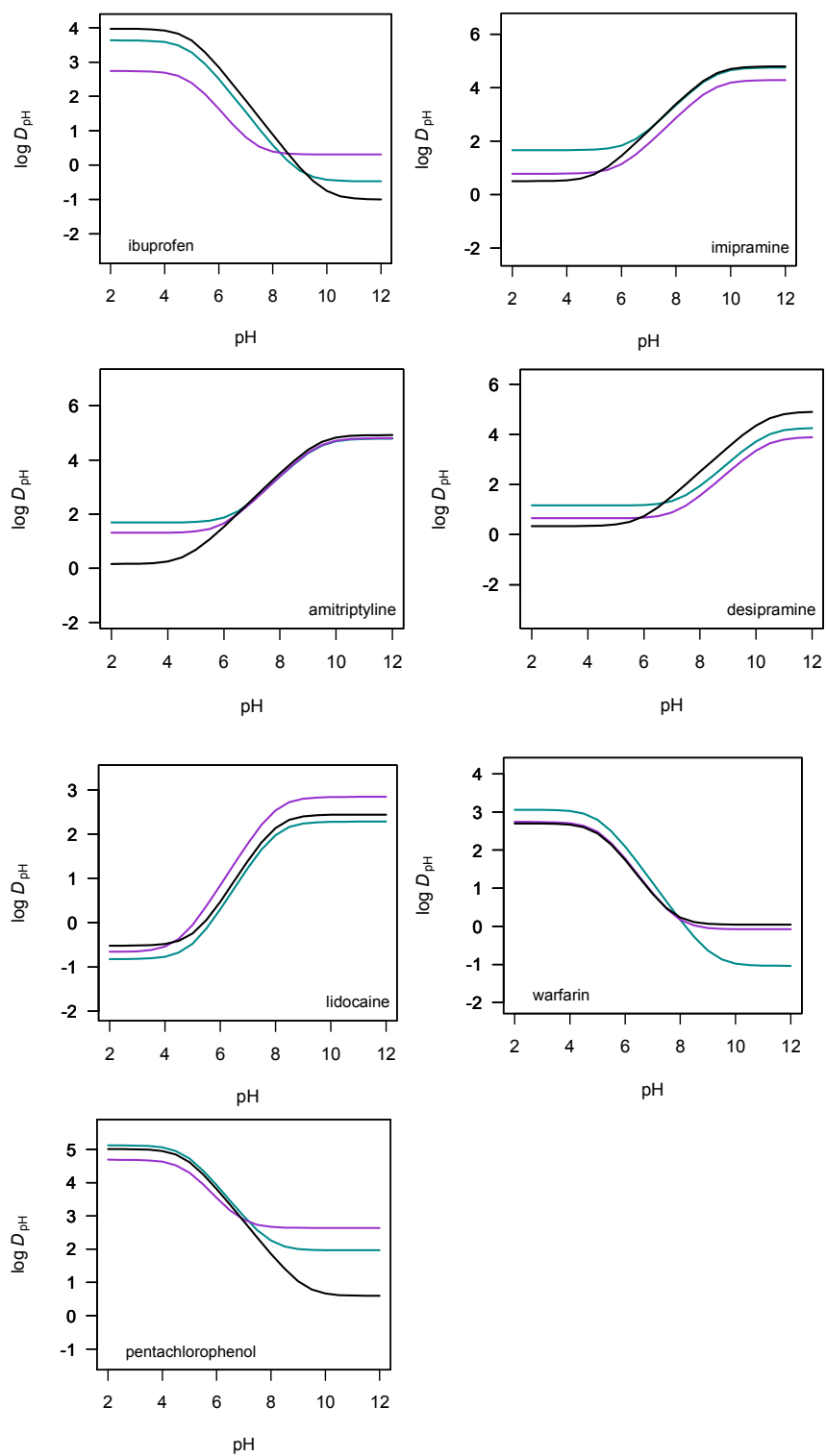
Figure S6. Comparison of *n*-octanol/water distribution coefficient profiles of selected amino acid analogues using ACD/I-Lab (cyan), ChemAxon (purple) and experimental data (black).
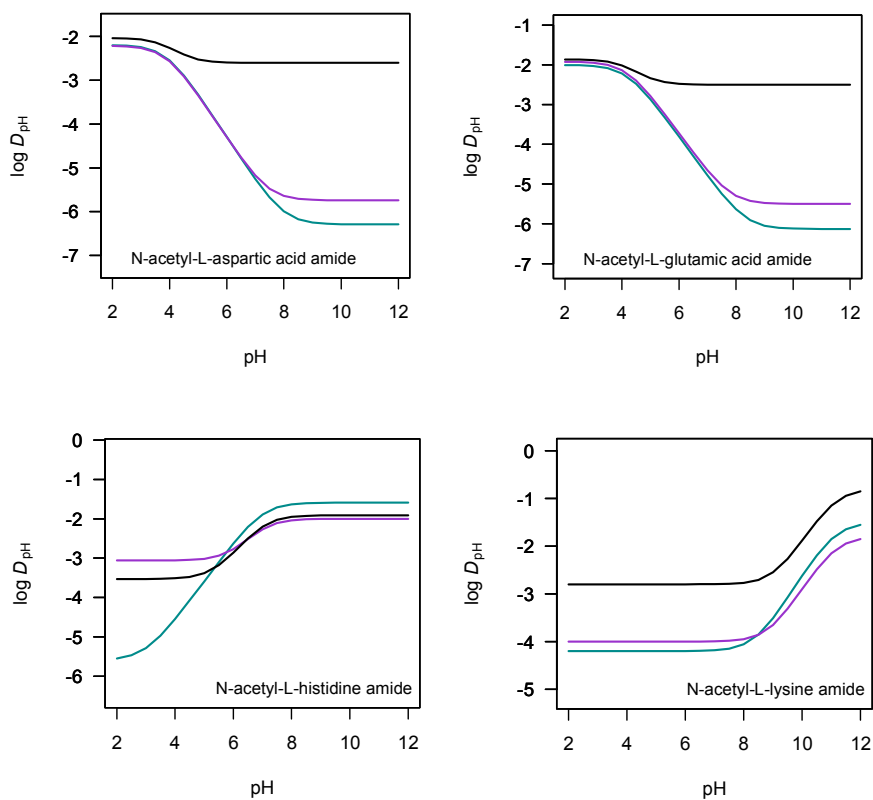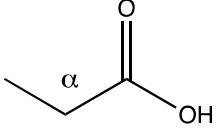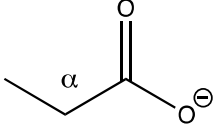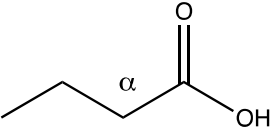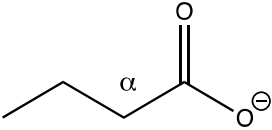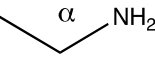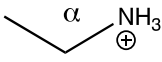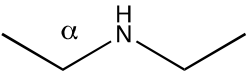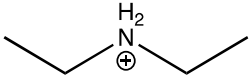
Table S1. Experimental data for the set of 35 ionizable compounds used to calibrate the MST solvation model for estimating the distribution coefficient (Data taken from refs. 1-3).
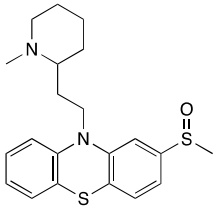
Species	Experimental pK_a	Experimental $\log P_N$	Experimental $\log D_{7.4}$
albendazole	4.21	2.70	3.29
amitriptiline	9.40	4.92	2.79
antipyrine	1.40	0.38	0.24
bumetanide	3.60	2.60	-0.11
caffeine	-0.92	-0.07	-0.04
clomipramine	9.40	5.19	3.28
clonidine	8.05	1.59	0.62
clozapine	7.50	3.23	2.99
cyclobenzaprine	8.47	5.20	2.90
desipramine	10.40	4.90	1.41
diazepam	3.30	2.82	2.74
diclofenac	4.15	4.51	1.14
diphenylamine	1.03	3.50	3.36
estradiol	10.71	3.67	4.01
fluconazole	2.94	0.40	0.50
fulbiprofen	4.22	4.16	0.91
ibuprofen	4.91	3.97	1.27
imipramine	9.40	4.80	2.51
indomethacin	4.50	4.27	0.76
lidocaine	8.01	2.44	1.61
loratadine	4.58	5.20	4.40
maleic acid	1.83	-0.48	-5.00
metoclopramide	9.27	2.62	0.53
metronidazole	2.60	-0.02	-0.07
mesoridazine	8.89	3.90	1.81
N-acetyl-L-aspartic amide	3.90	-2.03	-2.60
N-acetyl-L-glutamic amide	4.20	-1.86	-2.50
N-acetyl-L-histidine amide	7.00	-1.91	-3.53
N-acetyl-L-lysine amide	11.10	-0.80	-2.80
naproxen	4.15	3.18	0.32
paracetamol	9.38	0.46	0.30
pentachlorophenol	4.83	5.01	2.50
pentoxifylline	0.28	0.32	0.29
phenytoin	8.33	2.47	2.23
triflupromazine	9.40	5.54	3.39

Table S2. The substituent α-effect in ^{1}H-NMR for anionic and cationic organic compounds.

Neutral species δ_H (ppm) in α	Charged species δ_H (ppm) in α	diff (ppm) $\delta_{H,charged} - \delta_{H,neutral}$
2.18	2.38	0.2
2.16	2.33	0.2
2.65	3.06	0.4
2.65	3.04	0.4

http://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

S10

Additive scheme for acetyl-L-amino acid amide

log P_I of histidine. The partition coefficient for the ionic species of histidine (log P_I) was derived using the following additivity scheme:

Log P_I (*N*-acetyl-*L*-histidine amide) = (log D_1 (Ac-WL**H**LL) − log D_1 (Ac-WL**G**LL))$_{side\ chain}$ +
(log P_N (*N*-acetyl-*L*-glycine amide))$_{backbone+capping\ groups}$

where log D_1 (Ac-WL**H**LL) and log D_1 (Ac-WL**G**LL) are the *n*-octanol/water distribution coefficient at pH = 1 for a pentapeptide model containing at the center of the amino acid sequence the amino acid histidine (note that the imidazole ring will be doubly protonated at this pH) and glycine, respectively, and log P_N (*N*-acetyl-*L*-glycine amide) is the partition coefficient for the glycine amino acid analogue (AcNHCOCH$_2$CONH$_2$) at pH ≈ 7.

Experimental data for the ionizable side chains were taken from the scale at extreme pH reported by Wimley and White.[4] The value for the neutral backbone was taken from the scale at physiological pH reported by Fauchère.[5]

log P_I of acetyl-L-amino acid amides. The partition coefficient for the neutral species of ionizable amino acids was calculated by using the following expression:

Log P_N Ac-**X** amide = (log D_{pH} (Ac-WL**X**LL) − log D_{pH} (Ac-WL**G**LL)) $_{side\ chain}$ +
(log P_N (*N*-acetyl-*L*-glycine amide)) $_{backbone+capping\ groups}$

where log D_{pH} (Ac-WL**X**LL) and log D_{pH} (Ac-WL**G**LL) denote the *n*-octanol/water distribution coefficient at pH = 1 (for X = D and E) or 9 (for X= H and K) according to the scale by Wimley and White.

References

(1) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson. M.; Neveu, V.; et al. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nuc. Acids Res.* **2014**, *42*, D1091-1097.

(2) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem. Rev.* **1971,** *71*, 525–616.

(3) Wagner, B.; Fischer, H.; Kansy, M.; Seelig, A.; Assmus, F. Carrier Mediated Distribution System (CAMDIS): A New Approach for the Measurement of Octanol/Water Distribution Coefficients. *Eur. J. Pharm. Sci.* **2015,** *68*, 68–77.

(4) Wimley, W. C.; Creamer, T. P.; White, S. H. Solvation Energies of Amino Acid Side Chains and Backbone in a Famili of Host-Guest Pentapeptides. *Biochemistry* **1996**, *35*, 5109–5124.

(5) Fauchère, J. L.; Pliska, V. Hydrophobic Parameters π of Amino Acid Side Chains from the Partitioning of N-Acetyl-Amino-Acid Amides. *Eur. J. Med. Chem.* **1983**, *18*, 369–375.

Toward Refined Theoretical Models for the Description of Lipophilicity in Biomolecules

Appendix II. Article II:

Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations.

William J. Zamora, Josep Maria Campanera, F. Javier Luque, *Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations*, J. Phys. Chem. Lett. **2019**, 10, 883−889.

Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations

William J. Zamora, Josep M. Campanera,* and F. Javier Luque*

Department of Nutrition, Food Science and Gastronomy, Faculty of Pharmacy and Food Science, Institute of Biomedicine (IBUB) and Institute of Theoretical and Computational Chemistry (IQTCUB), Campus Torribera, University of Barcelona, 08921 Santa Coloma de Gramenet, Spain

S *Supporting Information*

ABSTRACT: Lipophilicity is a fundamental property to characterize the structure and function of proteins, motivating the development of lipophilicity scales. We report a versatile strategy to derive a pH-adapted scale that relies on theoretical estimates of distribution coefficients from conformational ensembles of amino acids. This is accomplished by using an accurately parametrized version of the IEFPCM/MST continuum solvation model as an effective way to describe the partitioning between *n*-octanol and water, in conjunction with a formalism that combines partition coefficients of neutral and ionic species of residues and the corresponding pK_a values of ionizable groups. Two weighting schemes are considered to derive solvent-like and protein-like scales, which have been calibrated by comparison with other experimental scales developed in different chemical/biological environments and pH conditions as well as by examining properties such as the retention time of small peptides and the recognition of antigenic peptides. A straightforward extension to nonstandard residues is enabled by this efficient methodological strategy.
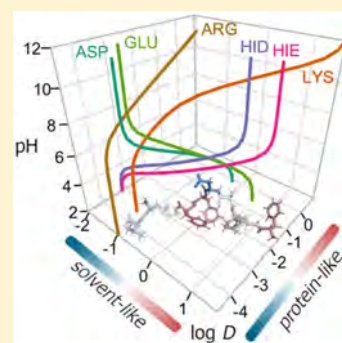
L ipophilicity is a cornerstone concept in chemistry and biology, as this property is crucial to understanding a variety of processes, such as the partitioning of molecules into immiscible solvents, the formation of host−guest complexes, the folding of proteins, and the stability of supramolecular aggregates.[1,2] In proteins, lipophilicity is mainly determined by the side chains of amino acids, and obtaining quantitative lipophilicity profiles of peptides and proteins is key to examine their structural and functional properties in biological environments. Accordingly, several strategies have been proposed to quantify the lipophilicity of amino acids, leading to lipophilicity scales that exploit the partitioning of small molecules between bulk solvents, the application of knowledge-based techniques to structural data, or experimental information derived from biological assays. (For comprehensive reviews, see refs 3−5.) Using these scales, lipophilicity profiles of peptides or proteins can be derived from the lipophilicity of single residues, generally assuming an additivity principle. Nevertheless, there are differences not only in the absolute magnitude of the residue lipophilicities but also in the relative values, giving rise to a variable degree of correlation between scales that reflects the differences between the material systems, methods, and experimental conditions that underlie the definition of each scale.

In this study, our aim is to develop a lipophilicity scale from theoretical computations that takes into account the structural dependence of the conformational preferences of amino acids as well as the influence of pH to provide a consistent description of pH-adapted lipophilicity profiles in peptides and proteins. Here attention is focused on the set of natural amino

acids, but the methodological strategy is intended to be easily adapted to nonstandard residues, such as nonproteinogenic residues, or to chemical modifications, such as phosphorylation, sulphonation, and nitrosation, which regulate enzyme activity and signaling processes. To achieve this goal, each residue has been characterized by its distribution coefficient (D_{pH}) using as a model system the corresponding *N*-acetyl-L-amino acid amides, taking into account the potential contribution of ionizable species at a given pH, as noted in eq 1, which has recently been shown to reproduce the pH-dependent lipophilicity profiles of amino acid analogues.[6]

$$\log D_{pH} = \log(P_N + P_I \cdot 10^\delta) - \log(1 + 10^\delta) \quad (1)$$

where P_N and P_I denote the partition coefficients of neutral and ionized species of an ionizable amino acid and δ is the difference between the pK_a of the ionizable group and the pH of the environment.

Let us note that the choice of *N*-acetyl-L-amino acid amides in this study enables a direct comparison with the experimental results reported by Fauchère and Pliska[7] because their experimental lipophilicity scaled was determined using these model systems in their study. The partition coefficients P_N and P_I were determined from theoretical computations using the B3LYP/6-31G(d) version of the quantum-mechanical IEFPCM-MST continuum solvation method,[8] which relies

Table 1. Solvent-Like (SolvL) and Protein-Like (ProtL) Lipophilicity Scales Based on the log D_{pH} Values Determined for *N*-Acetyl-L-amino Acid Amides at Physiological pH[a]

| residue | exp. pK_a | | log P_N | | log P_I | | log $D_{7.4}$[b] | |
	SolvL	ProtL	SolvL	ProtL	SolvL	ProtL	SolvL	ProtL
Ala			−1.16	−2.47			−1.16 (0.85)	−2.47 (0.66)
Arg	**12.5**[c]	**12.5**[c]	**−2.86**	**−3.66**	**−2.99**	**−7.38**	**−2.99 (−0.98)**	**−7.04 (−3.91)**
Asn			−2.98	−3.97			−2.98 (−0.97)	−3.97 (−0.84)
Asp	**3.90**[d]	**3.50**[e]	**−2.26**	**−3.18**	**−2.80**	**−8.54**	**−2.80 (−0.79)**	**−5.87 (−2.74)**
Cys	**9.83**[f]	**6.80**[e]	**−0.16**	**−1.47**	**−4.19**	**−5.78**	**−0.16 (1.85)**	**−2.17 (0.96)**
Gln			−2.22	−4.00			−2.22 (−0.21)	−4.00 (−0.87)
Glu	**4.20**[d]	**4.20**[e]	**−1.49**	**−3.79**	**−3.38**	**−6.20**	**−3.36 (−1.35)**	**−5.96 (−2.83)**
Gly			−2.01	−3.13			−2.01 (0.00)	−3.13 (0.00)
His (δ)	**7.00**[d]	**6.60**[e]	**−1.20**	**−4.67**	**−4.06**	**−5.97**	**−1.35 (0.66)**	**−4.56 (−1.43)**
His (ε)	**7.00**[d]	**6.60**[e]	**−0.72**	**−4.98**	**−4.06**	**−5.97**	**−0.87 (1.14)**	**−4.97 (−1.84)**
Ile			−0.50	−0.38			−0.50 (1.51)	−0.38 (2.75)
Leu			0.05	−1.36			0.05 (2.06)	−1.36 (1.77)
Lys	**11.1**[d]	**10.5**[e]	**−0.40**	**−2.19**	**−3.24**	**−6.81**	**−3.18 (−1.17)**	**−5.08 (−1.95)**
Met			−0.51	−1.83			−0.51 (1.50)	−1.83 (1.30)
Phe			0.61	0.86			0.61 (2.62)	0.86 (3.99)
Pro			−0.77	−1.44			−0.77 (1.24)	−1.44 (1.69)
Ser			−2.04	−4.12			−2.04 (−0.03)	−4.12 (−0.99)
Thr			−1.22	−3.01			−1.22 (0.79)	−3.01 (0.12)
Trp			0.33	0.16			0.33 (2.34)	0.16 (3.29)
Tyr	**10.3**[d]	**10.3**[e]	**−0.49**	**−1.80**	**−4.21**	**−9.59**	**−0.49 (1.52)**	**−1.80 (1.33)**
Val			−0.93	−1.68			−0.93 (1.08)	−1.68 (1.45)

[a]Experimental pK_a of side-chain ionizable groups and calculated partition coefficients of neutral (log P_N) and ionized (log P_I) residues are also given. [b]Values for ionizable residues are shown in bold. Log $D_{7.4}$ values relative to glycine are given in parentheses. [c]Ref 14. [d]Ref 15. [e]Ref 16. [f]Ref 17.
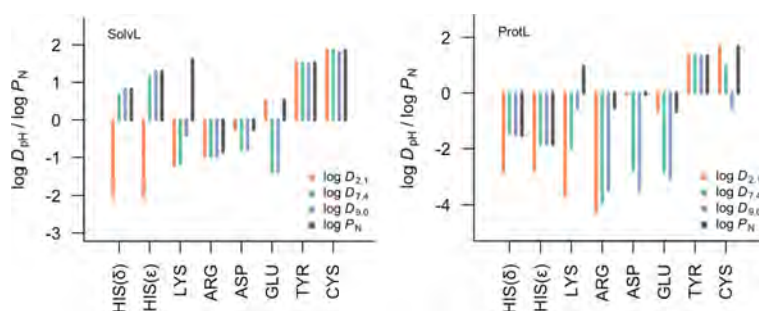
Figure 1. Representation of the pH dependence of the SolvL (left) and ProtL (right) lipophilicity scales for ionizable amino acids (values relative to Gly). Values determined at pH of 2.1, 7.4, and 9.0 are shown in orange, green, and blue, respectively, and the values of the neutral species (log P_N) are shown in black.

on the integral equation formalism (IEF) of the polarizable continuum model (PCM).[9,10] Following our previous study of the hydration free energy of the natural amino acids,[11] the backbone-dependent conformational library compiled by Drunback and coworkers[12−14] (http://dunbrack.fccc.edu) was used to extract the conformational preferences of residues, which defined the ensemble of structures used to estimate the log D_{pH} values from IEFPCM-MST calculations in *n*-octanol and water. (See the SI for a detailed description of the computational methods.)

Two schemes were explored for weighting the contribution of each conformational state to the differential solvation in the two solvents. In one case, P_N and P_I were determined using a Boltzmann's weighting scheme to the relative stabilities of the conformational species of a given residue in the two solvents, leading to the solvent-like scale (SolvL). In the second scheme, named protein-like scale (ProtL), the contribution of each

conformation was directly taken from the population distribution reported in the backbone-dependent conformational library. Therefore, these weighting schemes are expected to yield scales better suited for reflecting the lipophilic balance of amino acids well exposed to bulk solvent or in a protein-like environment, respectively. Finally, the effect of pH on the log D_{pH} values was introduced from the experimental pK_a values of ionizable residues in peptide models in aqueous solution[15,16] and in folded proteins[17,18] for the SolvL and ProtL scales, respectively.

The values of these lipophilicity scales for the amino acids at physiological pH are shown in Table 1. (ProtL data are averages of the log $D_{7.4}$ values determined separately for α-helix and β-sheet structures, which are reported in Table S1.) Taking Gly as reference, the ProtL scale comprises log $D_{7.4}$ values ranging from −3.91 (Arg) to 3.99 (Phe), reflecting the extreme values of hydrophilic residues (Arg, Asp, Glu, and

196
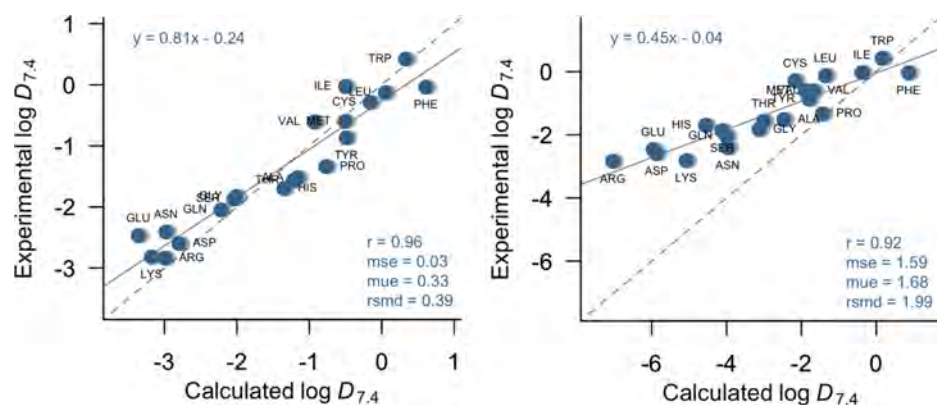
Figure 2. Comparison between (left) SolvL and (right) ProtL lipophilicity scales derived from the IEF/MST solvation model (expressed as log $D_{7.4}$) and Fauchère−Pliska experimental values for the 20 N-acetyl-L-amino acid amides (r, Pearson correlation coefficient; mse, mean signed error; mue, mean unsigned error; rmsd, root-mean-square deviation). Regression equations are shown in Table S8.

Table 2. Statistical Parameters of the Comparison of the SolvL and ProtL Scales with Other Lipophilicity Scales[a]

| scale[b]	SolvL				ProtL			
	mse[c]	mue	rsmd	r/p value	mse	mue	rsmd	r/p value
Bulk-Solvent-Adapted Scale								
Fauchère−Pliska	−0.20	0.36	0.46	0.94	0.36	0.98	1.28	0.92
				2×10^{-10}				6×10^{-9}
Eisenberg−McLachlan	−0.20	0.44	0.57	0.90	0.36	1.08	1.35	0.91
				3×10^{-8}				2×10^{-8}
Hopp−Woods	−0.49	0.60	0.74	0.91	0.07	0.84	1.08	0.89
				2×10^{-8}				9×10^{-8}
Wimley et al.	−0.60	1.02	1.16	0.59	0.04	1.24	1.64	0.61
				0.006				0.004
	−0.87[d]	0.92	1.03	0.87	−0.30	1.03	1.25	0.87
				2×10^{-6}				2×10^{-6}
Biological-Based Scale								
Moon−Fleming	−0.12	0.57	0.67	0.94	0.24	0.72	0.93	0.91
				4×10^{-10}				7×10^{-9}
Hessa et al.	−0.92	0.93	1.18	0.79	−0.36	1.08	1.46	0.82
				3×10^{-5}				6×10^{-6}
Knowledge-Based Scale								
Koehler et al.	−0.91	1.10	1.33	0.78	−0.35	1.55	1.87	0.80
				4×10^{-5}				2×10^{-5}
Janin et al.	−1.06	1.11	1.32	0.78	−0.51	1.36	1.71	0.74
				3×10^{-5}				2×10^{-4}
Consensus Scale								
Kyte−Doolittle	−0.81	1.43	1.71	0.72	−0.25	1.13	1.41	0.78
				3×10^{-4}				3×10^{-5}

[a]Comparison was made using the values adapted to the specific pH of each scale and relative to Gly. [b]Physiological pH was considered in all cases except for Wimley at al. and Moon−Fleming because the corresponding pH was fixed at 9.0 and 3.8 following the specific experimental conditions. [c]mse, mean signed error; mue, mean unsigned error; rmsd, root-mean-square deviation; r, Pearson correlation coefficient; p, statistical p value. mse, mue, and rmsd are given in log P_N/D units. [d]Values in this row were obtained upon the exclusion of Arg and Lys. Because this scale was built up using model pentapeptides (AcWL-X-LL) at pH 9.0, Arg and Lys formed a salt bridge with the terminal carboxyl group in n-octanol, as noted by ^{13}C NMR studies.[34]

Lys) and hydrophobic ones (Trp, Phe). (See also Figure S1.) These trends are also found in the SolvL scale, although the distribution of log $D_{7.4}$ values varies from −1.35 (Glu) to 2.62 (Phe). This trait is also found in other scales, as knowledge-based methods generally give rise to a narrower range of lipophilicites compared with other experimental scales.[19] In our case, this arises from the distinct weighting factors used in ProtL and SolvL scales, leading to larger differences in the log $D_{7.4}$ values of polar and ionizable amino acids, which show a

preference for extended conformations (Figure S2), likely reflecting the formation of stabilizing interactions (e.g., salt bridges) or the solvent exposure to bulk water in proteins.[20,21]

The sensitivity of the lipophilicity of ionizable residues to pH changes is shown in Figure 1, which compares the log D_{pH} values at pH 2.1, 7.4, and 9.0, chosen as representative values of the pH changes along the gastrointestinal tract. The hydrophilicity of acid/basic amino acids is enhanced at basic/acidic pH values, as expected from the predominance of the

197

ionic species. On the SolvL scale, it is worth noting the hydrophilic nature of protonated His at acidic pH and the slight hydrophobicity of protonated Glu. In contrast, the ProtL scale exhibits a higher sensitivity to pH, as noted in the large changes in the $\log D_{pH}$ values of Asp and Glu, which are decreased 2 to 3 $\log D_{pH}$ units upon deprotonation, the reduced hydrophilicity of Lys at basic pH, and the change from hydrophobic (at acid and physiological pH) to hydrophilic (at basic pH) Cys. This reflects the ability of these scales to present the pH influence on the lipophilicity of ionizable residues, which may be affected by the local environment in proteins.[22,23]

To calibrate the suitability of these scales, a comparison was made with the $\log D_{7.4}$ values reported by Fauchère and Pliska,[7] which were experimentally determined from the partitioning of N-acetyl-L-amino acid amides between n-octanol and water at physiological pH (Figure 2). A comparison with the SolvL values gives satisfactory results, as noted in a correlation coefficient (r) of 0.96 and a mean unsigned error (mue) of 0.33 $\log D_{7.4}$ units for a set of experimental values ranging from −3.36 to 0.61. The correlation coefficient is slightly worse ($r = 0.92$), and the mue increases to 1.68 for the ProtL scale. For the sake of comparison, the same analysis was performed by using $\log D_{7.4}$ values obtained from computations with the SMD solvation model,[24] in conjunction with the two weighting schemes, and the results also revealed a better performance for the solvent-adapted scheme ($r = 0.85$, mue = 0.83; Figure S3). On the contrary, the SolvL scale also performed better than the empirical estimates of $\log D_{7.4}$ obtained from ACD/ILab[25] ($r = 0.88$, mue = 0.60) and ChemAxon[26] ($r = 0.92$, mue = 0.65) when compared with the experimental values reported by Fauchère and Pliska (Figure S4).

Table 2 shows the comparison of the SolvL and ProtL lipophilicities with experimental scales, including four bulk-solvent-based scales (Fauchère–Pliska,[7] Eisenberg–McLachlan,[27] Hopp–Woods,[28] and Wimley et al.[29]), two biological-derived (Moon–Fleming[30] and Hessa et al.[31]) and two knowledge-based (Koehler et al.[19] and Janin et al.[32]) scales, and a consensus (Kyte–Doolittle[33]) scale. The bulk-solvent-based scales rely on experimental measurements of the transfer between n-octanol and water (Fauchère–Pliska, Eisenberg–McLachlan) at physiological pH or under basic conditions (pH 9.0; Wimley et al.) and between ethanol and the vapor phase (Hopp–Woods). Excellent correlations are found with Fauchère–Pliska, Eisenberg–McLachlan, and Hopp–Woods scales ($0.89 < r < 0.92$). A worse correlation ($r \approx 0.60$) is found in the comparison with Wimley et al. scale, but to large extent this can be attributed to the formation of salt bridges between Arg/Lys residues with the terminal carboxyl group in n-octanol for the AcWL-X-LL pentapeptides used as model systems, as noted by ^{13}C NMR studies.[34] The exclusion of Arg and Lys enhances the correlation coefficient to 0.87. On the contrary, the bulk-solvent-based lipophilicities are consistently closer to the values collected in the SolvL scale (mue of 0.36 to 0.92 $\log P/D$ units) than to the ProtL ones (mue of 0.84 to 1.24 $\log P/D$ units).

The correlation coefficients obtained with biological-, knowledge-based, and consensus scales are satisfactory ($0.74 < r < 0.94$; Table 2) but tend to be lower than the values obtained with the bulk-solvent-based transfer scales. This is not unexpected keeping in mind that the lipophilicites are derived from the statistical analysis of topological distributions of

residues in proteins (Koehler et al., Janin et al.) or from complex biochemically adapted assays, such as the transfer of amino acids from water to a phospholipid bilayer (Moon–Fleming), the recognition of artificial helices by the Sec61 translocon (Hessa et al.), or the combination of water-vapor transfer free energies with the interior–exterior distribution of amino acids in the consensus (Kyle–Doolittle) scale. Keeping in mind the notable differences in the material systems and protocols used to derive these experimental scales, the correlation coefficients obtained from the comparison with the SolvL scale are still remarkable.

The sensitivity of the results to the pH was examined by extending the comparison to the lipophilicities determined for the SolvL and ProtL scales at pH values of 3.8, 7.4, and 9.0. (Note that the acidic and basic pH values were chosen in the studies reported by Moon and Fleming and Wimley et al., respectively.) In general, there is little difference between the correlation coefficients obtained at pH 7.4 and 9.0 (Figure 3).
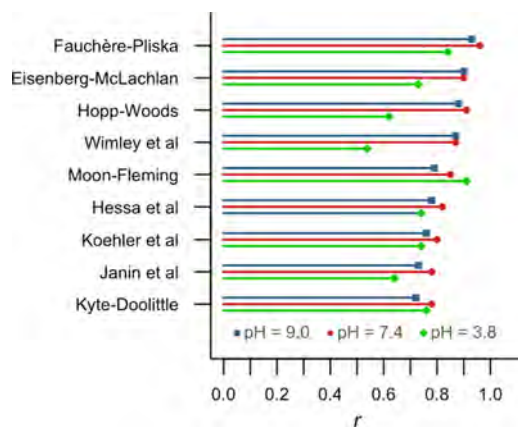
Figure 3. Representation of the Pearson correlation coefficient in the comparison of the SolvL scale with bulk-solvent-based scales and ProtL scale with biological-based, knowledge-based, and consensus lipophilicity scales at pH 3.8, 7.4, and 9.0 (shown as green, red, and blue lines, respectively).

However, a larger effect is found in the comparison of the $\log D_{3.8}$, as there is a general decrease in the correlation coefficient, which is remarkable for the bulk-solvent-based transfer scales, especially in the case of Hoop−Woods and Wimley et al. The only exception is found in the comparison with the Moon−Fleming scale, as the highest correlation coefficient is found for the ProtL values corrected at pH 3.8. These findings support the suitability of the SolvL/ProtL scales to account for the pH influence on the lipophilicity of amino acids.

The reliability of the SolvL/ProtL scales has been calibrated by comparing the cumulative lipophilicity with the (RP-HPLC) retention time determined for different sets of peptides.[35,36] Given the small size of the peptides (≤ 13 residues) and the lack of well-defined secondary structures, nonadditivity effects can be expected to play a minor role.[37] Accordingly, the cumulative lipophilicity was determined assuming an additive scheme (eq S3 in the SI Computational Methods).

The first test comprises eight 10-mer peptides with equal charge that differ in the content of hydrophobic residues (Table S2).[38] The SolvL cumulative lipophilicity yields a correlation coefficient of 0.96 (Figure 4A), which compares

198
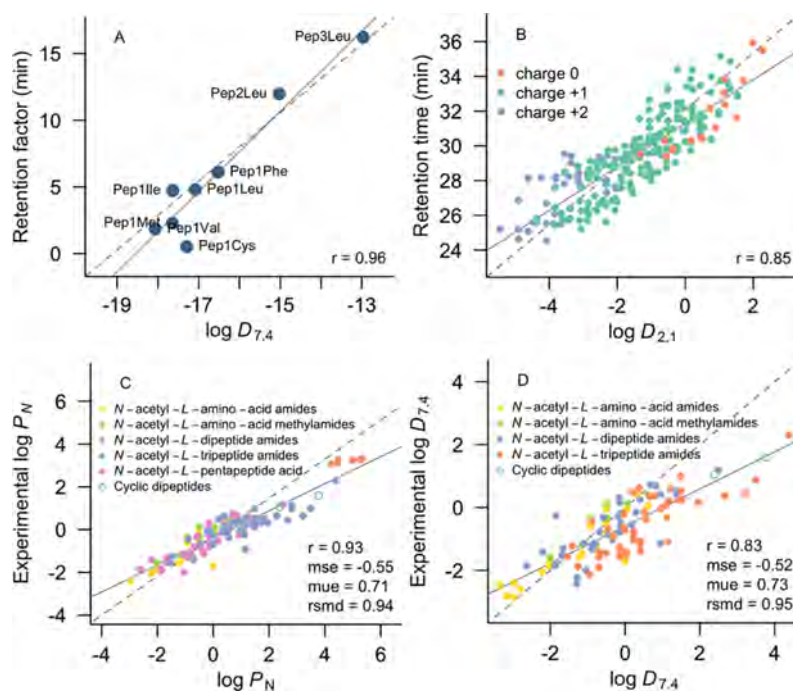
Figure 4. Relationship between the cumulative lipophilicities determined from the SolvL scale versus (A) the retention time for eight 10-mer peptides (pH 7.4; ref 38), (B) 248 unique 13-mer peptides (pH 2.1; refs 39 and 40), (C) log P_N for 118 random peptides (ref 42), and (D) log $D_{7.4}$ for 116 random peptides (ref 42). Regression equations are shown in Table S8.
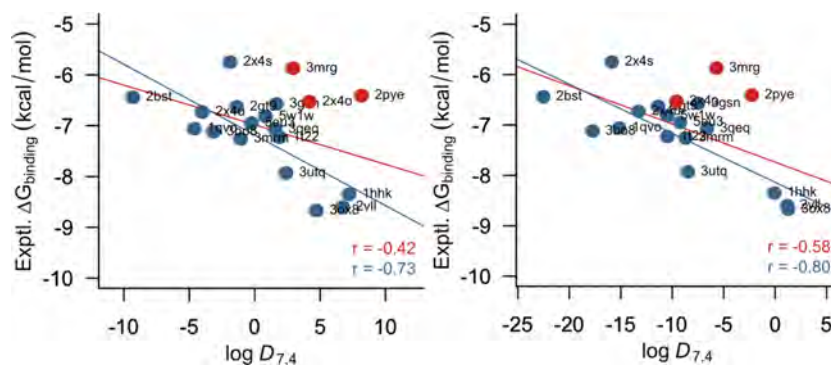
Figure 5. Relationship between the cumulative lipophilicities determined from (left) SolvL and (right) ProtL scales versus experimental binding affinities of MHC-bound peptides. Cys-containing peptides are indicated as red dots. Regression equations shown in Table S8.

with the value estimated from the hydrophobic surfaces of peptides derived from molecular dynamics simulations ($r = 0.97$),[38] whereas a slightly lower correlation was found for the ProtL scale ($r = 0.91$; Table S3). For this simple set of homogeneous peptides, most of the experimental lipophilicity scales generally yielded correlations >0.9 (Table S3).

A more challenging test is the set of 248 peptides with equal length but different net charge under the experimental acidic conditions (pH 2.1),[39,40] comprising 36 peptides with two charged amino acids (Arg combined with His or Lys), 105 peptides with a single charged residue (Arg, Lys, or His), and finally 17 neutral peptides. The SolvL cumulative lipophilicity correlates satisfactorily with the retention time determined for the whole set of peptides ($r = 0.85$; Figure 4B). Among bulk-solvent-based scales, Fauchère−Pliska, Eisenberg−McLachlan, and Hopp−Woods also provided reasonable correlations coefficients ($0.74 < r < 0.85$; Table S2 and Figure S6), but a

worse correlation was found for Wimley et al., although this may be attributed to the different pH used in this latter scale (pH 9.0) and the experimental assay conditions (pH 2.1). The performance of biological-based, knowledge-based, and consensus scales was also worse ($0.55 < r < 0.64$; Table S3 and Figure S5), but for Moon−Fleming ($r = 0.78$), it likely reflected the acidic pH conditions considered in the derivation of this lipophilicity scale.

Finally, given the relevance of partition (log P_N)/distribution (log $D_{7.4}$) coefficients for ADME properties of peptides,[41] the suitability of the SolvL scale was further checked for reproducing the differences in log P_N/log $D_{7.4}$ of a set of random peptides.[42] The SolvL-based additive scheme yielded promising results, as noted in the r values of 0.93 and 0.83 in reflecting the experimental range of log P_N and log $D_{7.4}$ for sets of 118 and 116 peptides, respectively (Figure 4C,D). Compared with experimental scales, a similar predictive

power was attained for Fauchère–Pliska and Eisenberg–McLachlan scales ($r \approx 0.90$) for the set of 118 log P_N data and for Hopp–Woods ($r \approx 0.88$) for the set of 116 log $D_{7.4}$ values but with a larger mue (~2.3 versus 0.7 for the SolvL scale; Tables S4 and S5).

In these test cases, the ProtL scale performed worse ($0.60 < r < 0.91$; Figure S6) than the SolvL one, suggesting that the Boltzmann-weighting scheme is better suited for describing the lipophilicity of residues in structureless peptides. However, one might expect an improved performance of the ProtL scale in the analysis of the lipophilic complementarity in peptide–protein and protein–protein complexes. To this end, we have examined the relationship between the ProtL cumulative lipophilicity and the experimental binding free energies of 19 peptides to MHC (HLA-A*02:01 allele) proteins (Table S6). These peptides were chosen subject to the availability of (i) precise structural information on the peptide–protein complex in the Protein Data Bank[43] and (ii) an estimate of the binding affinity in the Immune Epitope Database and Analysis Resource[44] (Table S6). The cumulative lipophilicity was determined taking into account the fraction of solvent-exposed area of the peptide residues in the MHC complex, supplemented with two correction parameters that account for the contribution due to the involvement of the backbone in hydrogen bonds[45] and to the burial of apolar residues from water to hydrophobic environments[30] (eq S4 in the SI Computational Methods).

The results show that the ProtL scale works better than the SolvL scale (correlation coefficients of 0.58 and 0.42, respectively; Figure 5) when the whole set of 19 peptides is considered, yielding correlation coefficients that are comparable to Moon–Fleming and Eisenberg–McLachlan scales (r of 0.61 and 0.51, respectively; Table S7). This correlation is remarkable keeping in mind the heterogeneity of the peptides and the uncertainty arising from the combination of data taken from different studies and determined using distinct experimental approaches. Furthermore, a significant improvement is observed upon the exclusion of the two Cys-containing peptides (PDB codes 3MRG and 2PYE), perhaps reflecting a quenching effect of cysteine in fluorescence assays.[46,47] Thus upon exclusion, the correlation coefficient of ProtL and SolvL scales increases up to 0.80 and 0.73, respectively, leading to regression equations with increased statistical significance (p values of 2×10^{-4} and 2×10^{-3}, respectively). Finally, let us note that this improvement outperforms the results obtained with the experimental scales ($r < 0.67$; Table S7).

Overall, the results point out the versatility of the SolvL/ProtL scales to examine the relationships between the lipophilicity and physicochemical properties of peptides under different pH conditions. From a methodological point of view, the strategy relies on the combination of an accurately parametrized version of continuum solvation models with an elaborate formalism to derive distribution coefficients from the partition of neutral and ionic species, in conjunction with the pK_a of ionizable groups. The simplicity of the computational strategy and the low cost of the required calculations permit a straightforward extension to nonstandard residues, such as the effect of chemical modifications on the lipophilicity maps of proteins, thus providing valuable information to explore biomolecular recognition and to modulate the properties of engineered polymeric materials.

ASSOCIATED CONTENT

ⓈSupporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpclett.9b00028.

> Detailed description of the computational strategy, tables and figures showing complementary information about the SolvL and ProtL scales, and their application to several test systems (PDF)

AUTHOR INFORMATION

Corresponding Authors
*J.M.C.: E-mail: campaxic@gmail.com.
*F.J.L.: E-mail: fjluque@ub.edu.
ORCID
William J. Zamora: 0000-0003-4029-4528
Josep M. Campanera: 0000-0002-6698-874X
F. Javier Luque: 0000-0002-8049-3567
Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

REFERENCES

(1) Tanford, C. The Hydrophobic Effect and the Organization of Living Matter. *Science* **1978**, *200*, 1012−1018.
(2) Ben-Amotz, D. Water-Mediated Hydrophobic Interactions. *Annu. Rev. Phys. Chem.* **2016**, *67*, 617−638.
(3) Simm, S.; Einloft, J.; Mirus, O.; Schleiff, E. 50 Years of Amino Acid Hydrophobicity Scales: Revisiting the Capacity for Peptide Classification. *Biol. Res.* **2016**, *49*, 31.
(4) Peters, C.; Elofsson, A. Why is the Biological Hydrophobicity Scale More Accurate than Earlier Experimental Hydrophobicity Scales? *Proteins: Struct., Funct., Genet.* **2014**, *82*, 2190−2198.
(5) MacCallum, J. L.; Tieleman, D. P. Hydrophobicity Scales: A Thermodynamic Looking Glass into Lipid-Protein Interactions. *Trends Biochem. Sci.* **2011**, *36*, 653−662.
(6) Zamora, W. J.; Curutchet, C.; Campanera, J. M.; Luque, F. J. Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus−Scrocco−Tomasi Continuum Solvation Calculations. *J. Phys. Chem. B* **2017**, *121*, 9868−9880.
(7) Fauchère, J. L.; Pliska, V. Hydrophobic Parameters Pi of Amino Acid Side Chains from the Partitioning of N-Acetyl-Amino Acid Amides. *Eur. J. Med. Chem.* **1983**, *18*, 369−375.
(8) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct.: THEOCHEM* **2005**, *727*, 29−40.
(9) Cancès, E.; Mennucci, B.; Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032−3041.
(10) Mennucci, B. Polarizable Continuum Model. *WIRES Comput. Mol. Sci.* **2012**, *2*, 386−404.

888

200

(11) Campanera, J. M.; Barril, X.; Luque, F. J. On the Transferability of Fractional Contributions to the Hydration Free Energy of Amino Acids. *Theor. Chem. Acc.* **2013**, *132*, 1−14.

(12) Dunbrack, R. L.; Karplus, M. Backbone-Dependent Rotamer Library for Proteins: Application to Side-Chain Prediction. *J. Mol. Biol.* **1993**, *230*, 543−574.

(13) Dunbrack, R. L.; Karplus, M. Conformational Analysis of the Backbone-Dependent Rotamer Preferences of Protein Sidechains. *Nat. Struct. Mol. Biol.* **1994**, *1*, 334−340.

(14) Shapovalov, M. V.; Dunbrack, R. L., Jr. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19*, 844−858.

(15) Arnold, M. R.; Kremer, W.; Lüdemann, H. D.; Kalbitzer, H. R. 1H-NMR Parameters of Common Amino Acid Residues Measured in Aqueous Solutions of the Linear Tetrapeptides Gly-Gly-X-Ala at Pressures between 0.1 and 200 MPa. *Biophys. Chem.* **2002**, *96*, 129−140.

(16) Kortemme, T.; Creighton, T. E. Ionisation of Cysteine Residues at the Termini of Model α-Helical Peptides. Relevance to Unusual Thiol pKaValues in Proteins of the Thioredoxin Family. *J. Mol. Biol.* **1995**, *253*, 799−812.

(17) Harms, M. J.; Schlessman, J. L.; Sue, G. R.; Garcia-Moreno E., B. Arginine Residues at Internal Positions in a Protein Are Always Charged. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 18954−18959.

(18) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A Summary of the Measured pKa Values of the Ionizable Groups in Folded Proteins. *Protein Sci.* **2008**, *18*, 247−251.

(19) Koehler, J.; Woetzel, N.; Staritzbichler, R.; Sanders, C. R.; Meiler, J. A Unified Hydrophobicity Scale for Multispan Membrane Proteins. *Proteins: Struct., Funct., Genet.* **2009**, *76*, 13−29.

(20) Musafia, B.; Buchner, V.; Arad, D. Complex Salt Bridges in Proteins: Statistical Analysis of Structure and Function. *J. Mol. Biol.* **1995**, *254*, 761−770.

(21) Tomlinson, J. H.; Ullah, S.; Hansen, P. E.; Williamson, M. P. Characterization of Salt Bridges to Lysines in the Protein G B1 Domain. *J. Am. Chem. Soc.* **2009**, *131*, 4674−4684.

(22) Isom, D. G.; Castañeda, C. A.; Cannon, B. R.; García-Moreno E., B. Large Shifts in pKa Values of Lysine Residues Buried Inside a Protein. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 5260−5265.

(23) André, I.; Linse, S.; Mulder, F. A. A. Residue-Specific pKa Determination of Lysine and Arginine Side Chains by Indirect ^{15}N and ^{13}C NMR Spectroscopy: Application to *apo* Calmodulin. *J. Am. Chem. Soc.* **2007**, *129*, 15805−15813.

(24) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378−6396.

(25) ACD/I-Lab. Advanced Chemistry Development, Inc.: Toronto, ON, Canada. http://www.acdlabs.com.

(26) ChemAxon, Budapest, Hungary. http://www.chemaxon.com.

(27) Eisenberg, D.; McLachlan, A. D. Solvation Energy in Protein Folding and Binding. *Nature* **1986**, *319*, 199−203.

(28) Hopp, T. P.; Woods, K. R. Prediction of Protein Antigenic Determinants from Amino Acid Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 3824−3828.

(29) Wimley, W. C.; Creamer, T. P.; White, S. H. Solvation Energies of Amino Acid Side Chains and Backbone in a Family of Host-Guest Pentapeptides. *Biochemistry* **1996**, *35*, 5109−5124.

(30) Moon, C. P.; Fleming, K. G. Side-Chain Hydrophobicity Scale Derived from Transmembrane Protein Folding into Lipid Bilayers. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10174−10177.

(31) Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S.; von Heijne, G. Recognition of Transmembrane Helices by the Endoplasmic Reticulum Translocon. *Nature* **2005**, *433*, 377−381.

(32) Janin, J. Surface and inside Volumne in Globular Proteins. *Nature* **1979**, *277*, 491−492.

(33) Kyte, J.; Doolittle, R. F. A Simple Method for Displaying the Hydropathic Character of a Protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(34) Wimley, W. C.; Gawrisch, K.; Creamer, T. P.; White, S. H. Direct Measurement of Salt-Bridge Solvation Energies Using a Peptide Model System: Implications for Protein Stability. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 2985−2990.

(35) Wilce, M. C. J.; Aguilar, M. I.; Hearn, M. T. W. Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Anal. Chem.* **1995**, *67*, 1210−1219.

(36) Biswas, K. M.; DeVido, D. R.; Dorsey, J. G. Evaluation of Methods for Measuring Amino Acid Hydrophobicities and Inter-actions. *J. Chromatogr. A* **2003**, *1000*, 637−655.

(37) König, G.; Bruckner, S.; Boresch, S. Absolute Hydration Free Energies of Blocked Amino Acids: Implications for Protein Solvation and Stability. *Biophys. J.* **2013**, *104*, 453−462.

(38) Amrhein, S.; Oelmeier, S. A.; Dismer, F.; Hubbuch, J. Molecular Dynamics Simulations Approach for the Characterization of Peptides with Respect to Hydrophobicity. *J. Phys. Chem. B* **2014**, *118*, 1707−1714.

(39) Houghten, R. A.; Degraw, S. T. Effect of Positional Environmental Domains on the Variation of High-Performance Liquid Chromatographic Peptide Retention Coefficients. *J. Chromatogr.* **1987**, *386*, 223−228.

(40) Reimer, J.; Spicer, V.; Krokhin, O. V. Application of Modern Reversed-Phase Peptide Retention Prediction Algorithms to the Houghten and DeGraw Dataset: Peptide Helicity and Its Effect on Prediction Accuracy. *J. Chromatogr. A* **2012**, *1256*, 160−168.

(41) Fosgerau, K.; Hoffmann, T. Peptide Therapeutics: Current Status and Future Directions. *Drug Discovery Today* **2015**, *20*, 122−128.

(42) Buchwald, P.; Bodor, N. Octanol-Water Partition of Non-zwitterionic Peptides: Predictive Power of a Molecular Size-Based Model. *Proteins: Struct., Funct., Genet.* **1998**, *30*, 86−99.

(43) Rose, P. W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Di Costanzo, L.; Duarte, J. M.; Dutta, S.; Feng, Z.; et al. The RCSB Protein data Bank: Integrative View of Protein, Gene and 3D Structural Information. *Nucleic Acids Res.* **2016**, *45*, D271−D281.

(44) Vita, R.; Mahajan, S.; Overton, J. A.; Dhanda, S. K.; Martini, S.; Cantrell, J. R.; Wheeler, D. K.; Sette, A.; Peters, B. The Immune Epitope Database (IEDB): 2018 Update. *Nucleic Acids Res.* **2019**, *47*, D339.

(45) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.

(46) Chen, Y.; Barkley, M. D. Toward Understanding Tryptophan Fluorescence in Proteins. *Biochemistry* **1998**, *37*, 9976−9982.

(47) D'Auria, S.; Staiano, M.; Kuznetsova, I.; Turoverov, K. K. The Combined Use of Fluorescence Spectroscopy and X-Ray Crystallography Greatly Contributes to Elucidating Structure and Dynamics of Proteins. In *Reviews in Fluorescence 2005*; Geddes, C. D., Lakowicz, J. R., Eds.; Springer: Boston, 2005; pp 25−61.

Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations

William J. Zamora, Josep Maria Campanera[*], F. Javier Luque[*]

Index

COMPUTATIONAL METHODS

SolvL and ProtL lipophilicity scales.

Following a previous study on the hydration free energy of the natural amino acids,[S1] the
N-acetyl-L-amino acid amides (CH_3-CO-NH-CHR-CONH$_2$) were chosen as molecular
models. Using the backbone-dependent conformational library reported by Dunbrack and
coworkers,[S2-S4] a total of 572 rotamers (i.e., conformers with a probability contribution
higher than 5% to the total conformational space of each residue) were compiled. These
structures were then used to compute the n-octanol/water transfer free energies, which
were performed with the B3LYP/6-31G(d) MST[S5] version of the IEF-PCM[S6] model.

Computation of the distribution coefficients at a given pH (log D_{pH}) was performed by
combining the partition coefficient of neutral and ionic species (for ionizable residues)
using Eq. S1.

$$\log D = \log\left(P_N + P_I \cdot 10^\delta\right) - \log(1 + 10^\delta) \qquad (S1)$$

where P_N and P_I denote the partition coefficient of the neutral and ionized species of the
amino acid, and δ is the difference between the pK_a of the ionizable group and the pH of
the environment.

Let us note that Eq. S1 represents one of the formalisms considered to estimate the pH-
dependent lipophilicity profile of small (bio)organic compounds,[S7] and was found to
reproduce satisfactorily the change in pH-dependent distribution coefficients for amino
acid analogues.

The contribution of the conformational species in water and n-octanol was accounted for

considering two weighting schemes, giving rise to the Solvent-like (SolvL) and Protein-like (ProtL) lipophilicities scales, respectively.

(i) In the SolvL scale, the contribution of each conformational state to the partition coefficient of the neutral/ionized species was determined using a Boltzmann weighting scheme, where the effective free energy was estimated by combining the internal energy of the conformer and its solvation free energy in water and n-octanol. To this end, the geometry of all rotamers was optimized at the B3LYP/6-31G(d) level of theory while keeping the backbone dihedrals fixed to the torsional values of the Dunbrack's library, and subsequently single-point calculations in the gas phase and in solution. The log D_{pH} was then computed using Eq. 1, adopting the pK_a values reported for ionizable residues from experimental peptide models in aqueous solutions.[S8,S9]

(ii) In the ProtL scale, the contribution of each conformation to the partition between the two solvents was determined by using the weights reported in the Dunbrack`s library, which reflect the rotameric distribution in a protein environment. The pK_as of ionizable residues were taken from values in folded proteins.[S10,S11]

For the sake of comparison, we also computed both approaches with the SMD model using the B3LYP/6-31G(d) level of theory.[S12] All calculations were performed using a locally modified version of Gaussian 09.[S13]

Comparison with experimental hydrophobicity scales.

Due to the diversity of experimental lipophilicity scales of amino acids, generally expressed in terms of transfer free energies, comparison was made by converting them to partition/distribution coefficients, which were subsequently normalized to Gly following Eq. S2.

$$\log P_{\mathrm{N}} / D_{\mathrm{pH}} = \frac{(-\Delta\Delta G_{\mathrm{transf,AA}} - \Delta\Delta G_{\mathrm{transf,Gly}})}{RT\ln 10} \qquad (S2)$$

where $\Delta\Delta G_{\mathrm{transf,AA}}$ is the transfer free energy of a given amino acid from the aqueous phase to the organic/biological environment, and $\Delta\Delta G_{\mathrm{transf,Gly}}$ is the transfer free energy of Gly.

Determination of the cumulative lipophilicity.

Most of the experimental scales present in the literature compute the lipophilicity of a given peptide as the sum of individual lipophilicity of the constituent amino acids relative to a reference residue, usually Gly or Ala. Since the MST solvation model gives atomic contributions to the transfer free energy,[S14-S16] we can separate the global lipophilicity in contributions corresponding to the backbone (*bb*), side-chain (*sc*), and the capping groups (*cg*). Combination of the *bb* and *sc* contributions yields the amino acid lipophilicity (reported in Table 1 in the manuscript), whereas the contribution of the capping groups has been estimated to be (N-terminus) CH_3CO- (log P_{N} = 0.20), NH_3^+- (log $D_{7.4}$ = -2.99), and (C-terminus) NH_2- (log P_{N} = -1.08), $NMe-$ (log P_{N} = 0.35), COO^-- (log $D_{7.4}$ = -4.89).

The cumulative lipophilicity of a peptide with N_{res} residues may be estimated by using Eq. S3.

$$\log(P_N / D_{pH})^{peptide} = \sum_{i=1}^{Nres} \log(P_N^i / D_{pH}^i)^{bb+sc} + \sum_{i=1}^{Ncg} \log(P_N^i / D_{pH}^i)^{cg} \qquad (S3)$$

where P_N^i / D_{pH}^i stands for the fragment (*bb*+*sc* or *cg*) partition/distribution coefficient, N_{res} and N_{cg} being the total number of residues and capping groups in the peptide.

For practical applications, this simple expression is convenient when there is no explicit knowledge about the 3D structure of peptides, as may occur in structureless peptides.

S4

For our purposes here, this is the expression adopted to evaluate the lipophilicity of small, flexible peptides in solution.

On the other hand, if the 3D structure of the peptide is known from experimental (X-ray, NMR) or computational (Molecular Dynamics) approaches, then the cumulative lipohilicity may be estimated taking into account the specific structural features of peptides/proteins, as noted in Eq. S4.

$$\log(P_N / D_{pH})^{peptide} = \sum_{i=1}^{Nres} \left(\lambda^i \cdot \log(P_N^i / D_{pH}^i)^{bb+sc} + \lambda^i \cdot \log(P_N^i / D_{pH}^i)^{cg} + \alpha^i + \beta^i \right) \ (S4)$$

In Eq. S4, λ^i stands for the fraction of solvent-exposed surface area (SASA) of the amino acid (*bb+sc*) or capping group (*cg*) according to the local structural environment of in a peptide/protein. For our purposes, the SASA was determined using NACCESS.[S17]

In addition, two correction factors were also introduced. The parameter α^i introduces a correction to the hydrophobic contribution when the backbone participates in a hydrogen bond (HB). This contribution can be estimated to amount, on average, to 0.73 (log *P* units) per HB.[S18] The occurrence of this kind of HBs in a given 3D structural model was determined with the DSSP program.[S19] Finally, the β^i factor accounts for a correction due to the burial of the side chain of hydrophobic residues (Ala, Leu, Ile, Val, Pro, Phe, Trp, Met and Tyr) from water to a lipophilic environment. This contribution has been estimated to be 0.023 kcal mol^{-1} Å^{-2} according to the studies reported by Moon and Fleming for the transfer of nonpolar side chains from water into a lipid bilayer.[S20] Therefore, the β^i term has been estimated from the fraction of the buried side chain with respect to the fully buried side chain, as noted in Eq. S5.

$$\beta^i = H_{res}^i \cdot (1 - \lambda^i)^{sc} \qquad\qquad (S5)$$

where H_{res}^i stands for the hydrophobic contribution (in log$P$ units) of a specific apolar residue, which was estimated as noted in Eq. S6.

$$H_{res}^i = \frac{0.023 \cdot SASA_{res}^{sc}}{2.303RT} \qquad\qquad (S6)$$

where $SASA_{res}^{sc}$ is the average SASA of a given residue type, R is the gas constant, and T is temperature.

The H_{res}^i values for nonpolar residues are given in Table S0.

Table S0. Average solvent accessible surface area for the side-chain of the hydrophobic residues and the hydrophobic effect contribution value when the side chain is fully buried.

Residue	Average SASA ($\mathring{A}^2$)	H_{res}^i (log P units)
Ala	69	1.2
Val	130	2.2
Leu	158	2.7
Ile	157	2.6
Met	166	2.8
Pro	115	1.9
Phe	188	3.2
Trp	232	3.9
Tyr	201	3.4

S6

References

(S1) Campanera, J. M.; Barril, X.; Luque, F. J. On the Transferability of Fractional Contributions to the Hydration Free Energy of Amino Acids. *Theor. Chem. Acc.* **2013**, *132*, 1–14.

(S2) Dunbrack, R. L.; Karplus, M. Backbone-Dependent Rotamer Library for Proteins: Application to Side-Chain Prediction. *J. Mol. Biol.* **1993**, *230*, 543–574.

(S3) Dunbrack, R. L.; Karplus, M. Conformational Analysis of the Backbone-Dependent Rotamer Preferences of Protein Sidechains. *Nat. Struct. Biol.* **1994**, *1*, 334–340.

(S4) Shapovalov, M. V.; Dunbrack, R. L. A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure* **2011**, *19*, 844–858.

(S5) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Javier Luque, F. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct. THEOCHEM* **2005**, *727*, 29–40.

(S6) Cances, E.; Mennucci, B.; Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032–3041.

(S7) Zamora, W. J.; Curutchet, C.; Campanera, J. M.; Luque, F. J. Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus–Scrocco–Tomasi Continuum Solvation Calculations. *J. Phys. Chem. B* **2017**, *121*, 9868–9880.

(S8) Arnold, M. R.; Kremer, W.; Lüdemann, H. D.; Kalbitzer, H. R. 1H-NMR Parameters of Common Amino Acid Residues Measured in Aqueous Solutions of the Linear Tetrapeptides Gly-Gly-X-Ala at Pressures between 0.1 and 200 MPa. *Biophys. Chem.* **2002**, *96*, 129–140.

(S9) Kortemme, T.; Creighton, T. E. Ionisation of Cysteine Residues at the Termini of Model α-Helical Peptides. Relevance to Unusual Thiol pK_a Values in Proteins of the Thioredoxin Family. *J. Mol. Biol.* **1995**, *253*, 799–812.

(S10) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A Summary of the Measured pK_a

Values of the Ionizable Groups in Folded Proteins. *Protein Sci.* **2009**, *18*, 247–251.

(S11) Harms, M. J.; Schlessman, J. L.; Sue, G. R.; Garcia-Moreno E., B. Arginine Residues at Internal Positions in a Protein Are Always Charged. *Proc. Natl. Acad. Sci.* **2011**, *108*, 18954–18959.

(S12) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(S13) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et Al. Gaussian 09, Revision D.01; Gaussian, Inc.: Wallingford CT, 2009.

(S14) Luque, F. J.; Curutchet, C.; Muñoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpí, J. L.; Orozco, M. Continuum Solvation Models: Dissecting the Free Energy of Solvation. *Phys. Chem. Chem. Phys*. **2003**, *5*, 3827–3836.

(S15) Ginex, T.; Muñoz-Muriedas, J.; Herrero, E.; Gibert, E.; Cozzini, P.; Luque, F. J. Development and Validation of Hydrophobic Molecular Fields Derived from the Quantum Mechanical IEF/PCM-MST Solvation Model in 3D-QSAR. *J. Comput. Chem*. **2016**, *37*, 1147-1162.

(S16) Vázquez, J.; Deplano, A.; Herrero, A.; Ginex, T.; Gibert, E.; Rabal, O.; Oyarzabal, J.; Herrero, E.; Luque, F. J. Development and Validation of Molecular Overlays Derived from Three-Dimensional Hydrophobic Similarity with PharmScreen. *J. Chem. Inf Model*. **2018**, *58*, 1596–1609.

(S17) Hubbard S.; Thornton, J. M. *NACCESS V.2.1.1* **1993**; http://wolf.bms.umist.ac.uk/naccess.

(S18) Pace, C. N.; Fu, H.; Fryar, K. L.; Landua, J.; Trevino, S. R.; Schell, D.; Thurlkill, R. L.; Imura, S.; Scholtz, J. M.; Gajiwala, K.; et al. Contribution of Hydrogen Bonds to Protein Stability. *Protein Sci.* **2014**, *23*, 652–661.

(S19) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.

(S20) Moon, C. P.; Fleming, K. G. Side-Chain Hydrophobicity Scale Derived from Transmembrane Protein Folding into Lipid Bilayers. *Proc. Natl. Acad. Sci.* **2011**, *108*, 10174–10177.

S8

Table S1. Protein-like (ProtL) Lipophilicity Scale Based on the $\log D_{\text{pH}}$ Values Determined for *N*-Acetyl-*L*-Amino Acid Amides at Physiological pH. The Lipophilicity Obtained for Conformational Distributions in α-Helix and β-Sheet Structures, the Experimental pK_a of Side Chain Ionizable Groups, and the Calculated Partition Coefficients of Neutral ($\log P_N$) and Ionized ($\log P_I$) Residues Are Also Given.

Residues	Exp. pK_a	$\log P_N$	$\log P_I$	$\log D_{7.4}$
ALA	**-**	**-2.47**	**-**	**-2.47**
α-helix	-	*-2.87*	-	*-2.87*
β-sheet	-	*-2.03*	-	*-2.03*
ARG		**-3.66**	**-7.38**	**-7.04**
α-helix	**12.51**	*-3.75*	*-8.09*	*-7.59*
β-sheet		*-3.49*	*-5.98*	*-5.98*
ASN	**-**	**-3.97**	**-**	**-3.97**
α-helix	-	*-4.09*	-	*-4.09*
β-sheet	-	*-3.39*	-	*-3.39*
ASP		**-3.18**	**-8.54**	**-5.87**
α-helix	**3.50**	*-3.26*	*-7.37*	*-5.63*
β-sheet		*-3.07*	*-10.07*	*-6.19*
CYS		**-1.47**	**-5.78**	**-2.17**
α-helix	**6.80**	*-2.06*	*-5.75*	*-2.76*
β-sheet		*-1.09*	*-5.81*	*-1.78*
GLN	**-**	**-4.00**	**-**	**-4.00**
α-helix	-	*-5.00*	-	*-5.00*
β-sheet	-	*-1.64*	-	*-1.64*
GLU		**-3.79**	**-6.20**	**-5.96**
α-helix	**4.20**	*-3.67*	*-6.42*	*-6.14*
β-sheet		*-4.03*	*-5.76*	*-5.58*
GLY	**-**	**-3.13**	**-**	**-3.13**
HID		**-4.67**	**-5.97**	**-4.56**
α-helix	**6.60**	*-5.12*	*-6.16*	*-5.00*
β-sheet		*-4.26*	*-5.79*	*-4.15*
HIE		**-4.98**	**-5.97**	**-4.97**
α-helix	**6.60**	*-5.49*	*-6.16*	*-5.46*
β-sheet		*-4.49*	*-5.79*	*-4.52*
ILE	**-**	**-0.38**	**-**	**-0.38**
α-helix	-	*-0.55*	-	*-0.55*
β-sheet	-	*-0.24*	-	*-0.24*
LEU	**-**	**-1.36**	**-**	**-1.36**
α-helix	-	*-1.59*	-	*-1.59*
β-sheet	-	*-1.09*	-	*-1.09*
LYS		**-2.19**	**-6.81**	**-5.08**
α-helix	**10.53**	*-2.32*	*-7.18*	*-5.29*
β-sheet		*-1.98*	*-6.16*	*-4.73*

MET	**-**	**-1.83**	**-**	**-1.83**
α-helix	-	*-2.06*	-	*-2.06*
β-sheet	-	*-1.56*	-	*-1.56*
PHE	**-**	**0.86**	**-**	**0.86**
α-helix	-	*2.23*	-	*2.23*
β-sheet	-	*-0.18*	-	*-0.18*
PRO	**-**	**-1.44**	**-**	**-1.44**
α-helix	-	*-1.42*	-	*-1.42*
β-sheet	-	*-1.45*	-	*-1.45*
SER	**-**	**-4.12**	**-**	**-4.12**
α-helix	-	*-3.21*	-	*-3.21*
β-sheet	-	*-4.92*	-	*-4.92*
THR	**-**	**-3.01**	**-**	**-3.01**
α-helix	-	*-3.33*	-	*-3.33*
β-sheet	-	*-2.80*	-	*-2.80*
TRP	**-**	**0.16**	**-**	**0.16**
α-helix	-	*0.51*	-	*0.51*
β-sheet	-	*-0.10*	-	*-0.10*
TYR		**-1.80**	**-9.59**	**-1.80**
α-helix	**10.33**	*-1.96*	*-9.65*	*-1.96*
β-sheet		*-1.69*	*-9.55*	*-1.69*
VAL	**-**	**-1.68**	**-**	**-1.68**
α-helix	-	*-2.19*	-	*-2.19*
β-sheet		*-1.38*	-	*-1.38*

S10

Table S2. Experimental RP-HPLC Retention Time for Eight Model Decapeptides and Cumulative Hydrophobicity Determined with the SolvL and ProtL Lipophilicity Scales.

Peptide [a]	Sequence	Retention factor k` (min)	log $D_{7.4}$	
			SolvL	**ProtL**
Pep1Leu	DKDKGGGGLG	4.80	-17.09	-34.04
Pep2Leu	DKDKGGGLLG	11.97	-15.03	-32.27
Pep3Leu	DKDKGGLLLG	16.22	-12.97	-30.50
Pep1Cys	DKDKGGGGCG	0.52	-17.30	-34.85
Pep1Ile	DKDKGGGGIG	4.73	-17.64	-33.06
Pep1Met	DKDKGGGGMG	2.27	-17.65	-34.51
Pep1Phe	DKDKGGGGFG	6.11	-16.53	-31.82
Pep1Val	DKDKGGGLVG	1.86	-18.07	-34.36

[a] Ref. 38.

Table S3. Correlation of Retention Time for Eight Model Decapeptides with the Same Charge,[38] and for 218 Peptides[39,40] with Three Different Charge States Using the Cumulative Hydrophobicity with Our Adaptive Hydrophobicity Scale and with Others Experimental Scales.

Scale	r p-value[a]	
	Ref. 38 (pH = 7.4)	**Refs. 39,40 (pH = 2.1)**
Fauchère-Pliska	0.96 2×10^{-4}	0.85 $< 1 \times 10^{-16}$
Eisenberg-McLachlan	0.95 3×10^{-4}	0.79 $< 1 \times 10^{-16}$
Hopp-Woods	0.99 7×10^{-6}	0.74 $< 1 \times 10^{-16}$
Wimley et al.	0.99 4×10^{-7}	0.36 4×10^{-9}
Moon-Fleming	0.99 3×10^{-6}	0.78 $< 1 \times 10^{-16}$
Hessa et al.	0.96 2×10^{-4}	0.61 $< 1 \times 10^{-16}$
Koehler et al.	0.76 0.03	0.64 $< 1 \times 10^{-16}$
Janin et al.	0.39 0.3	0.55 $< 1 \times 10^{-16}$
Kyte-Doolittle	0.93 8×10^{-4}	0.60 $< 1 \times 10^{-16}$
SolvL	0.96 2×10^{-4}	0.85 $< 1 \times 10^{-16}$
ProtL	0.91 1.8	0.80 $< 1 \times 10^{-16}$

[a] r: Pearson correlation coefficient, p: statistical p-value.

Table S4. Statistical Parameters of the Comparison[a] of the SolvL and ProtL Scale with Others Hydrophobicity Scales Against log P_N Values for 118 Random Peptides.

Scale	r p-value	mse	mue	rsmd
Fauchère-Pliska	0.90 $< 1 \times 10^{-16}$	-2.53	2.53	2.64
Eisenberg-McLachlan	0.89 $< 1 \times 10^{-16}$	-2.29	2.29	2.38
Hopp-Woods	0.74 $< 1 \times 10^{-16}$	-2.07	2.11	2.31
Wimley et al.	0.70 $< 1 \times 10^{-16}$	-1.54	1.67	1.81
Moon-Fleming	0.69 $< 1 \times 10^{-16}$	-0.80	1.12	1.34
Hessa et al.	0.22 0.02	0.29	0.98	1.29
Koehler et al.	0.45 3×10^{-7}	-0.35	0.87	1.12
Janin et al.	0.38 2×10^{-5}	-0.65	1.08	1.28
Kyte-Doolittle	0.50 6×10^{-9}	-2.85	3.00	3.60
ProtL	0.60 5×10^{-13}	1.35	1.68	2.00
SolvL	0.93 $< 1.0 \times 10^{-16}$	-0.55	0.71	0.94

[a] mse: mean signed error, mue: mean unsigned error, rmsd: root-mean square deviation, r: Pearson correlation coefficient, p: statistical p-value. mse, mue and rsmd are given in log P_N/D units.

S12

Table S5. Statistical Parameters of the Comparison[a] of the SolvL and ProtL Scale with Others Hydrophobicity Scales Against log $D_{7.4}$ Values for 116 Random Peptides.

Scale	r / p-value	mse	mue	rsmd
Fauchère-Pliska	0.76 $< 1 \times 10^{-16}$	-2.76	2.76	2.88
Eisenberg-McLachlan	0.75 $< 1 \times 10^{-16}$	-2.58	2.58	2.69
Hopp-Woods	0.88 $< 1 \times 10^{-16}$	-2.32	2.33	2.43
Wimley et al.	0.52 2×10^{-9}	-1.94	1.94	2.23
Moon-Fleming	0.79 $< 1 \times 10^{-16}$	-1.16	1.24	1.48
Hessa et al.	0.72 $< 1 \times 10^{-16}$	-0.22	0.60	0.73
Koehler et al.	0.76 $< 1 \times 10^{-16}$	-0.90	1.01	1.19
Janin et al.	0.61 4×10^{-13}	-1.12	1.21	1.38
Kyte-Doolittle	0.52 2×10^{-9}	3.04	3.17	3.76
ProtL	0.79 $< 1 \times 10^{-16}$	1.46	1.82	2.11
SolvL	0.83 $< 1 \times 10^{-16}$	-0.52	0.73	0.95

[a] mse: mean signed error, mue: mean unsigned error, rmsd: root-mean square deviation, r: Pearson correlation coefficient, p: statistical p-value. mse, mue and rmsd are given in log P_N/D units.

Appendix III. Book Chapter:

Implicit Solvation Methods in the Study of Ligand-Protein Interactions.

Zamora, W. J. Campanera, J. Luque, F. **(2015)**; Implicit Solvation Methods in the Study of Ligand-Protein Interactions. In C. Cavasotto (Ed.), *In silico Drug Discovery and Design:Theory, Methods, Challenges, and Applications.* by CRC Press.

9

Implicit Solvation Methods in the Study of Ligand–Protein Interactions

William Zamora, Josep M. Campanera, and F. Javier Luque

CONTENTS

9.1 Ligand–Receptor Interaction

The affinity between a small compound and its macromolecular target can be related to macroscopic observables through the laws of thermodynamics. Thus, the binding affinity can be expressed in terms of the equilibrium constant (K) for the formation of the ligand–receptor complex, which can be related to the difference in the standard Gibbs free energy between bound and unbound states (ΔG°; Equation 9.1).

$$\Delta G^\circ = -RT \ln K \qquad (9.1)$$

where R is the gas constant and T is the temperature.

The binding affinity reflects a subtle balance between a number of separate enthalpic and entropic contributions (Gohlke and Klebe 2002; Bissantz et al. 2010). The structural and chemical complementarity between the functional groups that are present at the binding interface renders the net stabilizing energy that is required to compensate unfavorable contributions to the binding. Thus, the binding between ligand and receptor is often accompanied by conformational changes, which can encompass a range of potential scenarios

such as the "induced fit" mechanism, the "conformational selection" process, or even more complex models that combine the selection of specific conformations with the induction of structural readjustments upon binding (Csermely et al. 2010; Spyrakis et al. 2011). Predicting the energy cost associated with conformational changes in the ligand has proved to be very challenging, as noted by the uncertainties associated with the choice of the level of theory used to determine the cost of selecting the bioactive conformation (Tirado-Rives and Jorgensen 2006; Butler et al. 2009).

The energy gain as a result of the seemingly favorable interactions formed in the complex must counterbalance the cost due to dehydration of the separate partners prior to their mutual interaction. For simple neutral organic compounds, the hydration-free energies are generally in a narrow range, as noted in the experimental values for the transfer from gas phase to water for compounds that mimic the side chain of noncharged amino acids, which vary from +2 to –11 kcal/mol (Table 9.1; Wolfenden et al. 1981). However, the hydration-free energy of charged compounds is much larger, as expected from the strengthening of the interactions with water molecules, leading to hydration-free energies of –77 kcal/mol for acetate anion and –71 kcal/mol for the protonated *n*-butylamine (Pliego and Riveros 2002). Hence, there must be a sizable compensation between the dehydration energy cost and binding site residues and the energy gain triggered upon burial of the ligand in the binding pocket.

Finally, the ligand–receptor interactions must also compensate for the entropy changes arising upon molecular association, such as the loss of translational and rotational degrees of freedom, the reduction in the accessible states for internal rotations of both ligand and protein, and the reorganization of water molecules upon formation of the complex. This can be illustrated by the fact that binding of amprenavir to HIV protease is accompanied by a configurational entropy loss of 26.4 kcal/mol, which primarily

TABLE 9.1

Experimental Hydration-Free Energies (ΔG_{hyd}; kcal mol^{-1}) of Organic Compounds Chosen as Analogs of the Side Chains of Neutral Amino Acids

Residue	Side Chain Analog	ΔG_{hyd}	Residue	Side Chain Analogue	ΔG_{hyd}
Ala	Methane	2.0	Leu	Isobutane	2.3
Ile	Butane	2.1	Met	Methyl ethyl sulfide	−1.5
Val	Propane	2.0	Phe	Toluene	−0.9
Phe	*p*-Cresol	−6.1	Trp	Methylindole	−5.9
His	Methylimidazole	−10.3	Ser	Methanol	−5.1
Thr	Ethanol	−5.1	Cys	Methanethiol	−1.2
Asn	Acetamide	−9.7	Gln	Propionamide	−9.4
Asp	Acetic acid	−6.7	Glu	Propionic acid	−6.5
Lys	*N*-butylamine	−4.3	Arg	*N*-propylguanidine	−10.9

arises from narrowness of the energy wells of bound amprenavir relative to free ligand (Chang et al. 2007).

The net balance between enthalpic and entropic components leads to ligand–protein binding affinities that generally fall between 10^{-2} and 10^{-12} M (Gohlke and Klebe 2002). Unfortunately, small uncertainties in determining the magnitude of the different free energy components may have a drastic impact on the accuracy of the binding affinity (Williams et al. 2004; Reynolds and Holloway 2011). Thus, an error of 1.36 kcal/mol changes the predicted binding constant (at 298 K) by one order of magnitude. Predicting with chemical accuracy the binding free energy is a formidable challenge to current computational methods due to the magnitude of the separate contributions to the binding free energy, and the compensation between enthalpic and entropic terms. However, this is a fundamental ingredient for the success of drug discovery, especially keeping in mind that the maximal free energy contribution per non-hydrogen atom in a drug-like ligand amounts to ~−1.5 kcal/mol (higher values per atom are found in the case of metals, small anions, and ligands that form covalent bonds; Kuntz et al. 1999).

The aim of this chapter is to examine the use of implicit solvation models in the calculation of the binding affinity of ligand–receptor complexes. To this end, the chapter is divided into two major sections. The first is focused on the use of implicit solvation models in the context of classical force field methods, dealing specifically with molecular mechanics Poisson–Boltmann surface area (MM-PBSA) and its Generalized Born counterpart (MM-GBSA). Attention is paid to the details of the underlying formalism and to the different strategies undertaken in order to improve the accuracy of the predicted binding affinities. In the second section, a brief overview of the application of implicit solvation methods in the framework of quantum mechanics is given in order to highlight the progressive development of novel implementations and their application in drug discovery.

9.2 Molecular Mechanics and Implicit Solvation Models

Free energy perturbation (FEP) and thermodynamic integration (TI) are the most valuable computational methods for the prediction of binding affinities of small drug-candidate compounds (Brandsdal et al. 2003; Chipot and Pohorille 2007; Jorgensen 2009). These techniques rely on the alchemical transformation of ligands (or amino acid residues in the wild-type protein and a mutated variant) in two states, which correspond to the ligand free in solution, and the ligand bound to the receptor. This transformation is performed by means of a series of simulations carried out at intermediate points along the transition path that connects the Hamiltonians of the initial and final states. As noted by Michel and Essex (2010), it seems reasonable to expect

that free energy calculations cannot predict binding free energies more accurately than solvation-free energies, where the uncertainties obtained for small organic compounds are approximately 1 kcal mol^{-1} (see also Merz 2010).

These calculations can provide the missing links between the experimental binding affinities and the atomic details of the protein–ligand complexes. However, when there are substantial differences in the chemical scaffold of the ligands, which differ by large substituents, or even when drastic mutations occur between the native protein and the mutated variant (e.g., tryptophan to alanine), the reliability and chemical accuracy of these calculations can be affected by convergence problems due to numerical instabilities and the limited conformational sampling. Hence, reliable computational schemes for the systematic prediction of ligand binding and mutagenesis effects are the subject of intense research (Pitera and van Gunsteren 2002; Steinbrecher et al. 2007; Lawrenz et al. 2011; Boukharta et al. 2014).

The high computational cost of these techniques is primarily due to the large number of intermediate states that must be defined in the alchemical transformation, but also to the explicit treatment of the molecular environment. These factors can be alleviated by treating solvent effects only implicitly using continuum solvent methods, and by considering only the endpoint states in the free energy calculations. These approximations lead to the so-called endpoint, implicit solvent-free energy methods, which encompass MM-PBSA and MM-GBSA. The main advantage of these methods is the huge reduction in the computational cost, which enables the screening of large datasets of ligands against a common receptor in a reasonable time span. Thus, MM-PB(GB)SA has been widely used in solving a broad range of topics valuable in ligand–receptor interactions, and specifically in drug discovery, such as determining hot spots in ligand-binding pockets and protein–protein interfaces, rescoring of docking poses, estimating binding affinities, and evaluating the stability of macromolecular assemblies. Nevertheless, the simplified description of the molecular system can also affect the chemical accuracy in predicting both the binding pose and the binding affinity, which makes it necessary to carry out a rigorous calibration of these methods.

9.2.1 Methodological Formalism of MM-PB(GB)SA Methods

In MM-PB(GB)SA, the binding free energy between ligand and receptor (ΔG_{bin}) is determined by combining three terms (Figure 9.1): the gas-phase free energy (ΔG_{MM}), the solvation-free energy (ΔG_{sol}), and the change in the configurational entropy ($-T\Delta S$) upon binding (Equation 9.2).

$$\Delta G_{bin} = \Delta G_{MM} + \Delta G_{sol} - T\Delta S \qquad (9.2)$$

The gas-phase component is determined from the molecular mechanics energy of the molecule, including bonded and nonbonded terms as implemented in a given force field. If the configurational space of the bound state
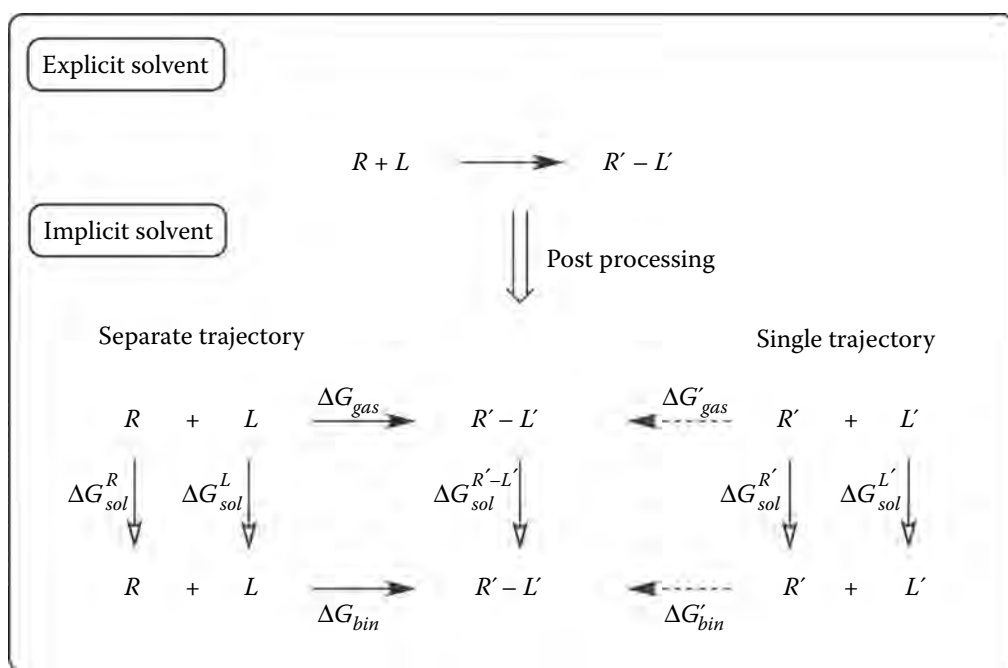
FIGURE 9.1

Thermodynamic cycle for the calculation of the binding affinity between ligand and receptor. Generally, MM-PB(GB)SA is used as a postprocessing method using representative snapshots taken from the trajectory sampled in a molecular dynamics simulation. In the single trajectory approach, ligand and receptor are taken from the snapshots sampled for the ligand-receptor complex (R'–L'). Other approaches use separate trajectories for receptor (R', R) and ligand (L', L). As noted in Equation 9.2, the binding affinity (ΔG_{bin}; $\Delta G'_{bin}$) combines the gas phase term (ΔG_{gas}; $\Delta G'_{gas}$), which combines the molecular mechanics (ΔG_{MM}) and entropic ($-T\Delta S$) terms, and the solvation contribution of complex ($\Delta G_{sol}^{R'-L'}$), receptor (ΔG_{sol}^{R}; $\Delta G_{sol}^{R'}$), and ligand (ΔG_{sol}^{L}; $\Delta G_{sol}^{L'}$).

is assumed to be representative of the configurations sampled by separate ligand and receptor, then the ΔG_{MM} term is merely given by the addition of Coulomb (ΔG_{elec}) and van der Waals (ΔG_{vdW}) contributions (Equation 9.3).

$$\Delta G_{MM} = \Delta G_{elec} + \Delta G_{vdW} \tag{9.3}$$

The solvation-free energy is divided into polar (ΔG_{sol-p}) and nonpolar (ΔG_{sol-np}) components (Equation 9.4). The polar term reflects the change in free energy for the transfer from the gas phase to the aqueous solvent, typically modeled as homogeneous medium characterized with dielectric constant of 1 and 78.4, respectively. This term is calculated by resorting to numerical methods for solving the Poisson–Boltzmann equation through a finite-difference approach, or alternatively by means of the GB theory (for a review, see Orozco and Luque 2000).

$$\Delta G_{sol} = \Delta G_{sol-p} + \Delta G_{sol-np} \tag{9.4}$$

In a continuum electrostatics model, a hydrated solute molecule is treated as a charge distribution in a low-dielectric cavity, which is embedded in a

high-dielectric medium representing water. The dependence between the charge distribution and the electric potential is then given by the Poisson equation (Equation 9.5).

$$\nabla \varepsilon(r) \nabla \phi(r) = -\rho(r) \tag{9.5}$$

where $\varepsilon(r)$ denotes the dielectric constant, $\phi(r)$ is the electric potential, and $\rho(r)$ is the charge distribution.

In the presence of an ionic atmosphere, Equation 9.5 adopts the form given by the nonlinear Poisson–Boltzmann equation, which under the assumption that $\varphi(r)$ is small can be linearized (using the approximation that sinh $\varphi(r) \approx \varphi(r)$; Equation 9.6).

$$\nabla \varepsilon(r) \nabla \phi(r) - \kappa^2 \phi(r) = -\rho(r) \tag{9.6}$$

where κ is the Debye–Hückel inverse screening length.

Equations 9.5 and 9.6 must be solved numerically. The finite-difference method solves the differential equations by discretizing the region of interest into grid points (typically a cubic grid). Accordingly, the solute partial charges are fractionally distributed among the nearby grid points, the dielectric constants are assigned to each grid point according to the geometry of the dielectric boundary, and the second derivatives of the potential at each grid point can be expressed in terms of the potentials at neighboring points. The coupled expressions for the potentials on the grid produce a linear system of equations that can be solved to yield the potential at each grid point. It is worth noting, however, that estimates of the electrostatic component from grid-based solvers of the Poisson equation inevitably contain numerical grid-discretization errors, and that a careful assessment of these errors must be performed (Harris et al. 2013). Other approaches, such as the finite element method or the boundary element method, are also available (for details, see Tomasi and Persico 1994).

The GB model offers a simpler, computationally less-expensive approach to the electrostatic component of the solvation-free energy (Equation 9.7; Still et al. 1990).

$$\Delta G_{sol-p} = \frac{1}{2}\left(1 - \frac{1}{\varepsilon_{out}}\right)\sum_{i,j} \frac{q_i q_j}{f_{GB}} \tag{9.7}$$

where q_i denotes the partial atomic charges of the solute, ε_{out} is the dielectric constant of the solvent environment, and f_{GB} stands for the screening function, which is generally expressed as noted in Equation 9.8 (for a review, see Bashford and Case 2000).

$$f_{GB}(r_{ij}) = \left[r_{ij}^2 + \alpha_i \alpha_j \exp\left(\frac{-r_{ij}^2}{(4\alpha_i \alpha_j)}\right)\right]^{1/2} \tag{9.8}$$

where r_{ij} is the interatomic distance between particles i and j, α_i stands for the effective Born radius of particle i.

The use of Equation 9.7 makes the calculation of the electrostatic solvation term to be the sum of pairwise interactions, thus making it suitable for implementation in molecular dynamics (MD) programs. Furthermore, the pairwise nature of the method also facilitates decomposition of free energies into individual atomic contributions (see below).

The nonpolar contribution (ΔG_{sol-np}) is generally estimated by using a linear expression with the solvent-accessible surface (*SAS*; Equation 9.9), which is intended to account for the contributions due to the cavity formation within the solvent and the change in nonpolar interactions between solute and solvent (Sitkoff et al. 1994).

$$\Delta G_{sol-np} = \gamma SAS + \beta \qquad (9.9)$$

Finally, the change in configurational entropy of the solute is usually estimated by means of a normal mode analysis of harmonic frequencies calculated at the MM level. This analysis can be performed for simplified structures containing the residues within a given sphere centered at the ligand, and the energy-minimized structures are obtained by using a distance-dependent dielectric, which is introduced to mimic the solvent dielectric (Kongsted and Ryde 2009; Genheden and Ryde 2011; Hou et al. 2011). However, this contribution is often neglected when the primary interest is the prediction of relative binding affinities between structurally similar ligands.

9.2.2 Computational Aspects of MM-PB(GB)SA Calculations

Calculation of the binding affinity between a ligand and its receptor can be performed using two computational approaches, which involves a single trajectory of the ligand–receptor complex or separate trajectories of the ligand–receptor complex, the receptor and the ligand (Figure 9.1; Wang et al. 2006). Although this latter approach is formally more rigorous, because it takes into account the differences in conformational flexibility of the bound and unbound states, the single trajectory strategy is usually adopted because it benefits from the cancellation of intramolecular contributions in the prediction of the binding affinity, especially in cases where no large structural differences are expected to occur upon binding.

MM-PB(GB)SA calculations are generally performed for ensembles of structures sampled along the trajectories obtained from MD simulations. Then, a set of representative structures is extracted from the trajectory, water molecules and counterions are subsequently removed, and the free energy is calculated as noted in Equation 9.2. At this point, it has been pointed out that selecting a relatively small number of representative snapshots may suffice to obtain an accurate prediction comparable to using the full MD trajectory (Lill and Thompson 2011).

Since a single MD simulation may often not provide a complete description of the conformational space available for the ligand–receptor complex (and even for the separate receptor), it is then unclear whether the binding affinity estimated from a single trajectory can be representative or not. Adler and Beroza (2013) have recently considered this issue. Thus, replicate MM-PBSA calculations were performed for four distinct ligand–receptor complexes. Separate trajectories were generated using nearly identical starting coordinates (1% randomly perturbed by 0.001 Å), and they were found to lead to significantly different calculated binding free energies. Thus, even though the binding affinity did converge in each separate run, the variation across separate runs implies that a single trajectory may inadequately sample the system. Hence, the authors recommend that combining MM-PB(GB)SA with multiple samples of the initial starting coordinates will lead to more accurate estimates of the binding affinity.

However, it is worth noting that the inclusion of specific structural water molecules has been found to be important for the accurate description of MM-PB(GB)SA energetics. For instance, it has been reported that the difference in binding affinity of nevirapine to the wild-type HIV-1 reverse transcriptase and the Y181C mutant was better discriminated upon inclusion of key water molecules as part of the protein (Treesuwan and Hannongbua 2009). Similarly, the protein–protein interaction between the T-cell receptor and its staphylococcal enterotoxin 3 (SEC3) binding partner was only effectively discriminated against two mutated SEC3 variants only when key explicit water molecules were included in the calculations (Wong et al. 2009). On the contrary, a protocol for the inclusion of water molecules that mediate ligand–protein interactions, denoted water-MM-PBSA, has been reported (Zhu et al. 2014), leading to improved correlation between the binding affinities estimated for a series of JNK3 kinase inhibitors and the experimental IC_{50} values compared to that obtained from classical MM-PBSA calculations.

The averaged contributions obtained from the whole set of snapshots enable to check the time convergence and internal consistency of the binding affinity and its free energy components (Stoica et al. 2008), while they take into account the effect due to conformational fluctuations of the molecular system. However, it has been advocated that the conformational sampling of the simulated system should be performed using simulations with explicit treatment of the solvent molecules, avoiding the use of continuum solvent simulations (Weis et al. 2006). Furthermore, the mixing of force fields for collecting the snapshots along the discrete MD simulation and for the MM-PB(GB)SA calculation is not recommended, as it may give inaccuracies (Weis et al. 2006).

Even though MM-PB(GB)SA has proven to be successful in various ligand–protein complexes, the results also demonstrate that the overall performance is highly system-dependent. For instance, a systematic analysis of 59 ligands interacting with six distinct receptors showed that MM-PBSA gives good predictions for homologous ligands and has a variable performance for ligands with diverse structures (Figure 9.2; Hou et al. 2011). Furthermore, MM-PBSA
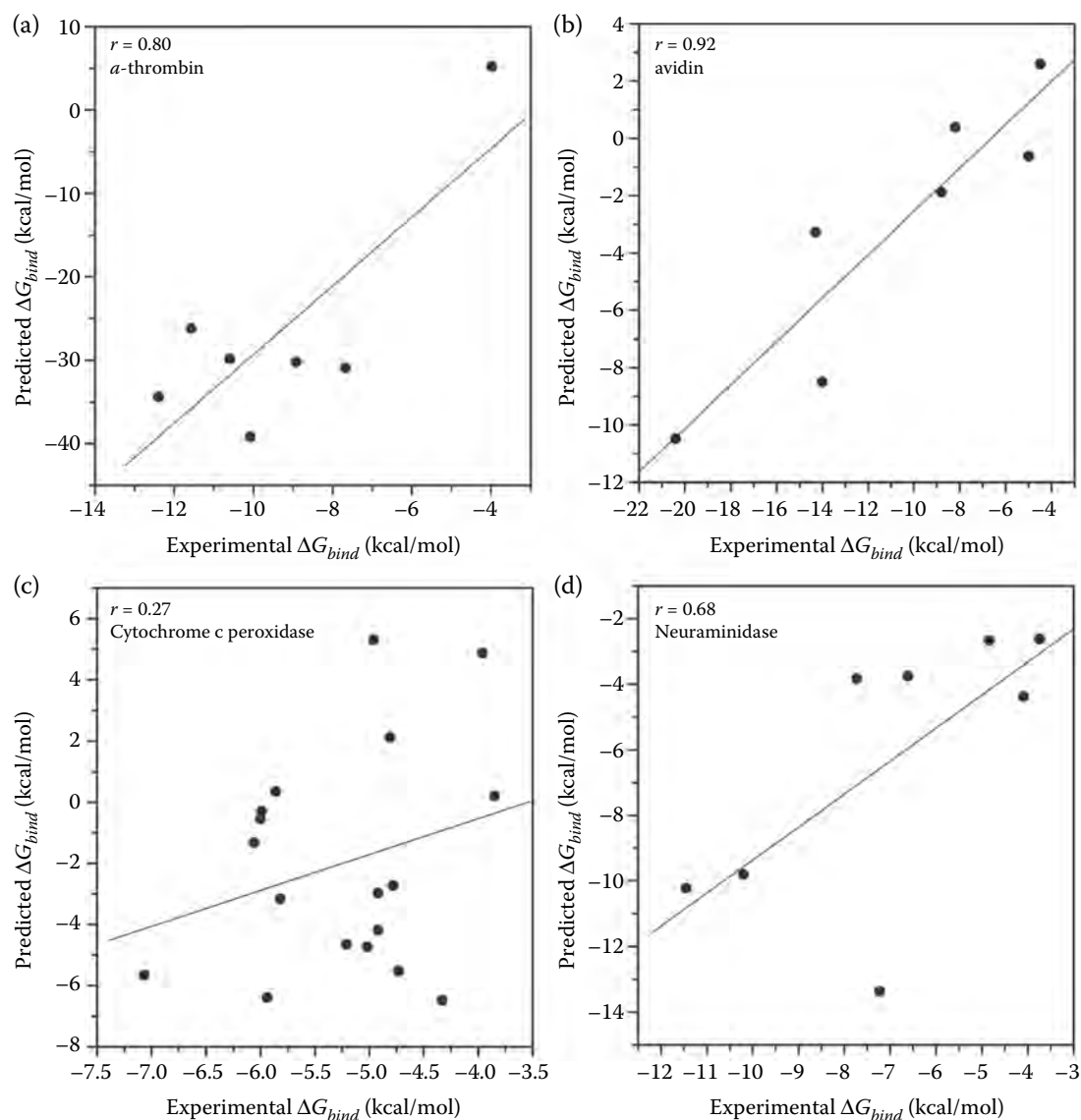
FIGURE 9.2

Correlations between the MM/PBSA binding affinities and the experimental values for (a) a-thrombin, (b) avidin, (c) cytochrome c peroxidase, (d) neuraminidase, (e) P450cam, and (f) penicillopepsin. (Reprinted with permission from Hou, T. et al. 2011. Assessing the performance of the MM/PBSA and MM/GBSA Methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* 51(1): 69–82. Copyright 2011, American Chemical Society.) (*Continued*)

predictions were found to be very sensitive to the solute dielectric constant, which is related to the physicochemical features of the binding interface. In fact, Hou et al. (2011) reported that for highly charged binding interfaces, a higher solute dielectric constant (ε_{in} ~ 4) is preferred, whereas for moderately charged or hydrophobic binding interfaces values of ε_{in} equal to 2 or 1, respectively, are more adequate. At this point, the authors suggested the change in the solvent-accessible surface area (SASA) of the groups involved in strong
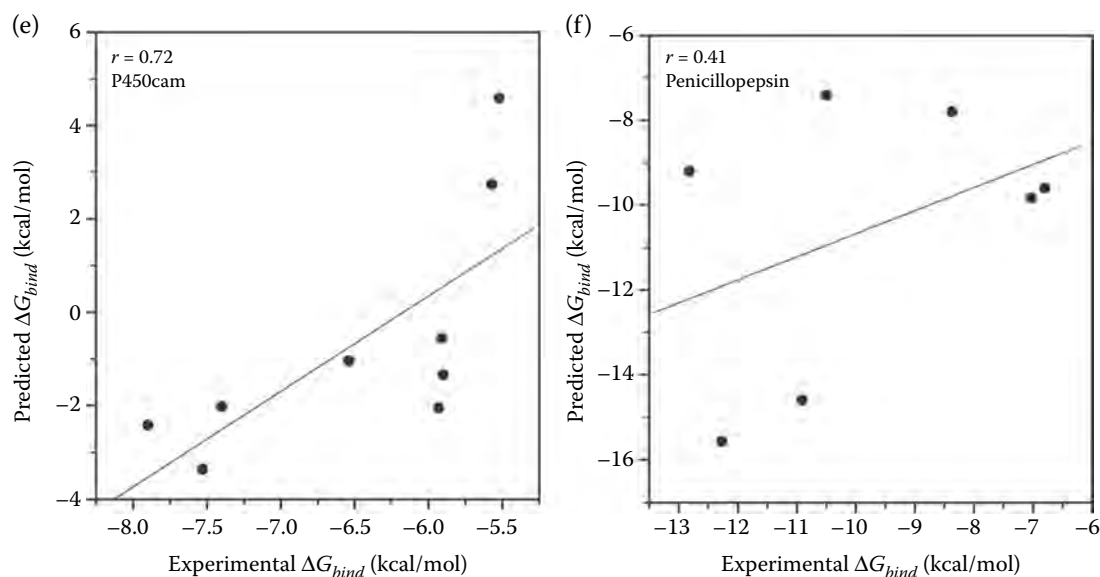
FIGURE 9.2 (Continued)
Correlations between the MM/PBSA binding affinities and the experimental values for (a) a-thrombin, (b) avidin, (c) cytochrome c peroxidase, (d) neuraminidase, (e) P450cam, and (f) penicillopepsin. (Reprinted with permission from Hou, T. et al. 2011. Assessing the performance of the MM/PBSA and MM/GBSA Methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* 51(1): 69–82. Copyright 2011, American Chemical Society.)

polar–polar interactions between ligand and receptor as a valuable guide to select the dielectric constant of the solute. Moreover, this study also concluded that inclusion of conformational entropy is crucial for predicting absolute binding free energies, but not for ranking the binding affinities of similar ligands.

Similar studies have been performed for MM-PB(GB)SA calculations for a total of 46 small molecules targeted to five different protein receptors (Xu et al. 2013). Attention was paid to the effect of (i) AMBER force fields (ff99, ff99SB, ff99SB-ILDN, ff03, and ff12SB), (ii) the timescale of MD simulations, and (iii) the impact of four different charge models (RESP, ESP, AM1-BCC, and Gasteiger) for small molecules.

In a separate work, Swanson et al. (2005) also examined the impact of solute charge, dielectric coefficient, and atomic radii on the accuracy in predicting the solvation-free energies. To this end, a set of 14 polyalanine peptides and a series of 20 nonzwitterionic *N*-acetyl-X-*N'*-methylamide dipeptides, with X representing one of the 20 standard amino acids, were subject to explicit solvent simulations, and the charging free energies were determined by means of FEP calculations. These data were then utilized for deriving two optimized sets of atomic radii, which were chosen to define either abrupt or cubic-spline smoother dielectric boundaries, to be used in conjunction with AMBER (parm99) charges. The optimized radii were found to offer increased accuracy of solvation energies and atomic forces in a test set of four protein-like polypeptides. The application of these optimized radii to the binding of

peptides to human class II MHC molecules was shown to reflect adequately the distinction between strong and for binding peptides (Cárdenas et al. 2010).

The aim of deriving parameters for implicit solvent models optimized in a system- or atom-specific manner on the basis of experimental data or more rigorous explicit solvent simulations has been adopted in other studies. For instance, the performance of PB calculations with regard to the TIP3P explicit solvent has been examined for a variety of systems of biochemical interest (Tan et al. 2006). The results support the transferability of empirically optimized parameters for the implicit solvent from small training molecules to large testing peptides. However, a computational strategy for optimizing the solute radii on the basis of forces and energies from explicit solvent simulations has been reported in the context of the AMBER partial charges and a spline-smoothed solute surface (Swanson et al. 2007). An alternative approach for deriving optimized radii for PB calculations has been undertaken by Yamagishi et al. (2014). The radii were optimized using results from explicit solvent simulations of amino acid templates and large peptides in the framework of the AMBER protein force field and using a smoothing dielectric function. Moreover, discrimination between radii assigned to N- and C-terminal residues from nonterminal ones was also considered.

In a different approach, Purisima and coworkers have developed the solvated interaction energy (SIE) method, which is an endpoint MM-PBSA-based scoring function that approximates the protein–ligand binding affinity by an interaction energy contribution and a desolvation free energy contribution (Naïm et al. 2007; Cui et al. 2008). Electrostatic solvation effects are calculated with the boundary element solution to the Poisson equation, while nonpolar solvation is based on change in the SAS. As in the single-trajectory approach, the free state is generally obtained by separation of both ligand and receptor from the ligand–receptor complex sampled along the MD trajectory. The SIE method has been carefully calibrated using a diverse set of ligand–protein complexes, including the calibration of parameters such as the dielectric constant, the surface tension coefficient, and the inclusion of an enthalpy–entropy compensating scaling factor. The SIE scoring function leads to a reasonable agreement between predicted and experimental binding affinities, as noted in the external testing against a curated dataset of 343 ligand–protein complexes, leading to a root-mean square error in the predicted binding affinities of 2.5 kcal mol^{-1} (Sulea et al. 2011).

9.2.3 Large-Scale Application of MM-PB(GB)SA Models

The advent of faster computers and automated procedures for preparation of ligands and receptors has promoted the use of MM-PB(GB)SA models in medium- and high-throughput screenings, making them valuable for reranking of docked poses. As an example, Brown and Muchmore (2009) reported a large-scale application to a set of 308 small-molecule ligands in complex with urokinase, PTP-1B, and Chk-1. Briefly, they use a GB implicit solvation model

during the computer-intensive ensemble-generating MD runs, whereas in the postproduction process a PB solver that employs a diffuse representation of the dielectric boundary (instead of the more common discrete transition between solute and solvent). Statistically significant correlations to experimentally measured potencies were found, leading to correlation coefficients for the three proteins in the range 0.72–0.83.

Greenidge et al. (2013) have validated an automated implementation of MM-GBSA using a large and diverse selection of 855 protein–ligand complexes. In particular, calculations were performed using the VSGB 2.0 energy model, which features an optimized implicit solvent model that includes physics-based corrections for hydrogen bonding, pi–pi interactions, self-contact interactions and hydrophobic contacts, and parameters were fit to a crystallographic database of 2239 single side chain and 100 11–13 residue loop predictions (Li et al. 2012). Calculations were performed using the KNIME-automated workflow. After carefully removing flawed structures, comparison of calculated and experimental binding affinities showed a significant correlation ($R^2 = 0.63$; Figure 9.3). The study also discussed the impact of ligand strain and water molecules, revealing that while inclusion of water molecules deteriorates the predictive quality, inclusion of ligand strain slightly improves the overall accuracy. In an independent study, the accuracy of the VSGB 2.0 energy model in predicting binding free energies was
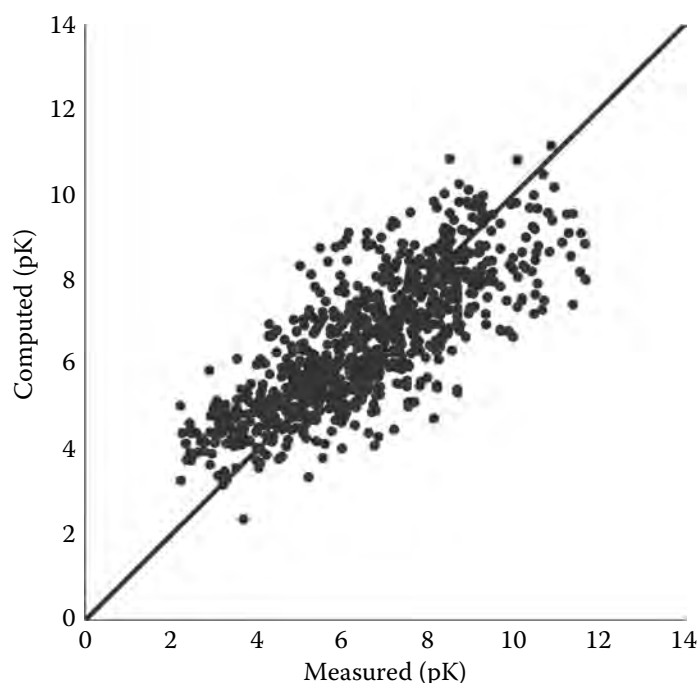

FIGURE 9.3

Comparison of computed and experimental binding affinities. (Reprinted with permission from Greenidge, P. A. et al. 2013. MM/GBSA binding energy prediction on the PDBbind data set: Successes, failures, and directions for further improvement. *J. Chem. Inf. Model.* 53(1): 201–209. Copyright 2013, American Chemical Society.)

also tested for 106 protein–ligand complexes (Mulakala and Viswanadhan 2013). The results indicate that this method may be approaching the accuracy required for absolute binding free energy determination, although through linear regression and without any conformational sampling. Furthermore, given the modest computational cost of these calculations, the MM-GBSA formalism may be poised toward generating physics-based scoring functions for docking.

Very recently, Greenidge et al. (2014) have shown that MM-GBSA can be used as an independent scoring function to assess the energetically preferred pose as generated with multiple scoring functions, and in multiple protein conformations. The results supported the role of MM-GBSA to distinguish between true and decoy poses of a ligand in addition to the rescoring of data sets.

A last example of the progressive large-scale application of MM-PB(GB)SA methods is the high-throughput virtual screening pipeline for *in silico* screening of virtual compound databases using high-performance computing (Zhang et al. 2014). This pipeline involves an automated receptor preparation scheme with unsupervised binding site identification, including receptor/target preparation, ligand preparation, VinaLC docking calculation, and MM-GBSA rescoring. The results demonstrate that MM-GBSA rescoring has higher average receiver operating characteristic (ROC) area under curve (AUC) values and consistently better early recovery of actives than Vina docking alone, though the enrichment performance is target-dependent.

9.3 Per-Residue Decomposition of the MM-PB(GB)SA Free Energy

The decomposition into per-residue and residue-pairwise contributions of the MM-PB(GB)SA binding free energy allows to unravel the network of energetic interactions that stabilize ligand–protein binding, thus providing insight into key features of binding (Gohlke et al. 2003). All the components of the binding affinity (Equation 9.2) can be decomposed with certain degree of approximation into per-residue and also residue-pairwise contributions according to the standard scheme given by Equation 9.10.

$$\Delta G_{bin} = \sum_{i=1}^{n} \Delta G^i = \sum_{i=1}^{n} \sum_{j \neq i}^{n} \Delta G^{i,j} \tag{9.10}$$

where n is the total number of residues, ΔG^i are the per-residue contributions, and $\Delta G^{i,j}$ are the residue-pairwise interaction contributions.

Under this scheme ΔG_{bin} can also be partitioned into the receptor and ligand components by summing the corresponding per-residue contributions of each fragment (Equation 9.11).

$$\Delta G_{bin} = \Delta G^{receptor} + \Delta G^{ligand} \tag{9.11}$$

It is worth noting that only the electrostatic (ΔG_{elec}) and van der Waals (ΔG_{vdW}) terms are strictly residue-pairwise decomposable, so that one-half of the pairwise interaction energy between two residues i and j is attributed to both of them. However, the solvation terms are not inherently decomposable, since the effective Born radii for GB and dielectric boundaries for PB are dependent on the surroundings (Miller et al. 2012).

Regarding the GB polar solvation term, $\Delta G_{sol-p}^{i,j}$, a pairwise descreening approximation was implemented by Onufriev et al. (Onufriev et al. 2000; Tsui and Case 2001) based on the improvement of the standard GB model (Hawkins et al. 1995) as noted in Equation 9.12.

$$\Delta G_{sol-p}^{i,j} = \sum_{l \in i}^{n_i} \sum_{k \in j}^{n_j} -\frac{1}{2}\left(\frac{1}{\varepsilon_{in}} - \frac{e^{-\kappa f_{GB}}}{\varepsilon_{out}} \right) \frac{q_l q_k}{f_{GB}} \tag{9.12}$$

where ε_{in} and ε_{out} are the solute and solvent dielectric constants, κ is the Debye–Hückel screening parameter to account for salt effects at low salt concentrations (Srinivasan et al. 1999).

Since f_{GB} depends on the effective Born radius (Equation 9.8), $\Delta G_{sol-p}^{i,j}$ is inherently nondecomposable, that is, the polar solvation interaction between residues i and j is affected by all other atoms in the system. Therefore, the binding free energies of receptor and ligand (Equation 9.11) become asymmetric, since the effective Born radius yields different values depending on the overall structure of either complex or receptor/ligand. A similar reasoning can be used in relation to the PB dielectric boundary to reach the conclusion that the PB polar solvation energy is neither inherently decomposable nor produces symmetric binding free energies. However, the nonpolar solvation term, ΔG_{sol-np}, also contains intrinsic difficulties in its geometry decomposition due to the nonlocal character of the SASA-dependent term used for its calculation (Gohlke et al. 2003), introducing asymmetry in the binding free energy between the protein and the ligand.

Regarding the configurational entropy, the decomposition at residue or residue-pairwise level remains still to be solved, though attempts to decompose the normal modes that contribute to the vibrational entropy into atomic contributions have been reported (Zoete and Michielin 2007). Generally, the configurational entropy decomposed at the residue level due to the loss of torsional freedom can be computed using the computational scheme adopted by Honig and coworkers (Froloff et al. 1997), which is based on the empirical scale of Pickett and Stemberg (1993). This procedure separates backbone and

side-chain components. For the backbone, an entropic penalty of 2 kcal mol^{-1} per residue is considered, whereas a variable value is computed for side-chain component depending on the solvent-exposed surface area (Doig and Sternberg 1995).

The MM-PB(GB)SA fragmental decomposition yields a high number of components that, combined with the systematic application to a set of protein–ligand complexes either from MD simulations or other sampling methods, can form voluminous energy matrices. The amount of data generated for this decomposition is vast and thus impedes univariate exploration. Alternatively, multivariate data analysis techniques such as partial least squares (PLS) or principal component analysis (PCA) have found their applicability to the in-depth exploration of the computed energy matrices in order to find significant residues or residue-pairwise contributions that govern the binding free energy.

The per-residue decomposition methodology has been widely applied to the study of protein–ligand binding free energy (Zoete and Michielin 2007; Berhanu and Masunov 2012; Laurini et al. 2013). However, the residue-pairwise decomposition has been less used, though recently several works have explored its potentiality. For instance, it has been used to elucidate the signal transmission mechanism in the allosteric regulation of protein kinases C by determining the differences in the residue-pairwise interaction profiles among six protein states of the mentioned protein (Seco et al. 2012). Furthermore, Pouplana and Campanera (2015) have used it to determine the relative importance of the hydrophobic fragments of Aβ oligomers in the oligomerization process of such peptides. As shown in the decompostion of the intermonomeric van der Waals free energy in Figure 9.4, the hydrophobic collapse in the formation of these oligomers is caused by hydrophobic interactions between three well-defined hydrophobic fragments: 31–35 (C-terminal hydrophobic region [CTHR]), 17–20 (central hydrophobic region [CHC]), and 12–14 (N-terminal hydrophobic region [NTHR]), ordered according to their importance.

9.4 Quantum Mechanics and Implicit Solvation Models

The use of simplified expressions in classical force fields is understandable in terms of providing an efficient sampling, as well as in facilitating the parametrization of the large number of functional groups that can be incorporated into drug-like molecules. However, these approximations also limit the accuracy of classical force fields in describing the intermolecular interactions that mediate the recognition between ligands and proteins. Thus, besides typical interactions such as salt bridges, standard hydrogen bonds, and van der Waals forces, a wider number of stabilizing interactions
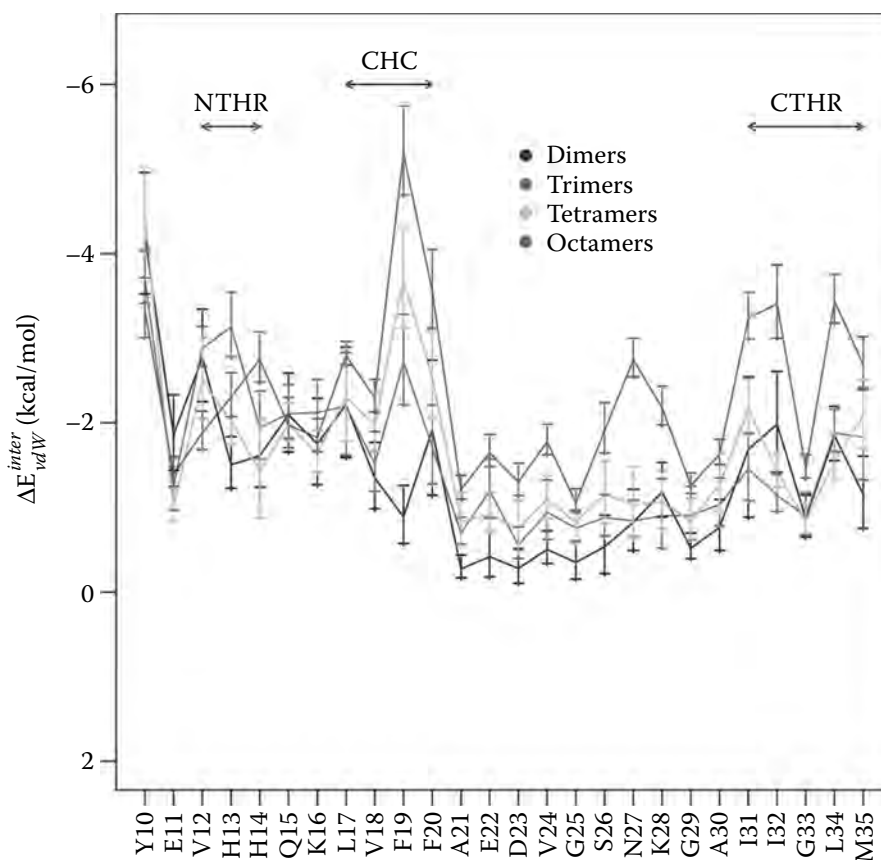
FIGURE 9.4
(**See color insert.**) Residue decomposition of the intermonomeric total stability free energy (kcal mol^{-1}) of different oligomers of β-amyloid peptide. (Reproduced from Pouplana, R. and J. M. Campanera. 2015. *Phys. Chem. Chem. Phys.* 17(4): 2823–2837. With permission from the PCCP Owner Societies.)

have been characterized in the last decades, including cation-π or anion-π complexes (Frontera et al. 2011), nonstandard hydrogen bonds (Hobza and Havlas 2000), and halogen bonding (Nguyen et al. 2004; Sarwar et al. 2010).

Quantum mechanical (QM) methods are the most accurate approach to the calculations of intermolecular interactions, and they form the basis for the parametrization of force fields. The continued increase in accuracy achieved by QM methods has also stimulated the implementation and usage of QM-based techniques for different applications in the study of ligand–protein complexes. Most of these applications follow the hybrid QM/MM computational scheme (Warshel 2003; Friesner and Guallar 2005), where the Hamiltonian of the whole system can be defined as the sum of three terms (Equation 9.13) corresponding to the QM subsystem ($\hat{H}_{QM}$), the MM subsystem ($\hat{H}_{MM}$), and the coupling between the QM and MM regions ($\hat{H}_{QM/MM}$).

$$\hat{H} = \hat{H}_{QM} + \hat{H}_{MM} + \hat{H}_{QM/MM} \qquad (9.13)$$

Although the major goal of QM/MM methods has been the study of reactive processes in condensed media or in enzymes, a wider range of applications is being explored in drug discovery, including the calculation of the ligand–protein interaction energy and the analysis of the energy components, and the rescoring of docking calculations (Hensen et al. 2004; Cho et al. 2005; Illingworth et al. 2008; Cho and Rinaldo 2009; Chaskar et al. 2014).

QM-based strategies have also been developed for the prediction of binding affinities of ligand–protein complexes. To this end, a variety of methodological strategies have been adopted, as will be illustrated by the representative cases presented below (Zhou et al. 2010; Barril and Luque 2012; Ilatovskiy et al. 2013; Mucs and Bryce 2013).

Balaz and coworkers have proposed a four-step strategy for the study of ligand-metalloprotein complexes (Khandelwal et al. 2005). The procedure involves docking of ligands, optimization of the complex, conformational sampling with constrained metal bonds, and a single point QM/MM calculation for the time-averaged structure. Finally, the QM/MM interaction energy, $\Delta\langle E_{QM/MM}\rangle$, is combined with a desolvation term in order to determine the binding free energy (Equation 9.14). After suitable parametrization against experimental data for a set of 28 hydroxamate inhibitors binding to zinc-dependent matrix metalloproteinase 9, Equation 9.14 was able to account for 90% of variance in the inhibition constants.

$$\Delta G_{bin} = \alpha\Delta\langle E_{QM/MM}\rangle + \gamma\Delta\langle SASA\rangle + \kappa \tag{9.14}$$

where $\Delta\langle SASA\rangle$ denotes the change in SAS upon complexation.

In a distinct study, the ability of QM/MM combined with the PBSA model has been utilized for the calculation of binding affinities for flexible ligands (Gräter et al. 2005). The method was tested for a set of 47 benzamidine derivatives binding to trypsin. The suitability of the computational strategy for automated ligand docking and scoring is supported by the accuracy in predicting the experimental range of binding energies, with a root-mean square error of 1.2 kcal mol^{-1}.

Das et al. (2009) followed a strategy based on the use of protein-polarized QM charges in GBSA calculations for nine protease inhibitors. In this work, the general expression of a GBSA model was adopted, but the ligand was described by assigning either MM charges or the protein-polarized ones as derived from QM/MM calculations. Moreover, attention was paid to the effect of including bridging water molecules that mediate hydrogen bonding with the ligand. The results showed that the binding free energies determined by using those polarized charges (and specific water molecules) showed higher correlation with antiviral IC$_{50}$ data. The importance of including polarization effects through QM/MM methods, combined with a van der Waals correction and a term accounting for desolvation, has also been highlighted for ligands binding to trypsin and cytochrome c peroxidase (Burger et al. 2011).

An elaborate scheme was reported by Raha and Merz (2004, 2005) with the aim to perform a large-scale validation of a QM-based scoring function for predicting the binding affinity of a diverse set of ligands. In this study, the binding affinity was determined as noted in Equation 9.15, where it is decomposed into the gas-phase interaction energy (ΔG_b^{gas}), and the change in solvation-free energy ($\Delta\Delta G_{solv}$) of the complex (ΔG_{solv}^{PL}) relative to protein (ΔG_{solv}^{P}) and ligand (ΔG_{solv}^{L}).

$$\Delta G_{bin} = \Delta G_b^{gas} + \Delta\Delta G_{solv} = \Delta G_b^{gas} + \Delta G_{solv}^{PL} - \Delta G_{solv}^{P} - \Delta G_{solv}^{L} \qquad (9.15)$$

The gas-phase interaction energy was determined as a sum of electrostatic and nonpolar interaction energies. The former was calculated using the divide-and-conquer method and the semiempirical AM1 or PM3 Hamiltonians, and the latter with the classical attractive component of the Lennard–Jones interaction potential. Furthermore, the entropic term was expressed as the addition of conformational and solvent entropy components. The former was estimated by considering a conformational penalty of 1 kcal mol^{-1} for each rotatable bond of the ligand and in the protein side chains frozen upon formation of the complex. The solvent entropy term accounts for the entropy gained by release of water molecules upon binding, and it was calculated from the buried surface area resulting upon complexation. Finally, the solvation-free energy term was determined using a QM self-consistent reaction field calculation for the complex, ligand, and protein. Finally, the weights of the different components were adjusted by fitting to experimental binding free energies. The method was shown to be effective as scoring function for predicting ligand poses docked to a protein target and for discriminating between native and decoy poses.

A related QM-based scheme based on the semiempirical QM PM6-DH2 method, which includes corrections dispersion energy and hydrogen bonds, has been proposed for the computation of binding affinities (Fanfrlik et al. 2010; Dobes et al. 2011a). Here, the binding affinity is determined by adding the PM6-DH2 interaction enthalpy evaluated in a continuum water environment using the COSMO model. The desolvation of the ligand was further refined by means of solvation model based on density (SMD) continuum calculations. Furthermore, the deformation contribution due to changes in protein and ligand upon binding was also considered. The method was successful in ranking 22 ligands binding to HIV-1 protease, and for the binding of 15 structurally diverse inhibitors to CDK2. Recently, the method has been extended to treat halogen bonding (Dobes et al. 2011b) as well as to treat noncovalent binding in protein–ligand complexes (Fanfrlik et al. 2013).

The MM/QM-COSMO strategy has been adopted to evaluate the binding affinity of phosphopeptide inhibitors of the Lck SH2 domain (Anisimov and Cavasotto 2011). Starting from MD trajectories of the complex, a QM postprocessing is made for a selection of representative snapshots, which

were first refined using the PM3 Hamiltonian and the COSMO continuum solvent model. The binding free energy was then determined as noted in Equation 9.16, where the first term in the right-hand side was determined using Equation 9.17, and the entropic term included changes in translational and rotational rigid body component and the change in vibrational entropy.

$$\Delta G_{binding} = \Delta \left\langle H^{COSMO} \right\rangle - T \Delta S^{RB} - T \Delta S^{int} \tag{9.16}$$

$$H^{COSMO} = E^{COSMO} + G_{np}^{solv} \tag{9.17}$$

where E^{COSMO} represents the PM3 QM energy (including vacuum and solvation energy components), and the nonpolar contribution (G_{np}^{solv}) is determined using a linear relationship with the change in SAS.

The binding affinities derived from MM/QM-COSMO calculations were compared with the results determined using MM-PBSA and MM-GBSA, as well as the SIE method. The MM/QM-COSMO method showed the best agreement both for absolute (average unsigned error of 0.7 kcal mol^{-1}) and relative binding free energies.

9.5 Conclusion

Despite substantial progresses made in the last years, predicting the binding free energy of ligand to their targets still remains a major challenge for computational chemistry. This conforms to the involvement of different enthalpic and entropic components, each playing a significant contribution, and to the important compensation between these thermodynamic quantities. Furthermore, the need to develop fast, yet accurate estimates of binding affinities, which may discriminate between strong and weak binders and between distinct poses of a given compound, is required for large-scale application in drug discovery. In this context, the use of implicit solvation methods represents a fundamental tool in the path toward novel computational strategies for the high-throughput analysis of ligand–receptor complexes.

In the classical framework, MM-PB(GB)SA methods are *a priori* well suited to attain the preceding goal due to the continuous development of more accurate force fields, and specially to the refinement of the crude approximations inherent in the description of solvent effects through implicit continuum models. Thus, among the wide range of applications achieved by MM-PB(GB)SA methods in the study of biomolecular systems, the large-scale application to virtual screening appears to be especially promising, as these methods are reaching the predictive accuracy that would be required to discriminate among large sets of compounds covering a wide range of binding affinities.

However, the availability of decomposition schemes permits to disclose the contribution of specific molecular determinants (i.e., chemical groups in the ligand or residues in the binding pocket) that play a distinctive role in the binding affinity, thus providing valuable guidelines to assist the structure-based drug design.

The investigation of compounds with small differences in the binding affinity seems still out of the realm of MM-PB(GB)SA methods, partly due to the limitations of the classical force field to account for the interactions formed between a ligand and its target, taking a proper accounting of electrostatic, induction, charge transfer, and dispersion effects, as well as from the simplified description of environmental effects. At this point, QM-based methods used directly for the modeled structures of ligand–protein complexes or in the framework of endpoint sampling techniques represent a promising alternative as a tool to develop and calibrate novel computational strategies designed to provide accurate estimates of binding affinities (Yilmazer and Korth 2013). Furthermore, the development of QM-based strategies can give rise to accurate tools for lead optimization, even though this option is seriously limited by the huge computational cost of high-level QM computations. This explains why most of the QM-based strategies devised for the study of ligand–protein complexes rely on semiempirical methods, often supplemented by suitable correction terms to assure the description of certain types of interactions. On the contrary, current efforts for making quantum chemistry codes more efficient and implementing them in powerful computational resources can be relevant to alleviate the computational requirements of QM-based strategies. Overall, it can be envisaged that QM-based approaches will be an increasingly used and valued tool in computational medicinal chemistry and structure-based drug discovery.

Acknowledgments

References

Adler, M. and P. Beroza. 2013. Improved ligand binding energies derived from molecular dynamics: Replicate sampling enhances the search of conformational space. *J. Chem. Inf. Model.* 53(8): 2065–2072.

Anisimov, V. M. and C. N. Cavasotto. 2011. Quantum mechanical binding free-energy calculation for phosphopeptide inhibitors of the Lck SH2 domain. *J. Comput. Chem*. 32(10): 2254–2263.

Barril, X. and F. J. Luque. 2012. Molecular simulation methods in drug discovery: A prospective outlook. *J. Comput.-Aided Mol. Des*. 26(1): 81–86.

Bashford, D. and D. A. Case. 2000. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem*. 51: 129–152.

Berhanu, W. M. and A. E. Masunov. 2012. Unique example of amyloid aggregates stabilized by main chain H-bond instead of the steric zipper: Molecular dynamics study of the amyloidogenic segment of amylin wild-type and mutants. *J. Mol. Model*. 18(3): 891–903.

Bissantz, C., B. Kuhn, and M. Stahl. 2010. A medicinal chemist's guide to molecular interactions. *J. Med. Chem*. 53(14): 5061–5084.

Boukharta, L., H. Gutiérrez-de-Terán, and J. Aqvist. 2014. Computational prediction of alanine scanning and ligand binding in G-protein coupled receptors. *PLoS Comput. Biol.* 10(4): e1003585.

Brandsdal, B. O., F. Österberg, M. Almlöf, I. Feierberg, V. Luzhkov, and J. Aqvist. 2003. Free energy calculations and ligand binding. *Adv. Protein Chem*. 66: 123–158.

Brown, S. P. and S. W. Muchmore. 2009. Large-scale application of high-throughput molecular mechanics with Poisson-Boltzmann surface area for routine physics-based scoring of protein-ligand complexes. *J. Med. Chem.* 52(10): 3159–3165.

Burger, S. K., D. C. Thompson, and P. W. Ayers. 2011. Quantum mechanics/molecular mechanics strategies for docking pose refinement: Distinguishing between binders and decoys in cytochrome c peroxidase. *J. Chem. Inf. Model*. 51(1): 93–101.

Butler, K. T., F. J. Luque, and X. Barril. 2009. Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem*. 30(4): 601–610.

Cárdenas, C., A. Bidon-Chanal, P. Conejeros, G. Arenas, S. Marshall, and F. J. Luque. 2010. Molecular modelling of class I and II alleles of the major histocompatibility complex of Salmo salar. *J. Comput.-Aided Mol. Des*. 24(12): 1035–1051.

Chang, C. A., W. Chen, and M. K. Gilson. 2007. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. USA* 104(5): 1534–1539.

Chaskar, P., V. Zoete, and U. F. Röhring. 2014. Toward on-the-fly quantum mechanical/molecular mechanical (QM/MM) docking: Development and benchmark of a scoring function. *J. Chem. Inf. Model*. 54(11): 3137–3152.

Chipot, C. and A. Pohorille. (Eds.), 2007. Free energy calculations. *Theory and Applications in Chemistry and Biology*. Springer Series in Chemical Physics, vol. 86. Berlin: Springer.

Cho, A. E., V. Guallar, B. J. Berne, and R. Friesner. 2005. Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem*. 26(9): 915–931.

Cho, A. E. and D. Rinaldo. 2009. Extension of QM/MM docking and its applications to metalloproteins. *J. Comput. Chem*. 30(16): 2609–2616.

Csermely, P., R. Palotai, and R. Nussinov. 2010. Inducedfit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends Biochem. Sci*. 35(10): 539–546.

Cui, Q., T. Sulea, J. D. Schrag, C. Munger, M.-N. Hung, M. Naïm, M. Cugler, and E. O. Purisima. 2008. Molecular dynamics—Solvent interaction energy studies of protein-protein interactions: The MP1-p14 scaffolding complex. *J. Mol. Biol*. 379(4): 787–802.

Das, D., Y. Koh, Y. Tojo, A. K. Gosh, and H. Mitsuya. 2009. Prediction of potency of protease inhibitors using free energy simulations with polarizable quantum mechanics based ligand charges and a hybrid water model. *J. Chem. Inf. Model.* 49(12): 2851–2862.

Dobes, P., J. Fanfrlik, J. Rezac, M. Otypeka, and P. Hobza. 2011a. Transferable scoring function based on semiempirical quantum mechanical PM6-DH2 method: CDK2 with 15 structurally diverse inhibitors. *J. Comput.-Aided Mol. Des.* 25(3): 223–235.

Dobes, P., J. Rezac, J. Fanfrlik, M. Otypeka, and P. Hobza. 2011b. Semiempirical quantum mechanical method PM6-DH2X describes the geometry and energetics of CK2-inhibitor complexes involving halogen bonds well, while the empirical potential fails. *J. Phys. Chem. B* 115(26): 8581–8589.

Doig, A. J. and M. J. E. Sternberg. 1995. Side-chain conformational entropy in protein folding. *Prot. Sci.* 4(11): 2247–2251.

Fanfrlik, J., A. K. Bronowska, J. Rezac, O. Prenosil, J. Konvalinka, and P. Hobza. 2010. A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and H-bonding: HIV-1 protease with 22 ligands. *J. Phys. Chem. B* 114(39): 12666–12678.

Fanfrlik, J., P. S. Brahmkshatriya, J. Rezac, A. Jilkova, M. Horn, M. Mares, P. Hobza, and M. Lepsik. 2013. Quantum mechanics-based scoring rationalizes the irreversible inactivation of parasitic *Schistosoma mansoni* cysteine peptidase by vinyl sulfone inhibitors. *J. Phys. Chem. B* 117(48): 14973–14982.

Friesner, R. A. and V. Guallar. 2005. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* 56: 389–427.

Froloff, N., A. Windemuth, and B. Honig. 1997. On the calculation of binding free energies using continuum methods: Application to MHC class I protein-peptide interactions. *Prot. Sci.* 6(6): 1293–1301.

Frontera, A., D. Quiñonero, and P. M. Deyà. 2011. Cation–π and anion–π interactions. *WIRES Comput. Mol. Sci.* 1(3): 440–459.

Genheden, S. and U. Ryde. 2011. Comparison of the efficiency of the LIE and MM/GBSA methods to calculate ligand-binding affinities. *J. Chem. Theory Comput.* 7(11): 3768–3778.

Gohlke, H. and G. Klebe. 2002. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed.* 41(15): 2644–2676.

Gohlke, H., C. Kiel, and D. A. Case. 2003. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *J. Mol. Biol.* 330(4): 891–913.

Gräter, F., S. M. Schwarzl, A. Dejaegere, S. Fischer, and J. C. Smith. 2005. Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics. *J. Phys. Chem. B* 109(20): 10474–10483.

Greenidge, P. A., C. Kramer, J.-C. Mozziconacci, and R. M. Wolf. 2013. MM/GBSA binding energy prediction on the PDBbind data set: Successes, failures, and directions for further improvement. *J. Chem. Inf. Model.* 53(1): 201–209.

Greenidge, P. A., C. Kramer, J.-C. Mozziconacci, and W. Sherman. 2014. Improved docking results via reranking of ensembles of ligand poses in multiple x-ray protein conformations with MM-GBSA. *J. Chem. Inf. Model.* 54(10): 2697–2717.

Harris, R. C., A. H. Boschitsch, and M. O. Fenley. 2013. Influence of grid spacing in Poisson-Boltzmann equation binding energy estimation. *J. Chem. Theory Comput.* 9(8): 3677–3685.

Hawkins, G. D., C. J. Cramer, and D. G. Truhlar. 1995. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* 246(1–2): 122–129.

Hensen, C., J. C. Hermann, K. Nam, S. Ma, J. Gao, and H.-D. Höltje. 2004. A combined QM/MM approach to protein–ligand interactions: Polarization effects of the HIV-1 protease on selected high affinity inhibitors. *J. Med. Chem.* 47(27): 6673–6680.

Hobza, P. and Z. Havlas. 2000. Blue-shifting hydrogen bonds. *Chem. Rev.* 100(11): 4253–4264.

Hou, T., J. Wang, Y. Li, and W. Wang. 2011. Assessing the performance of the MM/PBSA and MM/GBSA Methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* 51(1): 69–82.

Ilatovskiy, A. V., R. Abagyan, and I. Kufareva. 2013. Quantum mechanics approaches to drug research in the era of structural chemogenomics. *Int. J. Quantum Chem.* 113(12): 1669–1675.

Illingworth, C. J. R., G. M. Morris, K. E. B. Parkes, C. R. Snell, and C. A. Reynolds. 2008. Assessing the role of polarization in docking. *J. Phys. Chem. A* 112(47): 12157–12163.

Jorgensen, W. L. 2009. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* 42(6): 724–733.

Khandelwal, A., V. Lukacova, D. Comez, D. M. Kroll, S. Raha, and S. Balaz. 2005. A combination of docking, QM/MM methods, and MD simulation for the binding affinity estimation of metalloprotein ligands. *J. Med. Chem.* 48(17): 5437–5447.

Kongsted, J. and U. Ryde. 2009. An improved method to predict the entropy term with the MM/PBSA approach. *J. Comput.-Aided Mol. Des.* 23(2): 63–71.

Kuntz, I. D., K. Chen, K. A. Sharp, and P. A. Kollman. 1999. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. USA* 96(18): 9997–10002.

Laurini, E., V. Da Col, B. Wünsch, and S. Prici. 2013. Analysis of the molecular interactions of the potent analgesic S1RA with the σ1 receptor. *Bioorg. Med. Chem. Lett.* 23(10): 2868–2871.

Lawrenz, M., R. Baron, Y. Wang, and J. A. McCammon. 2011. Effects of biomolecular flexibility on alchemical calculations of absolute binding free energies. *J. Chem. Theory Comput.* 7(7): 2224–2232.

Li, J., R. Abel, K. Zhu, Y. Cao, S. Zhao, and R. A. Friesner. 2012. The VSGB 2.0 model: A next generation energy model for high resolution protein structure modelling. *Proteins: Struct., Funct., Bioinf.* 79(10): 2794–2812.

Lill, M. A. and J. J. Thompson. 2011. Solvent interaction energy calculations on Molecular Dynamics trajectories: Increasing the efficiency using systematic frame selection. *J. Chem. Inf. Model.* 51(10): 2680–2689.

Merz, K. M. Jr. 2010. Limits of free energy computation for protein-ligand interactions. *J. Chem. Theory Comput.* 6(5): 1769–1776.

Michel, J. and J. W. Essex. 2010. Prediction of protein–ligand binding affinity by free energy simulations: Assumptions, pitfalls and expectations. *J. Comput.-Aided Mol. Des.* 24(8): 639–658.

Miller, B. R., T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg. 2012. MMPBSA.py: An efficient program for end-state free energy calculations. *J. Chem. Theory Comput*. 8(9): 3314–3321.

Mucs, D. and R. A. Bryce. 2013. The application of quantum mechanics in structure-based drug design. *Expert Op. Drug Discov*. 8(3): 263–276.

Mulakala, C. and V. N. Viswanadhan. 2013. Could MM-GBSA be accurate enough for calculation of absolute protein/ligand binding free energies? *J. Mol. Graphics Model*. 46: 41–51.

Naïm, M., S. Bhat, K. N. Rankin, S. Dennis, S. F. Chowdhury, I. Siddiqi, P. Drabik et al. 2007. Solvated Interaction Energy (SIE) for scoring protein–ligand binding affinities. 1. Exploring the parameter space. *J. Chem. Inf. Model*. 47(1): 122–133.

Nguyen, H. L., P. N. Horton, M. B. Hursthouse, A. C. Legon, and D. W. Bruce. 2004. Halogen bonding: A new interaction for liquid crystal formation. *J. Am. Chem. Soc*. 126(1): 16–17.

Onufriev, A., D. Basford, and D. A. Case. 2000. Modification of the Generalized Born model suitable for macromolecules. *J. Phys. Chem B*. 104(15): 3712–3720.

Orozco, M. and F. J. Luque. 2000. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev*. 100 (11): 4187–4225.

Pickett, S. D. and M. J. E. Stemberg. 1993. Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol*. 231(3): 825–839.

Pitera, J. W. and W. F. Van Gunsteren. 2002. A comparison of non-bonded scaling approaches for free energy calculations. *Mol. Simul*. 28(1–2): 45–65.

Pliego, J. R., Jr. and J. M. Riveros. 2002. Gibbs energy of solvation of organic ions in aqueous and dimethyl sulfoxide solutions. *Phys. Chem. Chem. Phys*. 4(9): 1622–1627.

Pouplana, R. and J. M. Campanera. 2015. Energetic contributions of residues to the formation of early amyloid-β oligomers. *Phys. Chem. Chem. Phys*. 17(4): 2823–2837.

Raha, K. and K. M. Merz Jr. 2004. A quantum mechanics-based scoring function: Study of zinc ion-mediated ligand binding. *J. Am. Chem. Soc*. 126(4): 1020–1021.

Raha, K. and K. M. Merz Jr. 2005. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J. Med. Chem*. 48(14): 4558–4575.

Reynolds, C. A. and M. K. Holloway. 2011. Thermodynamics of ligand binding and efficiency. *ACS Med. Chem. Lett*. 2(6): 433–437.

Sarwar, M. G., B. Dragisic, L. J. Salsberg, C. Gouliaras, and M. S. Taylor. 2010. Thermodynamics of halogen bonding in solution: Substituent, structural, and solvent effects. *J. Am. Chem. Soc*. 132(5): 1646–1653.

Seco, J., C. Ferrer-Costa, J. M. Campanera, R. Soliva, and X. Barril. 2012. Allosteric regulation of PKCθ: Understanding multistep phosphorylation and priming by ligands in AGC kinases. *Proteins: Struct., Funct., Bioinf*. 80(1): 269–280.

Sitkoff, D., K. A. Sharp, and B. Honig. 1994. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem*. 98(7): 1978–1988.

Spyrakis, F., A. Bidon-Chanal, X. Barril, and F. J. Luque. 2011. Protein flexibility and ligand recognition: Challenges for molecular modelling. *Curr. Topics Med. Chem*. 11(2): 192–210.

Srinivasan, J., M. W. Trevathan, P. Beroza, and D. A. Case. 1999. Application of a pairwise Generalized Born model to proteins and nucleic acids: Inclusion of salt effects. *Theor. Chem. Acc*. 101(6): 426–434.

Steinbrecher, T., D. L. Mobley, and D. A. Case. 2007. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.* 127(21): 214108.

Still, W. C., A. Tempczyk, R. C. Hawley, and T. Hendrickson. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* 112(16): 6127–6129.

Stoica, I., S. K. Sadiq, and P. V. Coveney. 2008. Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. *J. Am. Chem. Soc.* 130(8): 2639–2648.

Sulea, T., Q. Cui, and E. O. Purisima. 2011. Solvated Interaction Energy (SIE) for scoring protein–ligand binding affinities. 2. Benchmark in the CSAR-2010 scoring exercise *J. Chem. Inf. Model.* 51(9): 2066–2081.

Swanson, J. M. J., S. A. Adcock, and J. A. McCammon. 2005. Optimized radii for Poisson-Boltzmann calculations with the AMBER force field. *J. Chem. Theory Comput.* 1(3): 484–493.

Swanson, J. M. J., J. A. Wagoner, N. A. Baker, and J. A. McCammon. 2007. Optimizing the Poisson dielectric boundary with explicit solvent forces and energies: Lessons learned with atom-centered dielectric functions. *J. Chem. Theory Comput.* 3(1): 170–183.

Tan, C., L. Yang, and R. Luo. 2006. How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis. *J. Phys. Chem. B* 110(37): 18680–18687.

Tirado-Rives, J. and W. L. Jorgensen. 2006. Contribution of conformer focusing to the uncertainty in predicting free energies for protein–ligand binding. *J. Med. Chem.* 49(20): 5880–5884.

Tomasi, J. and M. Persico. 1994. Molecular interactions in solution: An overview of methods based on continuous distributions of the solvent. *Chem. Rev.* 94(7): 2027–2094.

Treesuwan, W. and S. Hannongbua. 2009. Bridge water mediates nevirapine binding to wild type and Y181C HIV-1 reverse transcriptase—Evidence from molecular dynamics simulations and MM/PBSA calculations. *J. Mol. Graphics Model.* 27(8): 921–929.

Tsui, V. and D. A. Case. 2001. Theory and applications of the Generalized Born solvation model in macromolecular simulations. *Biopolymers* 56(4): 275–291.

Wang, J. M., T. J. Hou, and X. Xu. 2006. Recent advances in free energy calculations with a combination of molecular mechanics and continuum models. *Curr. Comput.-Aided Drug Des.* 2(3): 287–306.

Warshel, A. 2003. Computer simulations of enzyme catalysis: Methods, progress, and insights. *Annu. Rev. Biophys. Biomol. Struct.* 32: 425–443.

Weis, A., K. Katebzadeh, P. Söderhjelm, I. Nilsson, and U. Ryde. 2006. Ligand affinities predicted with the MM/PBSA method: Dependence on the simulation method and the force field. *J. Med. Chem.* 49(22): 6596–6606.

Williams, D. H., E. Stephens, D. P. O'Brien, and M. Zhou. 2004. Understanding noncovalent interactions: Ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angew. Chem. Int. Ed.* 43(48): 6596–6616.

Wolfenden, R., L. Andersson, P. M. Cullis, and C. C. B. Southgate. 1981. Affinities of amino acid side chains for solvent water. *Biochemistry* 20(4): 849–855.

Wong, S., R. E. Amaro, and J. A. McCammon. 2009. MM/PBSA captures key role of intercalating water molecules at a protein-protein interface. *J. Chem. Theory Comput*. 5(2): 422–429.

Xu, L., H. Sun, Y. Li, J. Wang, and T. Hou. 2013. Assessing the performance of MM/PBSA and MM/GBSA methods. 3. The impact of force fields and ligand charge models. *J. Phys. Chem.* B 117(27): 8408–8421.

Yamagishi, J., N. Okimoto, G. Morimoto, and M. Taiji. 2014. A new set of atomic radii for accurate estimation of solvation free energy by Poisson-Boltzmann solvent model. *J. Comput. Chem*. 35(29): 2132–2139.

Yilmazer, N. D. and M. Korth. 2013. Comparison of molecular mechanics, semi-empirical quantum mechanical, and density functional theory methods for scoring protein-ligand interactions. *J. Phys. Chem. B* 117(27): 8075–8084.

Zhang, X., S. E. Wong, and F. C. Lighstone. 2014. Toward fully automated high performance computing drug discovery: A massively parallel virtual screening pipeline for docking and molecular mechanics/generalized born surface area rescoring to improve enrichment. *J. Chem. Inf. Model*. 54(1): 324–337.

Zhou, T., D. Huang, and A. Caflisch. 2010. Quantum mechanical methods for drug design. *Curr. Top. Med. Chem*. 10(1): 33–45.

Zhu, Y.-L., P. Beroza, and D. R. Artis. 2014. Including explicit water molecules as part of the protein structure in MM/PBSA calculations. *J. Chem. Inf. Model*. 54(2): 462–469.

Zoete, V. and O. Michielin. 2007. Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein–protein association using MM-GBSA: Application to the TCR-p-MHC complex. *Proteins: Struct., Funct., Bioinf*. 67(4): 1026–1047.