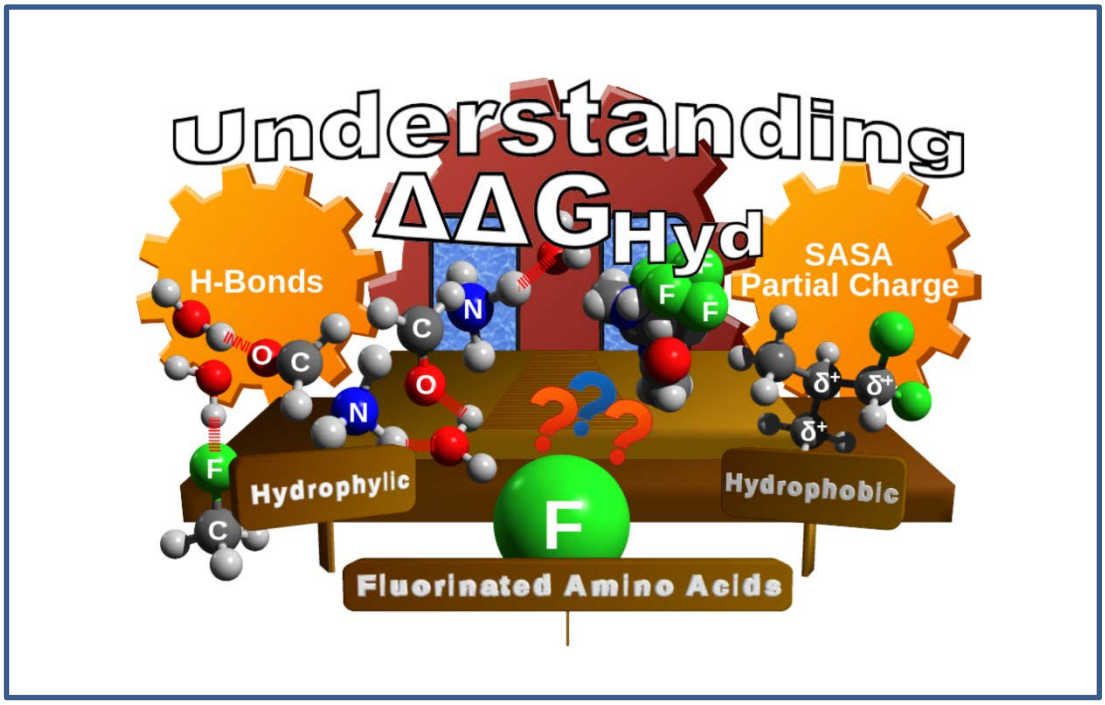
**Published as:**
Robalo, J. R., & Vila Verde, A. (2018). Unexpected Trends in the Hydrophobicity of Fluorinated Amino Acids Reflect Competing Changes in Polarity and Conformation. doi:10.26434/chemrxiv.7442648.v1.

# Unexpected Trends in the Hydrophobicity of Fluorinated Amino Acids Reflect Competing Changes in Polarity and Conformation

Robalo, João Ramiro & Vila Verde, Ana




The non-intuitive hydration free energy of fluorinated amino acids is calculated with molecular simulations and explained with an analytical model.

# Unexpected Trends in the Hydrophobicity of Fluorinated Amino Acids Reflect Competing Changes in Polarity and Conformation†

João R. Robalo and Ana Vila Verde*

Fluorination can dramatically improve the thermal and proteolytic stability of proteins and their enzymatic activity. Key to the impact of fluorination on protein properties is the hydrophobicity of fluorinated amino acids. We use molecular dynamics simulations, together with a new fixed-charge, atomistic force field, to quantify the changes in hydration free energy, $\Delta G_{Hyd}$, for amino acids with alkyl side chains and with 1 to 6 –CH→–CF side chain substitutions. Fluorination changes $\Delta G_{Hyd}$ by $-1.5$ to $+2$ kcal mol$^{-1}$, but the number of fluorines is a poor predictor of hydrophobicity. Changes in $\Delta G_{Hyd}$ reflect two main contributions: i) fluorination alters side chain-water interactions; we identify a crossover point from hydrophilic to hydrophobic fluoromethyl groups which may be used to estimate the hydrophobicity of fluorinated alkyl side-chains; ii) fluorination alters the number of backbone-water hydrogen bonds via changes in the relative side chain-backbone conformation. Our results offer a road map to mechanistically understand how fluorination alters hydrophobicity of (bio)polymers.

## 1 Introduction

The preferential interaction between apolar solutes in water – the hydrophobic effect – is a key factor driving protein folding[1,2], structural stability with respect to changes in temperature[3,4] (thermal stability) and interactions with other proteins and ligands[5,6]. The hydrophobic effect reflects the balance between solute-water interactions and direct, predominantly dispersive, solute-solute interactions[7,8]. Understanding how to use amino acid mutations to control the hydrophobic effect is critical to develop new protein-based drugs, biodevices and materials[9–13]. Simultaneously, minimizing changes in solute-solute packing upon mutations is desirable to ensure that protein structure – and thus function – is preserved[14,15]. This is, however, difficult with the limited pool of canonical hydrophobic amino acids because their side chains differ in structure and volume. Fluorinated versions of those amino acids, *i.e.* those where hydrogen atoms in side chain groups are substituted by fluorine (see Figure 1), can solve this problem while simultaneously enhancing other properties of interest[3,16–19]. Even fluorinating only a few residues may enhance the hydrophobicity and passive diffusion of peptides through membranes[20], the proteolytic resistance[21] and anti-microbial activity of proteins[22], in addition to tuning their thermal stability[23], making this synthetic approach of wide interest[24–28].

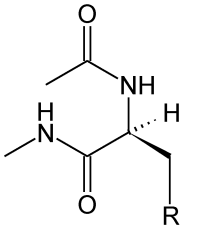
Still, a *caveat* of using fluorination to control protein properties remains: do we understand the factors influencing the hydrophobic effect involving fluorinated amino acids? The answer is, simply, that we do not. Here we focus on one factor contributing to the hydrophobic effect: solute-water interactions, for simplicity referred to as solute hydrophobicity. The hydrophobicity of solutes is usually characterized by their hydration free energies[8]. The hydrophobicity of fluorinated amino acids has been qualitatively estimated by considering the surface area of its side chain (the larger the surface area, the larger the hydrophobicity)[29,30] and its side chain polarity (the larger the polarity, the smaller the hydrophobicity)[23], but detailed mechanistic understanding is still lacking[16,19]. Understanding the origins of fluorination-induced changes in hydrophobicity depends critically on our ability to accurately quantify interactions between amino acids and their environment. We demonstrate that this quantification is now possible using molecular dynamics simulations and fixed-charge, all-atom models. The approach presented here is general, and may be used to investigate the hydrophobicity of any fluorinated (bio)polymer or small molecule.

## 2 Computational Methods

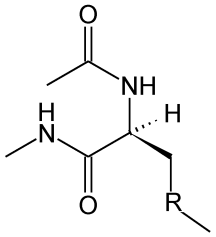
We used the TIP4P-Ew (ref. 31) water model, the AMBER14 (ff14sb; ref. 32) force field for the canonical amino acids and the GAFF force field for methane, ethane and propane. For the remaining amino acids and for the fluorinated small molecules, we used a force field developed by us (previous own work[33] and SI sections 1 and 2) based on AMBER14 (amino acids) or GAFF (small molecules). The main difference between our force field for fluorinated molecules and GAFF/AMBER14 lies in the Lennard-Jones (LJ) parameters of fluorine, and of hydrogen ($H_F$) bonded to a fluorinated carbon. The LJ parameters of fluorine were optimized to reproduce the hydration free energies of $CF_4$ and the molar volume of a 50% mix of $CF_4$ and $CH_4$; subsequently, those of $H_F$ were optimized to reproduce the hydration free energy and the molar volume of $CHF_3$. Atomic partial charges were obtained following the GAFF (small molecules)

*Max Planck Institute for Colloids and Interfaces, Department of Theory & Bio-systems, Science Park, Potsdam 14424 Germany. Fax: +49 (0)331 5679602; Tel: +49 (0)331 5679608; E-mail: ana.vilaverde@mpikg.mpg.de*

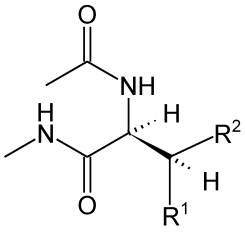
† Electronic Supplementary Information (ESI) available: Description of force fields and simulation details, hydration free energies, electrostatic potential in the vicinity of amino acid side chains, linear models to reproduce the calculated hydration free energies, radial distribution functions, conformational changes in fluorinated side chains. AMBER-format force fields for mono- and di-fluorinated amino acids available upon request.



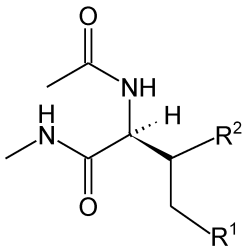
| | |
|---|---|
| $R = CH_3$ | Ethylglycine (**ETG**) |
| $R = CH_2F_1$ | 4-Monofluoroethylglycine (**E1G**) |
| $R = CHF_2$ | 4,4-Difluoroethylglycine (**E2G**) |
| $R = CF_3$ | 4,4,4-Trifluoroethylglycine (**E3G**) |



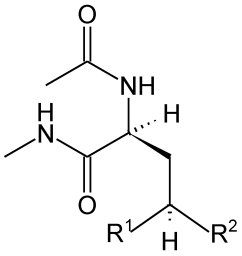
| | |
|---|---|
| $R = CH_2$ | Propylglycine (**PRG**) |
| $R = CF_2$ | 4,4-Difluoropropylglycine (**P2G**) |



| | |
|---|---|
| $R^1 = R^2 = CH_3$ | Valine (**VAL**) |
| $R^1 = CH_3$, $R^2 = CF_3$ | 4,4,4,-Trifluorovaline (**V3S**) |
| $R^1 = CF_3$, $R^2 = CH_3$ | 4',4',4',-Trifluorovaline (**V3R**) |
| $R^1 = R^2 = CF_3$ | 4,4,4,4',4',4'-Hexafluorovaline (**V6G**) |



| | |
|---|---|
| $R^1 = R^2 = CH_3$ | Isoleucine (**ILE**) |
| $R^1 = CH_3$, $R^2 = CH_2F$ | 4'-Monofluoroisoleucine (**I1G**) |
| $R^1 = CF_3$, $R^2 = CH_3$ | 5,5,5-Trifluoroisoleucine (**I3D**) |
| $R^1 = CH_3$, $R^2 = CF_3$ | 4',4',4'-Trifluoroisoleucine (**I3G**) |



| | | | |
|---|---|---|---|
| $R^1 = R^2 = CH_3$ | Leucine (**LEU**) | | |
| $R^1 = CH_3$, $R^2 = CH_2F$ | 5-Monofluoroleucine (**L1S**) | $R^1 = CH_3$, $R^2 = CF_3$ | 5,5,5-Trifluoroleucine (**L3S**) |
| $R^1 = CH_2F$, $R^2 = CH_3$ | 5'-Monofluoroleucine (**L1R**) | $R^1 = CF_3$, $R^2 = CH_3$ | 5',5',5'-Trifluoroleucine (**L3R**) |
| $R^1 = R^2 = CHF_2$ | 5,5,5',5'-Tetrafluoroleucine (**L4D**) | $R^1 = R^2 = CF_3$ | 5,5,5,5',5',5'-Hexafluoroleucine (**L6D**) |

**Fig. 1** Molecular structures, commonly used names and abbreviations for the amino acids under study. Each amino acid residue is capped at the N-terminus with an acetate group (ACE, $-COCH_3$) and at the C-terminus with an N-methyl group (NME, $-NHCH_3$). Abbreviations for fluorinated amino acids follow a three-character nomenclature: initial character of parent (non-fluorinated) amino acid name (E, P, V, I, L); number of fluorine atoms (1, 2, 3, 4, 6); fluorination site ($\delta$ carbon as D, $\gamma$ carbon as G or, in the case of chiral center formation following fluorination, R or S).

or AMBER (amino acids) protocols. We note that the requirement of compatibility with the protein force field implies that we retain the point charge representation. This representation has known shortcomings when modeling carbon-bound chlorine, bromine or iodine because these halogens exhibit $\sigma$-holes (a positive area in the electrodensity distribution of the halogen atom, surrounded by a negative belt) which this charge model cannot represent[34]. Carbon-bound fluorine, however, retains a negative potential in its entire surface and the level of anisotropy in its electronic charge distribution is small[35–37]. For that reason, single point charge models have been used to model the interactions of fluorinated sites with water and with other organic molecules: e.g., Schyman & Jorgensen[34], Ibrahim[38] and Ho[39] apply their models of $\sigma$-holes to carbon-bound chlorine, bromine and iodine, but retain a point charge representation for fluorinated sites. Point charge models perform less well in the case of fluorinated aryl groups because the $\sigma$-hole perturbation extends into the $\beta$-carbons; when modeling fluorinated aryl molecules, other promising models such as those based on permanent atomic multipole charges[37,40,41] should be considered. This extended perturbation occurs via the $\pi$ electrons[37], and consequently is not expected to be nearly as significant for fluorinated alkyl groups, which are the sole focus of the present work.

Free energy calculations were performed with Gromacs 5.0[42–48] and molecular dynamics simulations to calculate other observables were performed with AMBER 14[49]. All systems were assembled using the built-in tools of the software package used to perform the simulations. A summary of the most relevant parameters used during the production runs is given in SI Table 2. Simulations used a time-step of 2 fs and constraints (LINCS[50] in Gromacs, SHAKE[51] in Amber) were applied to all bonds involving hydrogen atoms. Integration of the equations of motion was done using a leap-frog Langevin algorithm. Van der Waals interactions were shifted to zero between 1.0 and 1.2 nm, and long-range dispersion corrections were applied to both pressure and energy. Long-range electrostatics were treated with the PME scheme with a 1.2 nm cutoff, a grid spacing of 0.1 nm (AMBER) or 0.12 nm (Gromacs) and a 4th (AMBER) or 6th (Gromacs) order interpolation. Production runs were done in the NpT ensemble. The Monte Carlo[49,52] (AMBER) or Berendsen[53] (Gromacs) barostats were used with a relaxation time of 1 ps for an isotropic coupling of system pressure to 1 bar; temperature coupling was handled by the leap-frog Langevin integrator with a collision frequency of 1 $ps^{-1}$ and a target temperature of 298 K.

Hydration free energies were calculated using Free Energy Perturbation (FEP) and Bennett Acceptance Ratio[54,55] (BAR), following the protocol we have previously adopted[33] and described in SI section 2.1. Briefly, in each case, simulations consisting of a single solute molecule in water were conducted, first decoupling Coulombic interactions and then LJ interactions. The coupling parameter $\lambda_C$ for the Coulombic interactions was scaled linearly and assumed 21 equally-spaced values between 0 and 1; decoupling the LJ interactions was done over 59 states, with unevenly-spaced $\lambda_{LJ}$ values between 0 and 1. At each state, we performed a steepest-descent minimization, BFGS minimization, 100 ps of NVT equilibration and 100 ps of NpT equilibration before collecting statistics for each state over 2 ns. Five independent production runs were performed for each amino acid.

Molecular dynamics simulations consisted of a steepest descent minimization, a conjugated gradient minimization, a 200 ps NVT heating from 0 K to 298 K, a 1 ns NpT equilibration and a 25 ns NpT production run; in the minimization, heating and equilibration steps the coordinates of the backbone atoms were restrained with a 20 kcal $mol^{-1}$ potential. These trajectories were used to extract the data used as input for Equation 1, and were also used as input for the APBS[56] software to estimate the electrostatic hydration free energy of the amino acids using the linearized Poisson-Boltzmann equation, as described in more detail in SI section 3.1.

## 3 Results & Discussion

**Change in $\Delta G_{Hyd}$ with fluorination depends on the chirality and location of the fluorinated site and on amino acid identity.** We calculated hydration free energies as a measure of the hydrophobicity of 16 fluorinated amino acids and their 5 non-fluorinated counterparts, totaling 21 aliphatic amino acids (see Figure 1). These free energies are shown in SI Table 4 and Figure 2; we reported some of these values in a prior publication[33]. The $\Delta\Delta G_{Hyd}$ values have an associated standard deviation of 0.1 to 0.3 kcal $mol^{-1}$ (see SI Table 4), enabling the precise detection of differences between amino acids.

The amino acids with one or more $-CH_3 \rightarrow -CF_3$ substitutions (here termed fully fluorinated) are, as expected, always more hydrophobic than their non-fluorinated counterparts (positive $\Delta\Delta G_{Hyd}$)[3,19,57], but the change in free energy is not constant per fluorinated group. Even more surprisingly, amino acids with $-CH_3 \rightarrow -CH_2F/-CHF_2$ and $-CH_2- \rightarrow -CF_2-$ substitutions (here termed partially fluorinated) display a range of $\Delta\Delta G_{Hyd}$ values from $-1.5$ kcal $mol^{-1}$ to $+1$ kcal $mol^{-1}$, regardless of the number of fluorine atoms. $\Delta\Delta G_{Hyd}$ depends strongly and non-intuitively on the chirality (R/S), location ($\gamma$ *vs.* $\delta$) of the fluorinated sites, and on the identity of the amino acid.

Experimental hydration free energies for amino acids are not available, so we cannot directly assess the accuracy of our predictions. To test the accuracy of our force field, we applied it to *all* fluorinated variants of methane, ethane and propane for which experimental hydration free energies could be found[58]. These small molecules are the closest analogues to the side chains of the amino acids investigated here. The free energies of hydration for the fluorinated small molecules are shown in Figure 3 and SI Table 4. We show also the free energies of hydration of methane, ethane and propane, to illustrate the accuracy of the AMBER force field for the alkyl side chains of amino acids. The force field for fluorinated molecules reproduces the experimental hydration free energy very well in most cases. The largest deviation is seen for $CH_3F$, whose $\Delta G_{Hyd}$ is 0.8 kcal $mol^{-1}$ too negative. The predicted hydration free energies for methane, ethane and propane, in contrast, are too positive by 0.5 to 0.8 kcal $mol^{-1}$. These results suggest that the predicted $\Delta\Delta G_{Hyd}$ for amino acids (Figure 2) may be systematically too negative by 0.5 to 0.8 kcal $mol^{-1}$. Differences between di- and trifluorinated amino acids should be well captured, but monofluorinated alkyl groups are likely excessively

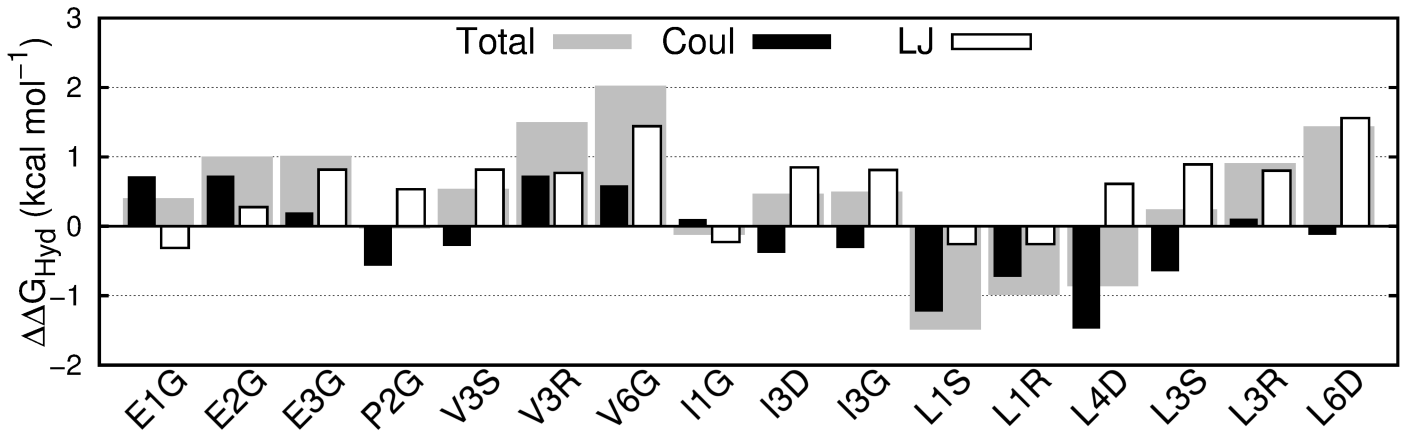



**Fig. 2** Coulombic contribution (Coul), Lennard-Jones contribution (LJ) and sum of the contributions (total) to the differences in hydration free energy ($\Delta\Delta G_{Hyd}$) between fluorinated and non-fluorinated amino acids. Each bar is the average of five independent simulations.

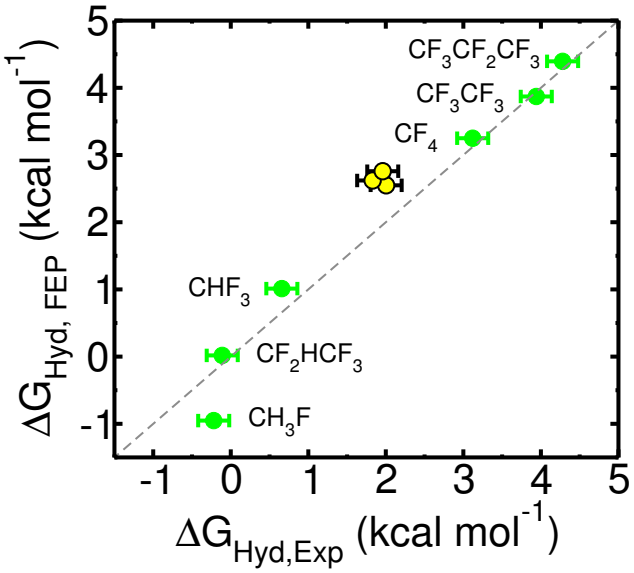



**Fig. 3** Free energy of hydration ($\Delta G_{Hyd}$) of the indicated fluorinated molecules (green), and of methane, ethane and propane (yellow), obtained with simulation (FEP) or experimentally (Exp; from ref. 58). $CF_4$ and $CHF_3$ were used during parameterization. The statistical uncertainty in the FEP values is of the size of the symbol. See SI Table 4 for free energy values.

hydrophilic in our model. Our main result – the large dependence of $\Delta\Delta G_{Hyd,FEP}$ on the identity of the amino acid and characteristics of the fluorinated site – is not affected by these force field shortcomings. Below we show that this dependence is in fact largely due to fluorination-induced conformational changes that alter the number of backbone-water hydrogen bonds. The AMBER force field for proteins, upon which we build the force field for fluorinated amino acids, has been extensively improved for decades and is able to reproduce experimentally-measured structure and dynamics of folded small proteins[59–63]. Fluorination-induced changes in conformation in our simulations result from steric hindrance and favorable electrostatic carbonyl-CF interactions, both of which can be captured, at least qualitatively, by all-atom, fixed charge force fields.

**The surprising variation in $\Delta\Delta G_{Hyd}$ is largely Coulombic in origin.** Decomposing the hydration free energy into Lennard-Jones and Coulombic contributions can be naturally done in simulations, and gives valuable insight into the origin of the observed trends. The Lennard-Jones contribution to $\Delta\Delta G_{Hyd}$ (Figure 2) is constant and positive *per* fluorinated group, positive but not constant for difluorinated amino acids and negative and constant for monofluorinated amino acids. Despite this variation in the LJ contribution to the free energy of hydration, the wide variation in $\Delta\Delta G_{Hyd}$, particularly in the case of the partially fluorinated amino acids, is actually dominated by the Coulombic contribution to the free energy. This contribution varies seemingly unpredictably (Figure 2), with each type of fluorination leading to either positive or negative $\Delta\Delta G^{Coul}_{Hyd}$: *e.g.*, compare E2G with P2G, E1G with L1S, L1R and I1G, V3S with V3R. Previous reports on the dependence of lipophilicity on fluorine-induced polarity changes support the idea that the contribution of electrostatics to hydrophobicity is far from intuitive[64,65]. Neither contribution shows visible correlation with local hydration around the fluorinated site, as measured by the radial distribution function of

water around the fluorinated sites (SI section 5.2.4 and Figure 8 A,B).

**Developing an analytical solvation model to understand how fluorination affects $\Delta G_{Hyd}$.** Can we understand the origin of these hydration free energies? To answer this question we model the fluorination-induced change in hydration free energy in terms of linear, multivariate models of the form $\Delta\Delta G_{Hyd} = \sum_i k_i \cdot \Delta Y_i$, where $k_i$ are fitting parameters and $Y_i$ are observables that should affect the hydration free energy and can be easily calculated in short molecular dynamics simulations. The LJ contribution to $\Delta\Delta G_{Hyd}$ is dominated by the energy (work) required to form a solute-sized cavity in water to accommodate the larger fluorine atoms. This contribution (Table 1) is proportional to the change in solvent accessible surface area (ΔSASA) of the amino acid[66,67], and can be easily quantified by measuring ΔSASA in molecular dynamics simulations, and multiplying it by the LJ component of $\Delta G_{Hyd}$ *per* surface area unit of methane, which is essentially identical to that of $CF_4$ (SI Table 9).

The polar contribution is dominated by the interaction of a distribution of atom-centered point charges with water. This contribution can be estimated in multiple manners[33,64,65,68]. We first attempted to model the polar contribution as the sum of three terms, one proportional to a global quantity, the dipole moment $\mu$, representing the molecular charge distribution, and the other two proportional to local quantities, the number of hydrogen bonds between water and amines ($h_{NH}$) or carbonyls ($h_{CO}$) in the backbone*. Fitting the $\Delta\Delta G_{Hyd}$ values calculated with FEP using a linear multivariate model (SI Equation 5) consisting of the sum of the contributions arising from changes in SASA, $\mu$, $h_{NH}$, and $h_{CO}$ due to fluorination was unsuccessful: the resulting fitting parameters had unphysical values, *e.g.*, a positive energetic contribution of the dipole moment, and overly large errors (SI Table 6). We also attempted to model our data using an analogous version of this model, but where the area term reflects the difference in hydration free energy between $CH_4$ and $CF_4$. This second model (SI Equation 6) has proven successful to understand hydrophobicity of tri- and hexa-fluorinated amino acids[33], but it fails (SI Figure 4) when applied to the partially fluorinated amino acids. Given that the contributions of solute-water hydrogen bonds and the Lennard-Jones interactions to the hydration free energies are well-known[33,66,68], we interpret these results as an inability of the molecular dipole moment to describe electrostatic solute-water interactions in a quantitative manner, at least for solutes as diverse as the current set of amino acids. We next attempted to characterize how fluorination alters the solute-water electrostatic interactions with another commonly used global descriptor: solving the linearized Poisson-Boltzmann (PB) equation, where the solvent is modeled as a continuum, as described in SI Section 3.1. Given the apparently simple problem we were trying to model, we were surprised to find that the electrostatic component of the hydration free energy calculated using PB hardly correlates with the reference FEP values (SI Figure 2 A). The absence of correlation suggests that it is imperative to model the solvent as discrete water molecules. Thus, when aiming for a correct characterization of how aqueous solvation of biopolymers changes with mutations, not only the solute but also the solvent must be modeled without the use of global or mean-field descriptors.

Backbone-water hydrogen bonds dominate the polar interactions between the backbone and water[68]; the unresolved issue is how to describe polar interactions between the side chain and water. We find that these can be characterized by the electrostatic potential, Φ, at the position of the water oxygen atoms in the hydration shell of the side chains, as described in SI Section 4. The corresponding probability distributions of Φ for the fluorinated amino acids, shown in SI Figure 2 B-F, show large negative potential regions together with, in some cases, regions of more positive potentials than observed for the parent amino acid. The more positive potential, almost exclusive to mono- and difluorinated species, arises from a larger exposure of the positively charged carbon skeleton of the side chain, left partially unshielded by hydrogen in mono- and difluorinated groups (see SI Section 4 and SI Figure 3). The negative potential region, observed for all amino acids, can be attributed to the fluorine atoms. Water molecules near fluorine atoms often assume configurations that meet the geometric criteria for HOH···FC hydrogen bonds, as discussed in SI section 5.2, so we consider that these weak hydrogen bonds exist. This interpretation is consistent with ab initio calculations, and spectroscopic measurements[69–72]; the strong correlation between $^{19}$F NMR isotropic chemical shifts and the type of fluorine-protein interactions observed in the Protein Data Bank also suggest that hydrogen bonds to fluorinated alkyl groups exist, and are strongest for groups with low degrees of fluorination[73], as we also observe (SI Table 7).

Our final model (Equation 1) reflects the above results. We capture the impact of the fluorination-induced differences on side chain-water interactions in two ways, *via* the average number of water–fluorine hydrogen bonds established by each amino acid (the $h_{CH_nF_m}$ terms, where $n = 0,1,2$ and $m = 1,2,3$, in Equation 1) and *via* the fluorination-induced change in the number of water molecules experiencing the positive potential region of the side chain (the $\Delta\Phi^+$ term in Equation 1). Fitting Equation 1 to the $\Delta\Delta G_{Hyd}$ data shown in Figure 2 yields the parameters in Table 1; see SI section 5.2 for details of the fit.

$$\begin{aligned}\Delta\Delta G_{Hyd} = {} & k_1\Delta A + k_2\Delta h_{CO} + k_3\Delta h_{NH} + k_4\Delta\Phi^+ \\ & + k_5 h_{CH_2F} + k_6 h_{CF_2} + k_7 h_{CF_3}\end{aligned} \quad (1)$$

Figure 4 shows the correlation between the FEP-calculated hydration free energies and the ones calculated using Equation 1. The model describes the changes in $\Delta G_{Hyd}$ following fluorination for most cases, with an average deviation between model and FEP results of only 0.26 kcal mol$^{-1}$. Poorer agreement occurs for E1G, for which it yields a value of $\Delta\Delta G_{Hyd}$ which deviates 0.5 kcal mol$^{-1}$ from the value given by the FEP simulations and actually has the wrong sign, and for E2G, L4D and L3S, for which deviations are 0.7, 0.5 and 0.5 kcal mol$^{-1}$ each. Our prior work suggests that the source of these large deviations might be

* Hydrogen bonds exist if the O···O distance <3.5 Å and the O-H···O angle is between 135° and 180°

| Fitting Parameter | Value | Error | P-value |
|---|---|---|---|
| $k_1$ ($\Delta$A; kcal mol$^{-1}$ Å$^{-2}$) | 0.053 | 0.001* | NA |
| $k_2$ ($\Delta h_{CO}$; kcal mol$^{-1}$ H-Bond$^{-1}$) | -3.590 | 0.350 | 0.000 |
| $k_3$ ($\Delta h_{NH}$; kcal mol$^{-1}$ H-Bond$^{-1}$) | -3.020 | 0.780 | 0.012 |
| $k_4$ ($\Delta\Phi^+$; kcal mol$^{-1}$ $H_2O^{-1}$) | 0.115 | 0.044 | 0.022 |
| $k_5$ ($h_{CH_2F}$; kcal mol$^{-1}$ H-Bond$^{-1}$) | -2.619 | 0.396 | 0.000 |
| $k_6$ ($h_{CF_2}$; kcal mol$^{-1}$ H-Bond$^{-1}$) | -1.808 | 0.270 | 0.000 |
| $k_7$ ($h_{CF_3}$; kcal mol$^{-1}$ H-Bond$^{-1}$) | -0.780 | 0.222 | 0.004 |

**Table 1** Values of the fitting parameters from Equation 1, and associated standard errors and P-values. NA: not applicable; * calculated via error propagation.

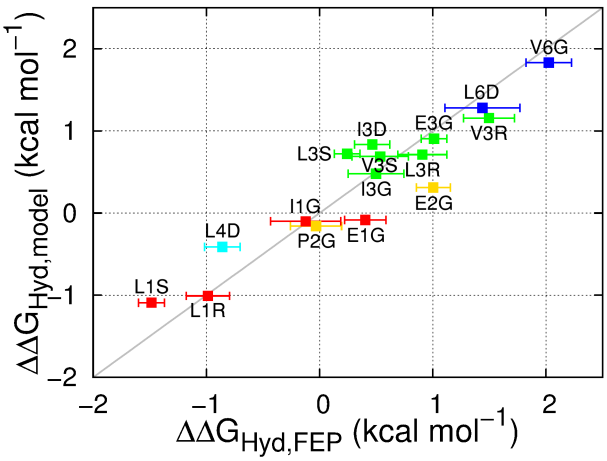



**Fig. 4** Correlation between the hydration free energy differences between fluorinated and non-fluorinated amino acids calculated with FEP ($\Delta\Delta G_{Hyd,FEP}$) or Equation 1 ($\Delta\Delta G_{Hyd,model}$). Data points are presented as mean ± standard deviation of five independent simulations. The color code indicates the number of fluorine atoms: red = one; yellow = two; green = three; cyan = four; blue = six. The gray line indicates perfect correlation.

entropic[33], and we speculate it might be related to changes in solute conformational entropy, which are not included in Equation 1. Clarifying this point is outside of the scope of the present work.

**Decomposing the contributions to $\Delta\Delta G_{Hyd}$.** The performance of the multivariate model is sufficiently good to enable insight into the mechanisms of fluorination-induced changes in solvation, by decomposing the individual contributions to $\Delta\Delta G_{Hyd}$ as seen in Figure 5.

*Surface area.* Within derivatives of the same amino acid, the surface area increases $\Delta\Delta G_{Hyd}$ proportionally to the number of fluorine atoms (SI Figure 5). The magnitude of the increase *per* fluorine atom depends on the amino acid identity, which implies that the hydration shells around each side chain are disturbed to a different degree, when accommodating the hydrogen → fluorine substitution – compare, *e.g.* E2G with P2G.

*Backbone-water hydrogen bonds.* The contribution of hydrogen bonds between water and carbonyl groups is always positive because fluorination reduces the number of these hydrogen bonds by steric blockage. The magnitude of this contribution, for fluorinated variants of a given amino acid, is again proportional to the number of fluorine atoms in the side chain. In contrast, the contribution of amine–water hydrogen bonds varies between positive and negative because fluorination may increase or decrease the number of these hydrogen bonds. Steric blockage occurs because of fluorine's large size and, for some amino acids, because fluorination changes the preferential conformation of the side chain as discussed in SI section 6. These results are consistent with previous reports indicating that CF and carbonyl groups interact favorably[74,75], and that changes in the conformational preference of fluorinated alkyl groups affect a molecule's lipophilicity, membrane permeability and inhibitory activity[25,76,77].

*Side-chain polarity.* The most interesting contributions arise from the polarity of the fluorinated side chain. The large, positive electrostatic potential affecting hydration waters adds an average +0.4 kcal mol$^{-1}$ to the $\Delta\Delta G_{Hyd}$ of all partially fluorinated amino acids, and a near-zero, positive, contribution to the trifluorinated amino acids; the largest deviations come from I1G and I3D, for which this contribution is +0.8 kcal mol$^{-1}$. As indicated above, this contribution stems from the partial shielding of the positively charged carbons occurring in partially fluorinated groups. Regarding the water-fluorine hydrogen bonds, they are much weaker than those with the carbonyl or amine groups, and decrease in stability in the order –$CH_2F$ > –$CHF_2$/–$CF_2$– > –$CF_3$, as indicated by the relative magnitudes of the relevant parameters in Table 1. These trends are expected: our simulations show that the distance between the water oxygen and the fluorine is smaller in groups with fewer fluorines (SI Table 7) indicating an increase in hydrogen bond strength, and other experiments and ab initio calculations have shown the same trends[73,78]. Despite the weakness of the hydrogen bonds between water and the di- and tri-fluorinated groups, they nevertheless play an important role: *e.g.*, they are present ≈15% of the time per fluorine in $CF_3$ groups (SI Table 8); for comparison, the number of water-methyl configurations per side chain CH group meeting the hydrogen bond criteria in the non-fluorinated amino acids is almost negligible (≈3%). The magnitude of the hydrogen bond contributions to $\Delta\Delta G_{Hyd}$ *per* fluorine atom also follows the order –$CH_2F$ > –$CHF_2$/–$CF_2$– > –$CF_3$, with the average values being −1.79, −0.60 and −0.13 kcal mol$^{-1}$ *per* fluorine, respectively†; the weak water-fluorine hydrogen bonds to –$CF_3$ contribute on average a non-negligible -0.39 kcal mol$^{-1}$ –$CF_3^{-1}$.

**Contributions of the side chain to $\Delta\Delta G_{Hyd}$ are largely constant per $CF_x$ substituent.** The values of the free energy contribution per water-fluorine hydrogen bond yielded by the multivariate model anticorrelate surprisingly well with the LJ contribution of the area *per* fluorine atom (SI Figure 6), suggesting that the energy cost associated with (overall repulsive) water–fluoromethyl LJ interactions is partially offset by the energy gain from the formation of water–fluorine hydrogen bonds. This point is illustrated in Figure 6, where the ratio of the area and water-fluorine hydrogen bond contributions is plotted against the number of fluorine atoms. It is clear that the contribution of each

† Values calculated from the relevant $h_{CH_mF_n}$ values in SI Table 8 and the corresponding $k$ parameter in Table 1.

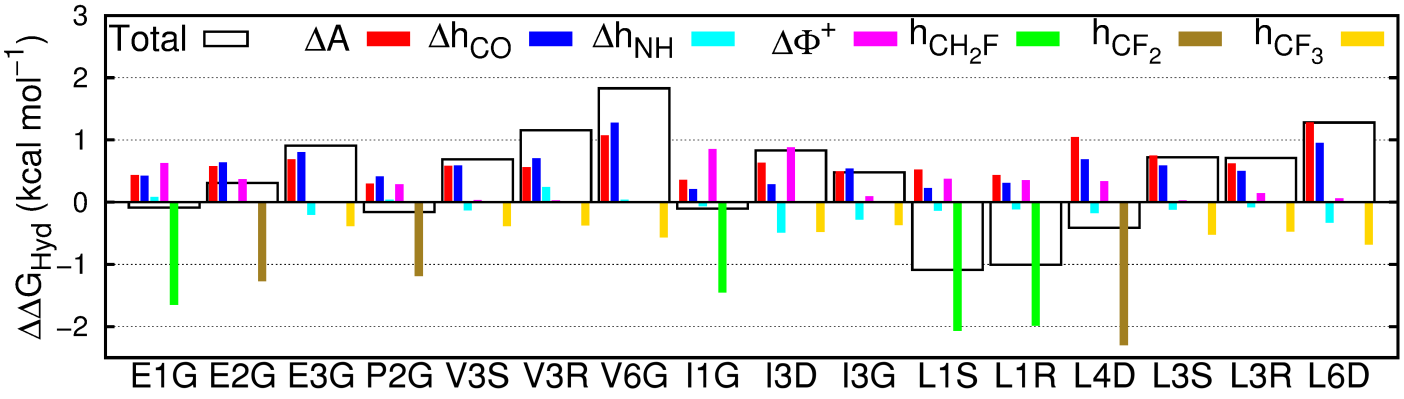



**Fig. 5** Contribution of the changes in surface area ($\Delta A$), hydrogen bonds between water and carbonyls ($\Delta h_{CO}$) or amines ($\Delta h_{NH}$), water molecules exposed to a large positive electrostatic potential ($\Delta\Phi^{+}$) and hydrogen bonds between water and fluorine in mono- ($h_{CH_2F}$), di- ($h_{CF_2}$) or trifluorinated ($h_{CF_3}$) amino acids to the total change in hydration free energy ($\Delta\Delta G_{Hyd}$), between fluorinated and non-fluorinated amino acids, calculated using Equation 1.

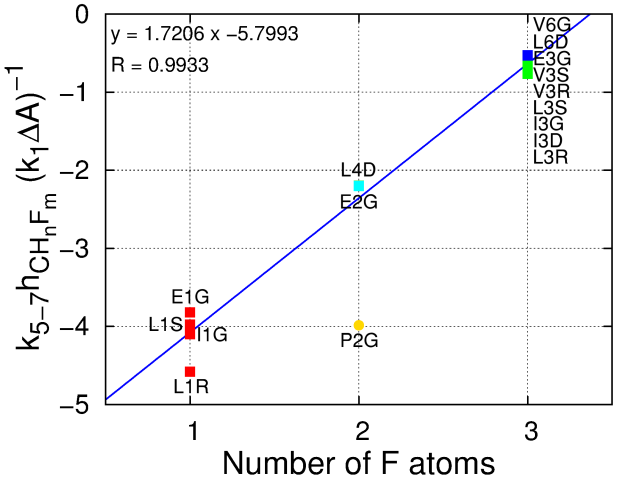



**Fig. 6** Ratio of the contribution of the water–fluorine hydrogen bonds to the contribution of the surface area to the $\Delta\Delta G_{Hyd}$ ($k_{5-7}h_{CH_nF_m}(k_1\Delta A)^{-1}$) *versus* the number of fluorine atoms *per* fluoromethyl group. Contributions are calculated using Equation 1. The color code indicates the number of fluorine atoms: red = one; yellow = two; green = three; cyan = four; blue = six. The blue line is a linear fit to the data points (excluding P2G, shown as a circle) with the corresponding equation and regression coefficient at the top left.

fluorinated group to the hydration free energy is fairly constant. P2G, bearing the only non-terminal fluorination site in this set of amino acids, has a much lower cavity-forming penalty, thereby escaping this trend. The notable dependency of $\Delta\Delta G_{Hyd}$ on the chirality and location of the fluorinated site and the identity of the the amino acids cannot be explained solely in terms of local changes in solvation around the fluorinated site.

For one hydrogen → fluorine substitution, the hydrogen bonding over-compensates the cavity-formation cost. As the number of substitutions increases, the penalty for cavity formation per fluorine is reduced, showing that the perturbation due to the insertion of the first fluorine atom is large but further insertions perturb the water network to a smaller extent. Simultaneously, the energy gain per fluorine from water–fluorine hydrogen bonds is decreased for multiple fluorine insertions, likely due to the less negative partial charges on fluorine and the fewer $H_F$ atoms, which interact favorably with water (see SI section 2.3). Interestingly, there is a crossover point, associated with the number of fluorine substitutions, after which the energy cost surpasses the energy gain; in other words, there is a transition between a *locally hydrophilic* moiety to a *locally hydrophobic* moiety. Inspecting Figure 6 we find that the number of fluorine atoms required to perform the transition is 2.8. This crossover point has associated uncertainty arising from the systematic deviations in the hydration free energies observed for small alkanes with the GAFF/Amber force field (see Figure 3 and discussion above). Even considering this uncertainty, it appears that trifluoromethyl groups, by themselves, impart only a small increase in amino acid hydrophobicity, because the positive Lennard-Jones component of the hydration free energy is partially offset by the weak but still favorable water-fluorine hydrogen bonds.

**RDFs of water around the fluorinated site do not give insight into local contributions to the $\Delta G_{Hyd}$.** Given that the impact of fluorination on local solvation is fairly constant for groups with the same number of fluorine atoms (Figure 6), we investigated whether these local changes in solvation correlate well with the radial distribution function (RDF) of water around methyl or fluoromethyl groups. Specifically, we calculated the

excess free energy, $\Delta\Delta G_{Shell,PMF}$, necessary to populate the first hydration layer of methyl or fluoromethyl groups from the potential of mean force associated with each radial distribution function. Our results (SI section 5.2.4, Figure 8 C) show that such a correlation is at best very weak. We could also not find correlations between $\Delta\Delta G_{Shell,PMF}$ and the Coulomb and Lennard-Jones components of the hydration free energy (SI Figure 8A,B). Radial distribution functions describe solvation along a single reaction coordinate, which may be insufficient to give quantitative or semi-quantitative insight into local hydrophobicity. In contrast, proximal radial distribution functions – i.e., those calculated perpendicular to the solute atoms – appear near universal for proteins[79] and have proven useful to estimate hydration free energies of small molecules and peptides[80,81]. Exploring their usefulness for fluorinated molecules is outside the scope of the present work.

## 4 Concluding Remarks

We present an all-atom, fixed-charge force field for amino acids with fluorinated alkyl side chains that is compatible with the AMBER force field for proteins. With it we investigate how mono-, di- and trifluorination alters amino acid solvation. Our predictions indicate that side chain fluorination alters the hydration free energy of amino acids in surprising ways: $\Delta\Delta G_{Hyd}$ strongly depends on the chirality and location of the fluorinated site and on the identity of the amino acid. Using a simple, analytical solvation model (Equation 1), we trace back these dependencies to the multiple mechanisms by which fluorination alters solvation of amino acids: there is a cost of introducing larger fluorine atoms, gains *and costs* associated with the higher polarity of fluorinated alkyl groups, and gains or costs from altering the number of *backbone*-water hydrogen bonds as a result of changed conformational preferences. For small molecules, it is often possible to predict the sign and even estimate the magnitude of the change in hydrophobicity upon fluorination. In contrast, for complex molecules 'the devil is in the details': the contribution of each mechanism to the overall hydrophobicity depends on conformational preferences and interactions between different parts of the molecule, making rules-of-thumb insufficient. For example, monofluorination *does not* always make amino acids more hydrophilic; similar increases in the solvent-exposed surface area of different molecules *do not* imply that the molecules will experience similar increases in hydrophobicity. Solvent accessible surface area descriptors of hydration free energies remain useful for complex molecules, but only when other contributions are properly accounted for.

The solvation model given by Equation 1 and applied here to amino acids can also be used to interpret molecular dynamics results of other small molecules containing the same functional groups, and extended for other functional groups. The model is also directly relevant for proteins: together with short molecular dynamics simulations of proteins in the folded and unfolded ensembles, it can be used to gain insight into how fluorination-induced changes in protein-water interactions contribute to changes in the free energy of folding. Good sampling of the folded protein ensemble can easily be achieved in many cases with molecular dynamics simulations; to sample the unfolded protein ensemble, one can take advantage of a number of algorithms[82–86]. Future work by our group will attempt to extend the solvation model to include mechanisms by which fluorination alters intra-protein non-bonded interactions. This extension is necessary to obtain a complete picture of the mechanisms by which fluorination alters the thermal stability of proteins.

Molecular dynamics studies with custom-tailored force fields and phenomenological models based on discrete rather than mean-field descriptors are key to gain mechanistic insight on solvation, as exemplified in this work. The force field and model we present lay the foundation to interpret how fluorination alters the hydrophobicity of other (bio)polymers. Further improving these force fields and our understanding of solvation will require the experimental measurement of free energies of hydration for amino acids or molecules of similar complexity. If the vapor pressure of the pure compound in liquid form is known together with its aqueous solubility, the free energy of hydration can be immediately calculated assuming ideality[58]. We deliberately restricted our study to amino acids for which synthesis protocols exist[19], and we hope that a direct comparison between experiment and simulation will be possible in the future.

## Acknowledgments

JRR was funded by the International Max Planck Research School on Multiscale Bio-Systems. We thank Marco Ehlert and René Genz from the IT team of the MPIKG for their help. We thank Dr. Valerio Molinari and Dr. Alexandra Latnikova for their critical reading of the manuscript.

## References

1 R. Zhou, X. Huang, C. J. Margulis and B. J. Berne, *Science*, 2004, **305**, 1605–1609.
2 R. C. Harris and B. M. Pettitt, *J. Phys. Cond. Matter*, 2016, **28**, 083003.
3 E. N. G. Marsh, *Acc. Chem. Res.*, 2014, **47**, 2878–2886.
4 C. N. Pace, H. Fu, K. L. Fryar, J. Landua, S. R. Trevino, B. A. Shirley, M. M. Hendricks, S. Iimura, K. Gajiwala, J. M. Scholtz and G. R. Grimsley, *J. Mol. Bio.*, 2011, **408**, 514–528.
5 S. Ye, B. Loll, A. A. Berger, U. Mülow, C. Alings, M. C. Wahl and B. Koksch, *Chem. Sci.*, 2015, **6**, 5246–5254.
6 A. M. Davis and S. J. Teague, *Angew. Chem. Int. Ed.*, 1999, **38**, 736–749.
7 D. Ben-Amotz, *J. Phys. Chem. Lett.*, 2015, **6**, 1696–1701.
8 D. Ben-Amotz, *Annu. Rev. Phys. Chem.*, 2016, **67**, 617–638.
9 K. A. Brogden, *Nat. Rev. Microbiol.*, 2005, **3**, 238–250.
10 M. Zasloff, *Nature*, 2002, **415**, 389–395.
11 S. T. Henriques, M. N. Melo and M. A. R. B. Castanho, *Biochem. J.*, 2006, **399**, 1–7.
12 Y.-W. Cui, H.-Y. Zhang, J.-R. Ding and Y.-Z. Peng, *Sci. Rep.*, 2016, **6**, 24825.
13 B. J. Harris, X. Cheng and P. Frymier, *J. Phys. Chem. B*, 2016, **120**, 599–609.
14 B. C. Buer, J. L. Meagher, J. A. Stuckey and E. N. G. Marsh, *Proc. Natl. Acad. Sci. USA*, 2012, **109**, 4810–4815.

15 B. C. Buer, J. L. Meagher, J. A. Stuckey and E. N. G. Marsh, *Protein Sci.*, 2012, **21**, 1705–1715.

16 A. A. Berger, J.-S. Völler, N. Budisa and B. Koksch, *Acc. Chem. Res.*, 2017, **50**, 2093–2103.

17 F. Agostini, J.-S. Völler, B. Koksch, C. G. Acevedo-Rocha, V. Kubyshkin and N. Budisa, *Angew. Chem. Int. Ed.*, 2017, **56**, 9680–9703.

18 N. C. Yoder and K. Kumar, *Chem. Soc. Rev.*, 2002, **31**, 335–341.

19 M. Salwiczek, E. K. Nyakatura, U. I. Gerling, S. Ye and B. Koksch, *Chem. Soc. Rev.*, 2012, **41**, 2135–2171.

20 M. Oliver, C. Gadais, J. García-Pindado, M. Teixidó, N. Lensen, G. Chaume and T. Brigaud, *RSC Adv.*, 2018, **8**, 14597–14602.

21 V. Asante, J. Mortier, H. Schlüter and B. Koksch, *Bioorg. Med. Chem.*, 2013, **21**, 3542–3546.

22 H. Meng and K. Kumar, *J. Am. Chem. Soc.*, 2007, **129**, 15615–15622.

23 C. Jäckel, M. Salwiczek and B. Koksch, *Angew. Chem. Int. Ed.*, 2006, **45**, 4198–4203.

24 T. Liang, C. N. Neumann and T. Ritter, *Angew. Chem. Int. Ed.*, 2013, **52**, 8214–8264.

25 E. P. Gillis, K. J. Eastman, M. D. Hill, D. J. Donnelly and N. A. Meanwell, *J. Med. Chem.*, 2015, **58**, 8315–8359.

26 G. Pupo, F. Ibba, D. M. H. Ascough, A. C. Vicini, P. Ricci, K. E. Christensen, L. Pfeifer, J. R. Morphy, J. M. Brown, R. S. Paton and V. Gouverneur, *Science*, 2018, **360**, 638–642.

27 J.-S. Völler, M. Dulic, U. I. Gerling-Driessen, H. Biava, T. Baumann, N. Budisa, I. Gruic-Sovulj and B. Koksch, *ACS Cent. Sci.*, 2016, **3**, 73–80.

28 M. Imiołek, G. Karunanithy, W.-L. Ng, A. J. Baldwin, V. E. Gouverneur and B. G. Davis, *J. Am. Chem. Soc.*, 2018.

29 R. L. Baldwin, *Proc. Natl. Acad. Sci. USA*, 2014, **111**, 13052–13056.

30 D. Chandler, *Nature*, 2005, **437**, 640–647.

31 H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, *J. Chem. Phys.*, 2004, **120**, 9665–9678.

32 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.

33 J. R. Robalo, S. Huhmann, B. Koksch and A. Vila Verde, *Chem*, 2017, **3**, 881 – 897.

34 W. L. Jorgensen and P. Schyman, *J. Chem. Theory Comput.*, 2012, **8**, 3895–3901.

35 P. Politzer, J. S. Murray and T. Clark, *Phys. Chem. Chem. Phys.*, 2010, **12**, 7748–7757.

36 P. Metrangolo, J. S. Murray, T. Pilati, P. Politzer, G. Resnati and G. Terraneo, *Cryst. Growth Des.*, 2011, **11**, 4238–4246.

37 K. El Hage, T. Bereau, S. Jakobsen and M. Meuwly, *J. Chem. Theory Comput.*, 2016, **12**, 3008–3019.

38 M. A. A. Ibrahim, *J. Phys. Chem. B*, 2012, **116**, 3659–3669.

39 M. Carter, A. K. Rappé and P. S. Ho, *J. Chem. Theory Comput.*, 2012, **8**, 2461–2473.

40 T. Bereau, C. Kramer and M. Meuwly, *J. Chem. Theory Comput.*, 2013, **9**, 5450–5459.

41 F. Hédin, K. El Hage and M. Meuwly, *J. Chem. Inf. Model*, 2016, **56**, 1479–1489.

42 H. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43 – 56.

43 E. Lindahl, B. Hess and D. van der Spoel, *J. Mol. Model.*, 2001, **7**, 306–317.

44 D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.

45 B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.

46 S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics*, 2013, **29**, 845–854.

47 S. Páll, M. J. Abraham, C. Kutzner, B. Hess and E. Lindahl, International Conference on Exascale Applications and Software, 2014, pp. 3–27.

48 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**, 19–25.

49 D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham III, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu and P. A. Kollman, *AMBER 14*, 2014.

50 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

51 J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.

52 M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, 1989.

53 H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola and J. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

54 C. H. Bennett, *J. Comput. Phys.*, 1976, **22**, 245–268.

55 M. R. Shirts, E. Bair, G. Hooker and V. S. Pande, *Phys. Rev. Lett.*, 2003, **91**, 140601.

56 N. A. Baker, D. Sept, S. Joseph, M. J. Holst and J. A. McCammon, *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 10037–10041.

57 C. Gadais, E. Devillers, V. Gasparik, E. Chelain, J. Pytkowicz and T. Brigaud, *ChemBioChem*, 2018, **19**, 1026–1030.

58 A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer and D. G. Truhlar, *Minnesota Solvation Database – version 2012*, 2012.

59 D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan and W. Wriggers, *Science*, 2010, **330**, 341–346.

60 K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, 2010, **78**, 1950–1958.

61 K. Lindorff-Larsen, N. Trbovic, P. Maragakis, S. Piana and D. E.

Shaw, *J. Am. Chem. Soc.*, 2012, **134**, 3787–3791.
62 K. A. Beauchamp, Y.-S. Lin, R. Das and V. S. Pande, *J. Chem. Theory Comput.*, 2012, **8**, 1409–1414.
63 S. Piana, J. L. Klepeis and D. E. Shaw, *Curr. Opin. Struct. Biol.*, 2014, **24**, 98–105.
64 Q. A. Huchet, B. Kuhn, B. Wagner, H. Fischer, M. Kansy, D. Zimmerli, E. M. Carreira and K. Müller, *J. Fluorine Chem.*, 2013, **152**, 119–128.
65 Q. A. Huchet, B. Kuhn, B. Wagner, N. A. Kratochwil, H. Fischer, M. Kansy, D. Zimmerli, E. M. Carreira and K. Müller, *J. Med. Chem.*, 2015, **58**, 9041–9060.
66 C. Tan, Y.-H. Tan and R. Luo, *J. Phys. Chem. B*, 2007, **111**, 12263–12274.
67 V. H. Dalvi and P. J. Rossky, *Proc. Nat. Acad. Sci. USA*, 2010, **107**, 13603–13607.
68 C. A. Hunter, *Chem. Sci.*, 2013, **4**, 1687–1700.
69 S. Mondal, B. Biswas, T. Nandy and P. C. Singh, *Phys. Chem. Chem. Phys.*, 2017, **19**, 24667–24677.
70 W. Caminati, S. Melandri, I. Rossi and P. G. Favero, *J. Am. Chem. Soc.*, 1999, **121**, 10098–10101.
71 H.-J. Schneider, *Chem. Sci.*, 2012, **3**, 1381–1394.
72 J. Thomas, N. A. Seifert, W. Jäger and Y. Xu, *Angew. Chem. Int. Ed.*, 2017, **56**, 6289–6293.
73 C. Dalvit and A. Vulpetti, *ChemMedChem*, 2011, **6**, 104–114.
74 J. A. Olsen, D. W. Banner, P. Seiler, B. Wagner, T. Tschopp, U. Obst-Sander, M. Kansy, K. Müller and F. Diederich, *ChemBioChem*, 2004, **5**, 666–675.
75 D. O'Hagan, *Chem. Soc. Rev.*, 2008, **37**, 308–319.
76 Q. A. Huchet, W. B. Schweizer, B. Kuhn, E. M. Carreira and K. Müller, *Chem. Eur. J.*, 2016, **22**, 16920–16928.
77 Q. A. Huchet, N. Trapp, B. Kuhn, B. Wagner, H. Fischer, N. A. Kratochwil, E. M. Carreira and K. Müller, *J. Fluorine Chem.*, 2017, **198**, 34–46.
78 R. E. Rosenberg, *J. Phys. Chem. A*, 2012, **116**, 10842–10849.
79 B. Lin and B. M. Pettitt, *J. Chem. Phys.*, 2011, **134**, 106101–.
80 B. Lin, K.-Y. Wong, C. Hu, H. Kokubo and B. M. Pettitt, *J. Phys. Chem. Lett.*, 2011, **2**, 1626–1632.
81 S.-C. Ou and B. M. Pettitt, *J. Phys. Chem. B*, 2016, **120**, 8230–8237.
82 T. P. Creamer, R. Srinivasan and G. D. Rose, *Biochemistry*, 1995, **34**, 16245–16250.
83 N. C. Fitzkee and G. D. Rose, *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 12497–12502.
84 J. Estrada, P. Bernado, M. Blackledge and J. Sancho, *Bmc Bioinformatics*, 2009, **10**, 104.
85 Y. Seki, Y. Shimbo, T. Nonaka and K. Soda, *J. Chem. Theory Comput.*, 2011, **7**, 2126–2136.
86 S. A. Ali, M. I. Hassan, A. Islam and F. Ahmad, *Curr. Protein Pept. Sci.*, 2014, **15**, 456–476.

# SUPPLEMENTAL INFORMATION

João R. Robalo,[1] Ana Vila Verde[1*]

[1] Department of Theory & Bio-systems, Max Planck Institute for Colloids and Interfaces, Science Park, Potsdam 14424 Germany.

[*] Correspondence: ana.vilaverde@mpikg.mpg.de

## 1 FORCE FIELDS

In the TIP4P-Ew (ref. 1) water model, used in this work, the hydrogen has a charge $q_H = +0.52422\ |e|$, oxygen carries no charge, and a negative charge of magnitude $q_M = -1.04844\ |e|$ is located along the direction bisecting the region between the two hydrogens, 0.1250 Å away from the oxygen atom. Lennard-Jones (LJ) parameters are zero for hydrogen atoms; oxygen atoms have $\varepsilon_{OO} = 0.680946$ kJ mol$^{-1}$ and $\sigma_{OO} = 3.16435$ Å; see SI Equation 1 for the functional form of the LJ potential.

We use the AMBER14 (ref. 2) force field for the canonical amino acids, and a force field developed by us (previous own work[3] and present work) compatible with AMBER14 for the remaining ones. The force field for ethylglycine and for tri- and hexa-fluorinated amino acids, together with a detailed description of the procedure used to develop it, is given elsewhere[3]. Here and in SI section 2 we give only the relevant details for the parameterization of mono- and difluorinated amino acids. Bonded parameters for the fluorinated amino acids were taken from the parent non-fluorinated amino acid; non-bonded, LJ, parameters are from AMBER14 (ref. 2), save for $\varepsilon_{FF}$, $\sigma_{FF}$, $\varepsilon_{H_F}$ and $\sigma_{H_F}$ LJ parameters, where $F$ indicates a fluorine atom covalently bound to a carbon, and $H_F$ indicates a hydrogen atom covalently bound to a fluorinated carbon; see SI Equation 1 for the definition of these parameters. LJ parameters used for methyl and fluoromethyl groups are presented in SI Table 1. For fluorine we use LJ parameters ($\varepsilon_{FF}$ and $\sigma_{FF}$) optimized against the free energy of hydration of $CF_4$ and the molar volume of a liquid 1:1 mixture of $CH_4$:$CF_4$, as described elsewhere[3]. LJ parameters for fluorocarbon-bound hydrogen, $H_F$, were optimized to describe the free energy of hydration of $CHF_3$ and the molar volume of liquid $CHF_3$, as described in SI Section 2.

| | $\varepsilon_{i,i}$ (kJ mol$^{-1}$) | $\sigma_{i,i}$ (Å) |
|---|---|---|
| C (aliphatic, GAFF[4]) | 0.4577300 | 3.3996700 |
| F (ref. 3) | 0.0908200 | 2.8000000 |
| H (aliphatic, GAFF) | 0.0656888 | 2.6495300 |
| $H_F$ (optimized, this work) | 0.1746000 | 2.0000000 |

**Table 1** LJ parameters $\varepsilon$ and $\sigma$ for the carbon, fluorine and hydrogen atoms in methane/methyl/methylene and fluorinated methane/methyl/methylene groups used in this work. All final results with partially fluorinated alkyl groups use the $H_F$ parameters optimized in this work.

Non-bonded Van der Waals interactions between any two particles, $i$ and $j$, over an inter-particle distance $r_{i,j}$ are calculated via the LJ potential as

$$V_{i,j}^{LJ} = 4\varepsilon_{i,j}\left(\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{6}\right) \tag{1}$$

with

$$\varepsilon_{ij} = \sqrt{\varepsilon_{ii}\varepsilon_{jj}} \text{ and } \sigma_{ij} = \frac{\sigma_{ii}+\sigma_{jj}}{2} \tag{2}$$

following the Lorentz-Berthelot combination rules[5–11].

Atomic partial charges for the fluorinated methane derivatives were obtained with the Merz-Singh-Kollman Restrained Electrostatic Potential (RESP) methodology on a single conformation for each molecule, following the standard GAFF protocol. The charges were calculated at the RHF/6-31G* level of theory from a two-step geometry optimization (MP2/6-31G* optimization followed by RHF/6-31G* optimization). Quantum mechanical calculations were performed with the Gaussian 03 software[12], and RESP fitting of the partial charges was performed with the Antechamber package[13].

The atomic partial charges for all non-canonical amino acids are obtained via a multi-configuration RESP fitting procedure, described in greater detail in ref. 3, performed over 200 conformations (100 $\alpha$-helical and 100 $\beta$-strand) per amino acid.

## 2 Simulation Details

The results discussed in the main text are obtained from Free Energy Perturbation (FEP) simulations – to calculate the hydration free energies of the amino acids – and molecular dynamics (MD) simulations – all other observables, *e.g.* solvent accessible surface area, dipole moment, numbers of hydrogen bonds – of single copies of the amino acids in water. In addition, FEP simulations of a single $CHF_3$ in water were used to calculate hydration free energies in the parameterization of fluorocarbon-bound hydrogen ($H_F$). Also used in the parameterization were simulations of liquid $CHF_3$, to calculate its molar volume, and additionally simulations of gas phase $CHF_3$, to calculate its vaporization enthalpy. Obtaining the partial charges for the partially fluorinated amino acids required initial simulations of single copies of the amino acids (with non-optimized charges) in water, to obtain multiple configurations that were then used in the multi-configuration RESP fit, as described in ref. 3. The simulation boxes used for each type of simulation are illustrated in SI Figure 1; periodic boundary conditions in all directions were used in all cases. All systems were assembled using the built-in tools of the software package used to perform the simulations. A summary of the most relevant parameters used during the production runs is given in SI Table 2. Simulations used a time-step of 2 fs and constraints (LINCS[15] in Gromacs, SHAKE[16] in Amber) were applied to all bonds involving hydrogen atoms. Integration of the equations of motion was done using a leap-frog Langevin algorithm. Van der Waals interactions were shifted to zero between 1.0 and 1.2 nm, and long-range dispersion corrections were applied to both pressure and energy.

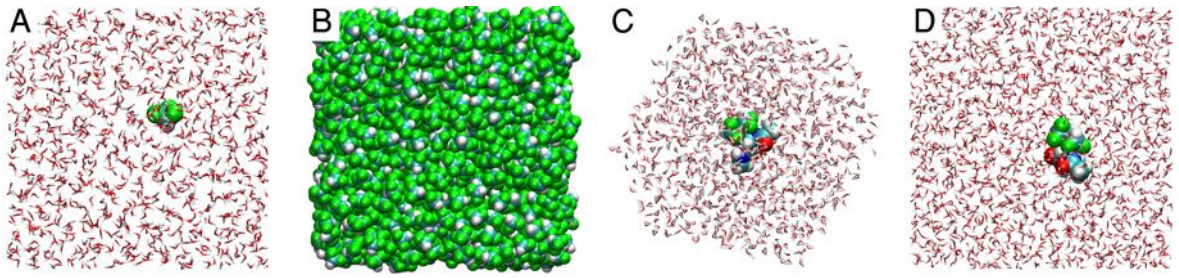



**Fig. 1** Liquid phase simulation boxes of A) one $CHF_3$ molecule in water; B) pure $CHF_3$; C,D) L4D, capped with ACE and NME, in an octahedral (C) or cubic (D) box. Fluorinated molecules are shown as van der Waals surfaces, water as narrow tubes.

| | $CH_3F/CH_2F_2/CHF_3$ ($aq$) | $CHF_3$ ($l$) | AA ($aq$) | AA ($aq$) | AA ($aq$) |
|---|---|---|---|---|---|
| Obs | $\Delta G_{Hyd}$ | $\Delta H_{Vap}$, $V_{Mol}$ | RESP fit | $\Delta G_{Hyd}$ | Analysis |
| Box | cubic (A) | cubic (B) | trunc. oct (C) | cubic (D) | trunc. oct. (C) |
| # mol | 1070 ($H_2O$) | 1150 ($CHF_3$) | 1000 ($H_2O$) | 1750 ($H_2O$) | 1000 ($H_2O$) |
| p (bar) | 1 | 1 | 1 | 1 | 1 |
| T (K) | 298 | 173 | 298 | 298 | 298 |
| t (ns) | 1* | 10 | 50 | 2* | 25 |
| # rep | 3 | 3 | 1 | 5 | 1 |
| Package | Gromacs 5.0[5–11] | Gromacs 5.0 | AMBER 14[14] | Gromacs 5.0 | AMBER 14 |

* per $\lambda$ value

**Table 2** Simulation details for the production runs of the systems: $C_nH_m$ ($aq$) = one $C_nH_m$ molecule in water; $CHF_3$ ($l$) = pure $CHF_3$; AA ($aq$) = one capped amino acid in water. Obs = calculated observable; Box = box type (corresponding panel in Figure 1); # mol = number of molecules; p = pressure; T = temperature; t = production run length; # rep = number of independent runs (with different velocity assignments in the NVT equilibration step); Package = software package used in the simulations.

For the parameterization of the Lennard-Jones $\sigma_{H_F}$ and $\varepsilon_{H_F}$, each FEP and liquid phase simulation (molar volume and enthalpy of vaporization) consisted of a steepest-descent minimization, BFGS minimization, 100 ps of NVT equilibration, 100 ps of NpT equilibration and production run; gas phase simulations (enthalpy of vaporization) consisted only of the production run. FEP simulation of the amino acids also used the same equilibration procedure. Production runs were done in the NpT ensemble (liquid simulations) or the NVT ensemble (for gas phase). Long-range electrostatics were treated with the PME[17] scheme with a 1.2 nm cutoff, a grid spacing of 0.12 nm and a 6th order interpolation. The Berendsen barostat[18] was used with a time constant of 1 ps for coupling pressure to 1 bar; temperature coupling was enforced in the leap-frog Langevin integrator with a time constant of 1 ps.

Simulations for the second iteration of the RESP fit (extracting multiple configurations of the amino acids) consisted of a steepest descent minimization, a conjugated gradient minimization, a 200 ps NVT heating from 0 K to 298 K, and a 1 ns NpT equilibration; in the minimization, heating and equilibration steps the coordinates of the backbone atoms were restrained with a 20 kcal $mol^{-1}$ potential. The production step was run in the NpT ensemble, keeping the same restraints on the backbone atoms. Long-range electrostatics were treated with the PME scheme with a 1.2 nm cutoff, a grid spacing of 0.1 nm and a 4th order interpolation. The Monte Carlo barostat[14,19] was used with a relaxation time of 1 ps for an isotropic coupling of system pressure to 1 bar; temperature coupling was handled by the leap-frog Langevin integrator with a collision frequency of 1 $ps^{-1}$. Simulations for the calculation of the input parameters in Equation 1 in the main text followed the same protocol except that no

restraints were imposed on the backbone atoms; these trajectories were used as input for the APBS[20] software to estimate the electrostatic hydration free energy of the amino acids using the linearized Poisson-Boltzmann equation, as described in more detail below.

### 2.1 Free Energy Perturbation

The potential form of the scaled intermolecular non-bonded interactions follows SI Equation 3, with the soft-core parameter $\alpha_{LJ}$ set to 0.5 for any $\lambda \neq \{0,1\}$ and zero otherwise.

$$V_{NB}(\lambda_C,\lambda_{LJ}) = \sum_i \sum_j (1-\lambda_C)\frac{q_i q_j}{r_{ij}} + \frac{(1-\lambda_{LJ})4\varepsilon_{ij}}{\left[\alpha_{LJ}\lambda_{LJ} + (r_{ij}/\sigma_{ij})^6\right]^2} - \frac{(1-\lambda_{LJ})4\varepsilon_{ij}}{\alpha_{LJ}\lambda_{LJ} + (r_{ij}/\sigma_{ij})^6} \tag{3}$$

Hydration free energies were calculated using Free Energy Perturbation (FEP) and Bennett Acceptance Ratio[21,22] (BAR), following the protocol we have previously adopted[3]. In each case, simulations consisting of a single solute molecule in water were conducted, first decoupling Coulombic interactions and then LJ interactions. For the parameterization of hydrogen LJ coefficients, the coupling parameter $\lambda_C$ for the Coulombic interactions adopted the value of 0.00 (fully coupled), 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95 and 1.00 (fully decoupled); the resulting Coulombic potential is scaled linearly with $\lambda$. $\lambda_{LJ}$ values were 0.00 (fully coupled), 0.06, 0.12, 0.18, 0.24, 0.30, 0.36, 0.42, 0.46, 0.50, 0.52, 0.54, 0.56, 0.58, 0.60, 0.64, 0.68, 0.72, 0.76, 0.80, and 1.00 (fully decoupled), for a total of 42 simulations; the hydration free energies for methane and fluorinated methane derivatives presented in SI Table 4 were calculated using 41 $\lambda_{LJ}$ states: 0.00 (fully coupled), 0.03, 0.06, 0.09, 0.12, 0.15, 0.18, 0.21, 0.24, 0.27, 0.30, 0.33, 0.36, 0.39, 0.42, 0.44, 0.46, 0.48, 0.50, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59, 0.60, 0.62, 0.64, 0.66, 0.68, 0.70, 0.72, 0.74, 0.76, 0.78, 0.80, 0.90 and 1.00 (fully decoupled), adding to a total of 62 simulations. For the calculation of amino acid hydration free energies, the $\lambda_C$ were kept from the methane simulations. $\lambda_{LJ}$ adopted the values 0.00 (fully coupled), 0.03, 0.06, 0.09, 0.12, 0.15, 0.18, 0.21, 0.24, 0.27, 0.30, 0.33, 0.36, 0.39, 0.42, 0.44, 0.46, 0.48, 0.50, 0.51, 0.52, 0.53, 0.54, 0.55, 0.56, 0.57, 0.58, 0.59, 0.60, 0.61, 0.62, 0.63, 0.64, 0.65, 0.66, 0.67, 0.68, 0.69, 0.70, 0.71, 0.72, 0.73, 0.74, 0.75, 0.76, 0.77, 0.78, 0.79, 0.80, 0.82, 0.84, 0.86, 0.88, 0.90, 0.92, 0.94, 0.96, 0.98 and 1.00 (fully decoupled), for a total of 80 Coulomb- and LJ-decoupling steps. An increase in both the number of states and the simulation time (compared to the simulations for methane derivatives, refer to SI Table 2) was necessary due to poor convergence for $\lambda_{LJ}$ values above 0.6. Convergence was assumed when the entropy difference between adjacent states, calculated by analyzing the trajectory at each $\lambda$ state with the adjacent states' Hamiltonians, did not exceed 0.2 kcal mol$^{-1}$.

### 2.2 Enthalpy of Vaporization and Molar Volume

Following the protocol of Gough and co-workers[23], the enthalpy of vaporization per mole, $\Delta H$, of pure $CHF_3$ was calculated as

$$\Delta H = \Delta E + \Delta(pV) \approx \Delta E + RT \tag{4}$$

assuming that the gas phase systems behave ideally. In this expression, $\Delta E = E_g - E_l$, $p$ is the pressure, $V$ is the volume and $R$ is the ideal gas constant. $E_l$ is the potential energy per mole calculated from a simulation of a cubic box of pure $CHF_3$ in the liquid phase, and $E_g$ is the equivalent quantity calculated using a simulation box containing only one molecule of $CHF_3$ (gas phase) and otherwise the same simulation parameters. For each system, the molar volume was calculated by dividing the average volume of the liquid phase simulation box by the total number of molecules, then multiplying by Avogadro's constant.

### 2.3 Parameterizing the $H_F$ Atom

Given the known impact of the highly electronegative fluorine element in adjacent atoms, parameters for a fluorocarbon-bound hydrogen atom are required. To this end, we calculated the hydration free energy and molar volume of the $CHF_3$ molecule for all combinations of $\varepsilon$ equal to (0.0656888*,0.15, 0.20, 0.25,0.30) kJ mol$^{-1}$ and $\sigma$ equal to (1.0, 1.5, 2.0, 2.64953†, 3.0) Å, totaling 25 ($\varepsilon$, $\sigma$) pairs. We then fit each of the calculated data sets and intersected the resulting surfaces with the experimental value of the hydration free energy or molar volume of $CHF_3$, given in SI Table 3. The intersection of both data sets yields the optimized parameters given in SI Table 1.

| | $\Delta G_{Hyd}$ kcal mol$^{-1}$ | $\Delta H_{Vap}$ kcal mol$^{-1}$ | $V_{Mol}$ cm$^3$ mol$^{-1}$ |
|---|---|---|---|
| $CH_3F$ | -0.952 ± 0.026‡ (-0.22 Ref. 24,25) | - | - |
| $CH_2F_2$ | -0.461 ± 0.001♯ ( - ) | - | - |
| $CHF_3$ | 1.012 ± 0.029** (0.659 Ref. 24,25) | 3.470 ± 0.010 (4.25 Ref. 23) | 44.453 ± 0.188 (46.1 Ref. 23) |

‡ $\Delta G_{Hyd}$ calculated in TIP3P water of $CH_3F$: -1.023 kcal mol$^{-1}$;
♯ $\Delta G_{Hyd}$ calculated in TIP3P water of $CH_2F_2$: -0.747 kcal mol$^{-1}$;
** $\Delta G_{Hyd}$ calculated in TIP3P water of $CHF_3$: 0.687 kcal mol$^{-1}$;

**Table 3** Calculated hydration free energy ($\Delta G_{Hyd}$), enthalpy of vaporization ($\Delta H_{Vap}$) and molar volume ($V_{Mol}$) for $CHF_3$ and hydration free energy for $CH_2F_2$ and $CH_3F$, using our optimized parameters; target values are shown within parenthesis. For the hydration free energy, solubilities are converted to the Ostwald coefficient[24], from which the standard free energy of solvation can be calculated[25]; for the vaporization enthalpy and molar volume of $CHF_3$, experimental values are taken from Gough and colleagues[23]. Results are shown as mean ± standard deviation of five independent simulations.

It would also be of interest to have an estimate of the optimal parameters for hydrogen in mono- and di-fluorinated carbons, which would be more relevant considering the fluorination patterns in the amino acids we are studying. Unfortunately, relevant experimental data on these species is sparse: for $CHF_3$, used in our parameterization, there are experimentally obtained values of its hydration free energy, molar volume and enthalpy of vaporization; for $CH_3F$ there is only the hydration free energy,

* Original value in the GAFF[4] force field for carbon-bound hydrogen.
† Original value in the GAFF[4] force field for carbon-bound hydrogen.

for $CH_2F_2$ there is no data. Hydration free energies, enthalpies of vaporization and molar volumes, calculated with our parameters for $CHF_3$, $CH_2F_2$ and $CH_3F$, are given in SI Table 3. Our parameters follow the expected trend of increasingly positive hydration free energy with increasing degree of fluorination, though the hydration free energy for $CH_3F$ is off by 0.7 kcal $mol^{-1}$. Because of the lack of experimental data, we cannot comment on the quality of our parameters for $CH_2F_2$.

It is worthwhile noting that the polar hydrogen atom we have parameterized is a main contributor to the more hydrophilic character of partially fluorinated methane relative to either $CF_4$ or $CH_4$, as becomes obvious by inspecting SI Tables 3 and 4. This effect of the polar hydrogen partially explains why the proportionality between surface area and hydration free energy no longer stands for partially fluorinated methane derivatives: each of the methane derivatives in SI Table 3 has a higher surface area than $CH_4$ while having a much lower hydration free energy (the hydration free energy of methane is 2.549 kcal $mol^{-1}$; see SI Table 4). In addition, the non-uniform charge distribution in the partially fluorinated molecules also increases the hydrophilicity of these molecules relative to $CH_4$ or $CF_4$.

## 3 Hydration free energies of fluorinated and non-fluorinated amino acids

The hydration free energy and differences in hydration free energy between fluorinated and non-fluorinated amino acids are presented in SI Table 4; also presented are the Coulombic and LJ components of the hydration free energy. Values for ethylglycine (ETG), valine (VAL), isoleucine (ILE), leucine (LEU) and all fully fluorinated amino acids (E3G, V3S, V3R, V6G, I3D, I3G, L3S, L3R, L6D) were retrieved from ref. 3 and are repeated here for the convenience of the reader.

### 3.1 Estimating electrostatic hydration free energies using Poisson-Boltzmann theory

We tested whether it was possible to estimate the electrostatic hydration free energy of the amino acids by solving the linearized Poisson-Boltzmann (PB) equation via a multigrid, finite difference, method[26,27], as implemented in the APBS[20] software. For the aqueous phase, default options for the APBS package were used, namely a multigrid level of four, with 129 grid points *per* processor in each direction (x, y, z; 0.1 Å grid spacing) and a multiple Debye-Hückel boundary condition. The solute dielectric constant was set to one, the solvent dielectric constant to 63 (corresponding to the TIP4P-Ew water model at 298 K[1]), with a cubic B-spline discretization of the charge mapping. The solvent-excluded volume was defined from the molecular surface, with atomic radii adopted from ref. 28; the solvent probe radius was taken as 1.4 Å (corresponding to water). The temperature was set at 298.15 K. Calculations in vacuum were performed by setting the solvent dielectric constant to one and the hydration free energy was calculated as the difference in free energy from the gas state to the solvated state. For each solvation free energy calculation, 25 configurations were extracted from a 25 ns simulation of the solvated amino acid. SI Table 5 holds the results of the PB calculations. SI Figure 2 A makes the comparison between PB and FEP hydration free energies. This figure clearly demonstrates that the PB approach does not capture the relative differences in the electrostatic hydration free energy

| | $\Delta G_{Hyd}$ | $\Delta G_{Hyd,Coul}$ | $\Delta G_{Hyd,LJ}$ | $\Delta\Delta G_{Hyd}$ | $\Delta\Delta G_{Hyd,Coul}$ | $\Delta\Delta G_{Hyd,LJ}$ |
|---|---|---|---|---|---|---|
| $CH_4$ | 2.549 ± 0.022 | 0.003 ± 0.001 | 2.546 ± 0.022 | | | |
| $CH_3F$ | -0.952 ± 0.026 | -2.862 ± 0.009 | 1.910 ± 0.022 | -3.501 ± 0.049 | -2.866 ± 0.011 | -0.636 ± 0.044 |
| $CH_2F_2$ | -0.461 ± 0.001 | -2.956 ± 0.012 | 2.495 ± 0.017 | -3.010 ± 0.024 | -2.959 ± 0.013 | -0.051 ± 0.039 |
| $CHF_3$ | 1.012 ± 0.029 | -2.069 ± 0.007 | 3.082 ± 0.019 | -1.537 ± 0.052 | -2.073 ± 0.008 | 0.536 ± 0.041 |
| $CF_4$ | 3.258 ± 0.033 | -0.379 ± 0.003 | 3.637 ± 0.028 | 0.709 ± 0.055 | -0.382 ± 0.004 | 1.091 ± 0.050 |
| $CH_3CH_3$ | 2.621 ± 0.021 | 0.0024 ± 0.000 | 2.618 ± 0.021 | | | |
| $CHF_2CH_3$ | 0.018 ± 0.010 | -2.823 ± 0.010 | 2.841 ± 0.015 | | | |
| $CF_3CF_3$ | 3.871 ± 0.037 | -0.386 ± 0.002 | 4.257 ± 0.038 | | | |
| $CH_3CH_2CH_3$ | 2.759 ± 0.047 | -0.002 ± 0.000 | 2.761 ± 0.047 | | | |
| $CF_3CF_2CF_3$ | 4.394 ± 0.055 | -0.378 ± 0.002 | 4.772 ± 0.055 | | | |
| ETG* | -14.125 ± 0.054 | -15.976 ± 0.037 | 1.851 ± 0.024 | | | |
| E1G | -13.722 ± 0.128 | -15.260 ± 0.106 | 1.538 ± 0.050 | 0.403 ± 0.182 | 0.716 ± 0.143 | -0.313 ± 0.075 |
| E2G | -13.121 ± 0.096 | -15.248 ± 0.114 | 2.127 ± 0.023 | 1.004 ± 0.150 | 0.728 ± 0.151 | 0.276 ± 0.047 |
| E3G* | -13.115 ± 0.060 | -15.781 ± 0.059 | 2.666 ± 0.022 | 1.010 ± 0.114 | 0.195 ± 0.096 | 0.815 ± 0.046 |
| PRG | -14.025 ± 0.067 | -16.098 ± 0.050 | 2.074 ± 0.030 | | | |
| P2G | -14.056 ± 0.159 | -16.665 ± 0.126 | 2.608 ± 0.066 | -0.032 ± 0.226 | -0.566 ± 0.176 | 0.535 ± 0.096 |
| VAL* | -13.871 ± 0.120 | -15.965 ± 0.122 | 2.094 ± 0.025 | | | |
| V3S* | -13.335 ± 0.129 | -16.244 ± 0.111 | 2.909 ± 0.057 | 0.536 ± 0.249 | -0.279 ± 0.233 | 0.815 ± 0.082 |
| V3R* | -12.375 ± 0.105 | -15.240 ± 0.153 | 2.865 ± 0.052 | 1.496 ± 0.226 | 0.725 ± 0.275 | 0.771 ± 0.078 |
| V6G* | -11.846 ± 0.082 | -15.381 ± 0.105 | 3.535 ± 0.030 | 2.025 ± 0.202 | 0.584 ± 0.227 | 1.441 ± 0.055 |
| ILE* | -13.700 ± 0.100 | -16.066 ± 0.101 | 2.366 ± 0.043 | | | |
| I1G | -13.823 ± 0.209 | -15.963 ± 0.204 | 2.140 ± 0.025 | -0.123 ± 0.309 | 0.103 ± 0.306 | -0.226 ± 0.068 |
| I3D* | -13.235 ± 0.056 | -16.448 ± 0.077 | 3.213 ± 0.048 | 0.465 ± 0.156 | -0.382 ± 0.178 | 0.847 ± 0.091 |
| I3G* | -13.202 ± 0.146 | -16.379 ± 0.094 | 3.177 ± 0.075 | 0.498 ± 0.246 | -0.313 ± 0.196 | 0.811 ± 0.118 |
| LEU* | -13.811 ± 0.070 | -16.109 ± 0.050 | 2.298 ± 0.030 | | | |
| L1S | -15.296 ± 0.046 | -17.335 ± 0.056 | 2.040 ± 0.017 | -1.485 ± 0.116 | -1.226 ± 0.106 | -0.259 ± 0.047 |
| L1R | -14.797 ± 0.121 | -16.838 ± 0.062 | 2.041 ± 0.068 | -0.987 ± 0.191 | -0.729 ± 0.112 | -0.258 ± 0.098 |
| L4D | -14.670 ± 0.087 | -17.580 ± 0.064 | 2.910 ± 0.045 | -0.859 ± 0.157 | -1.471 ± 0.114 | 0.612 ± 0.075 |
| L3S* | -13.568 ± 0.042 | -16.757 ± 0.063 | 3.190 ± 0.047 | 0.243 ± 0.112 | -0.648 ± 0.113 | 0.892 ± 0.077 |
| L3R* | -12.903 ± 0.147 | -16.002 ± 0.129 | 3.099 ± 0.033 | 0.908 ± 0.217 | 0.107 ± 0.180 | 0.801 ± 0.063 |
| L6D* | -12.373 ± 0.262 | -16.231 ± 0.175 | 3.858 ± 0.089 | 1.438 ± 0.332 | -0.122 ± 0.226 | 1.560 ± 0.119 |

**Table 4** Free energy of hydration ($\Delta G_{Hyd}$), Coulombic ($\Delta G_{Hyd,Coul}$) and Lennard-Jones ($\Delta G_{Hyd,LJ}$) components of the free energy of hydration for small molecules and amino acids; differences in the free energy of hydration ($\Delta\Delta G_{Hyd}$), Coulombic ($\Delta\Delta G_{Hyd,Coul}$) and Lennard-Jones components of the free energy of hydration ($\Delta\Delta G_{Hyd,LJ}$) between fluorinated and non-fluorinated molecules. All quantities were calculated using free energy perturbation and are shown as mean ± standard deviation, of five independent simulations, in kcal $mol^{-1}$. * retrieved from ref. 3. Related to Figure 2 in the main text.

between the amino acids, and does not give insight into these systems.

## 4 Calculating the electrostatic potential in the vicinity of the side chains

The electrostatic potential was calculated on a grid (0.1 Å of spacing) around the side chain of each amino acid, with conformations extracted from a 25 ns simulation of each amino acid in water, using only the partial charges of the side chain atoms. Then, the electrostatic potential was interpolated to the coordinates of oxygen atoms belonging to water molecules within a radius of 5.5 Å of the side chain carbon atoms; this radius corresponds to the first minimum of the radial distribution function of water oxygen atoms near methane's carbon atom, that is, it corresponds to the farther limit of the first hydration shell of methane. The probability density distributions of the electrostatic potential, $\Phi$, at these positions is shown in SI Figure 2 B-F. For each fluorinated amino acid, the $\Delta\Phi^{+}$ term in Equation 1 in the main text corresponds to the number of water molecules, within 5.5 Å of the side chain carbon atoms, that experience a potential which is more positive than that of the waters at the $90^{th}$ percentile of the $\Phi$ probability distribution of the corresponding non-fluorinated amino acid. SI Figure 3 shows the electrostatic potential $\Phi$ around the side chain of ETG, E1G, E2G and E3G. Compared to ETG, the negative potential on the fluorinated derivatives arises from the fluorine atoms, the positive potential from the exposed carbon skeleton, especially from the fluorinated carbon (the

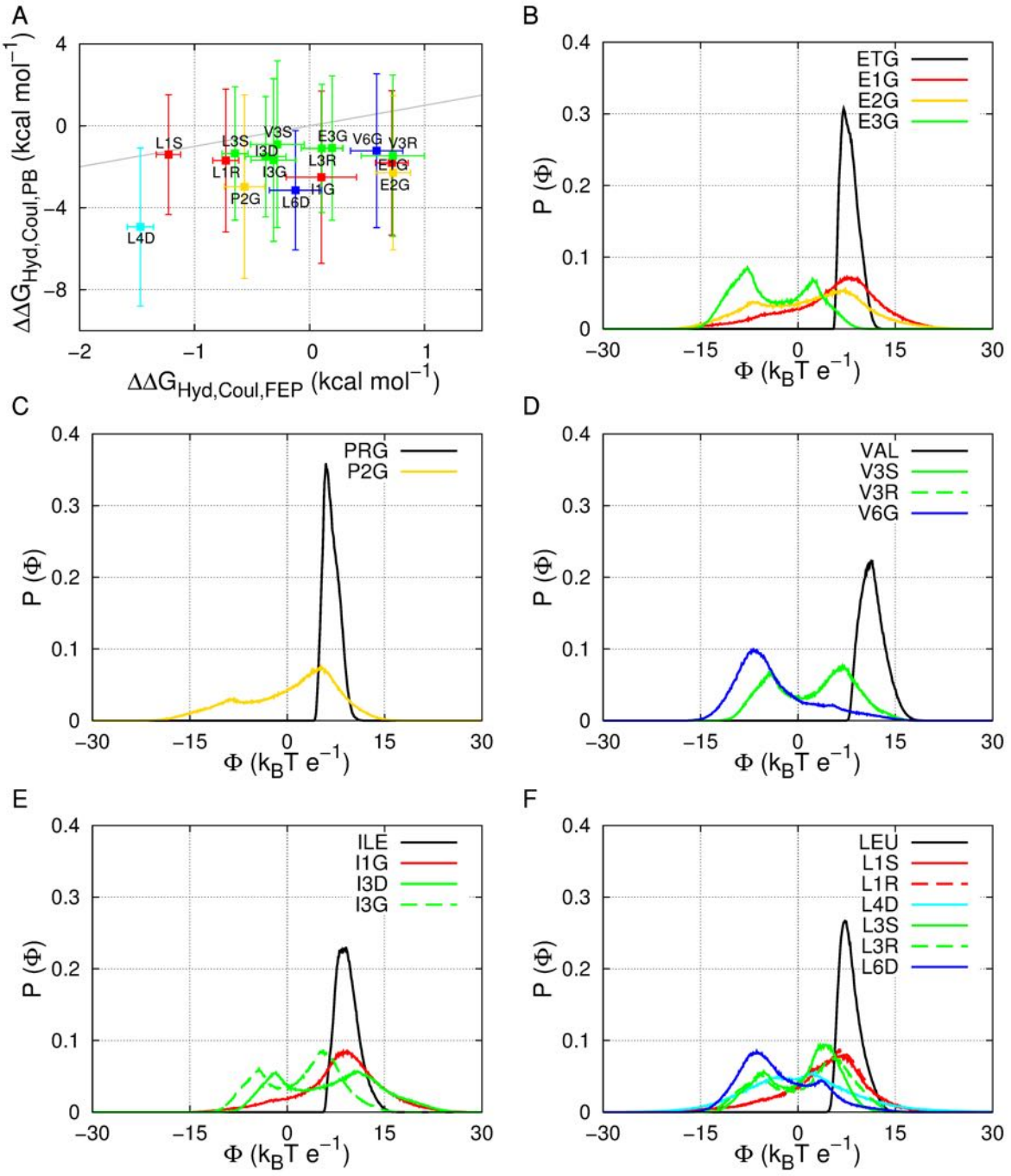



**Fig. 2** A) Free energy of hydration calculated by the linearized Poisson-Boltzmann equation ($\Delta\Delta G_{Hyd,PB}$) *versus* the free energy of hydration calculated by free energy perturbation method ($\Delta\Delta G_{Hyd,FEP}$); data shown as mean $\pm$ standard deviation of 25 independent calculations (PB) or five independent simulations (FEP); the gray line indicates perfect correlation; B-F) probability density associated with the electrostatic potential induced by the side chain atoms only, at the position of the oxygen atoms of water molecules in the hydration shell of the side chains of: B) ethylglycine and derivatives, C) propylglycine and derivatives, D) valine and derivatives, E) isoleucine and derivatives and F) leucine and derivatives. The color code indicates the number of fluorine atoms in the side chain: red = one; yellow = two; green = three; cyan = four; blue = six. The hydration shell is composed of all water molecules within 5.5 Å of the side chain carbon atoms.

| | $\Delta G_{Hyd,Coul,PB}$ | $\Delta\Delta G_{Hyd,Coul,PB}$ |
|---|---|---|
| ETG | -18.281 ± 1.497 | |
| E1G | -20.086 ± 2.027 | -1.805 ± 3.525 |
| E2G | -20.572 ± 2.270 | -2.290 ± 3.767 |
| E3G | -19.365 ± 2.030 | -1.083 ± 3.528 |
| PRG | -17.612 ± 1.865 | |
| P2G | -20.583 ± 2.608 | -2.971 ± 4.474 |
| VAL | -17.568 ± 1.617 | |
| V3S | -18.466 ± 2.448 | -0.898 ± 4.066 |
| V3R | -19.031 ± 2.323 | -1.463 ± 3.941 |
| V6G | -18.782 ± 2.139 | -1.214 ± 3.756 |
| ILE | -17.277 ± 1.664 | |
| I1G | -19.785 ± 2.543 | -2.508 ± 4.207 |
| I3D | -18.777 ± 1.272 | -1.500 ± 2.935 |
| I3G | -18.945 ± 2.305 | -1.668 ± 3.968 |
| LEU | -18.241 ± 1.444 | |
| L1S | -19.643 ± 1.483 | -1.403 ± 2.927 |
| L1R | -19.933 ± 2.050 | -1.692 ± 3.494 |
| L4D | -23.175 ± 2.411 | -4.934 ± 3.855 |
| L3S | -19.586 ± 1.815 | -1.345 ± 3.259 |
| L3R | -19.343 ± 1.689 | -1.102 ± 3.134 |
| L6D | -21.385 ± 1.467 | -3.144 ± 2.911 |

**Table 5** Coulombic component of the free energy of hydration calculated using the linearized form of the Poisson-Boltzmann equation ($\Delta G_{Hyd,Coul,PB}$) and differences in the Coulombic component of the free energy of hydration calculated using the linearized form of the Poisson-Boltzmann equation ($\Delta\Delta G_{Hyd,Coul,PB}$). All quantities are shown as mean ± standard deviation in kcal mol$^{-1}$. Calculations were performed over amino acid conformations obtained from 25 equally spaced frames extracted from a 25 ns simulation.

hydrogen atoms in $-CH_3$ and, *e.g.*, $-CH_2F$ have a charge of similar magnitude; the carbon atoms have charges of opposite sign, negative in $-CH_3$ and positive in $-CH_2F$). The number of water molecules in the hydration shell was calculated with the Amber software package[14], the electrostatic potential was calculated using the VMD software package[29].

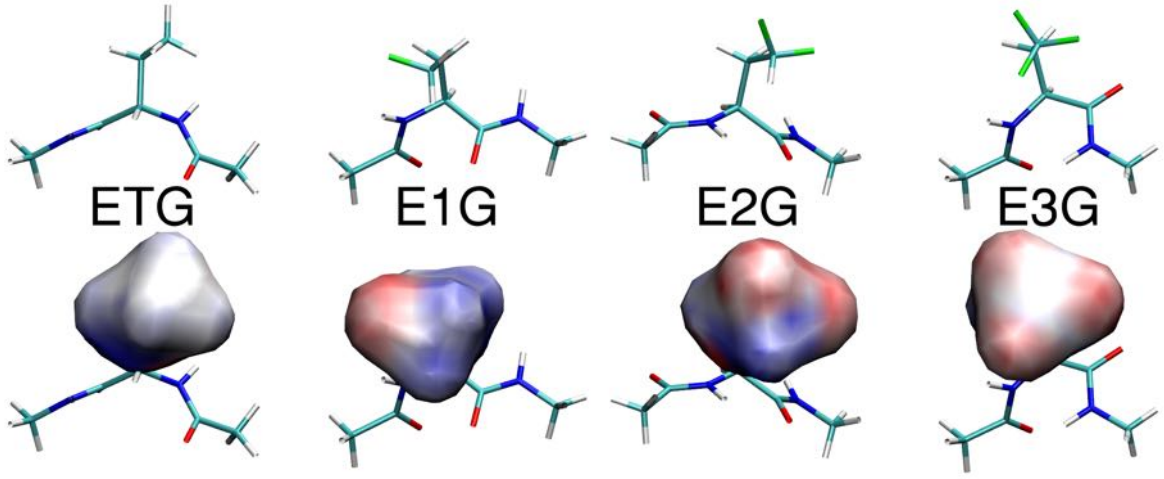



**Fig. 3** Molecular structure (upper row) and electrostatic potential (lower row) for a single conformation of the side chain of ETG, E1G, E2G and E3G. The surfaces where the electrostatic potential is mapped were calculated using a probe of radius 5.5 Å. The color code for the electrostatic potential map is the same for all molecules, scaling from red (negative) to blue (positive).

## 5 Modeling the hydration free energy of fluorinated amino acids

### 5.1 Two linear, multivariate models that fail to explain how fluorination alters hydration free energies of amino acids

We attempted to fit the $\Delta\Delta G_{Hyd}$ for all amino acids using SI Equation 5:

$$\Delta\Delta G_{Hyd} = k_1\Delta A + k_2\Delta h_{CO} + k_3\Delta h_{NH} + k_4\Delta\mu \tag{5}$$

In this equation, the coefficient $k_1$ of the ΔA term is the LJ component of the hydration free energy *per* surface area unit of methane and the coefficients $k_{2-4}$ for the water–carbonyl hydrogen bond ($\Delta h_{CO}$), the water–amine hydrogen bond ($\Delta h_{NH}$) and the dipole moment ($\Delta\mu$) terms are freely optimized in the fitting procedure. SI Table 6 holds the values of the fitting parameters, errors and P-values resulting from this attempt. As described in the main text, this model proved unsuccessful: the contributions of the water-backbone hydrogen bonds are unphysical, with hydrogen bonds with amines being much stronger than with carbonyls, and increases in dipole moment increasing amino acid hydrophobicity.

| Fitting Parameter | Coefficient | Error | P-value |
|---|---|---|---|
| $k_1$ (ΔA; kcal mol$^{-1}$ Å$^{-2}$)) | 0.053 | 0.001* | NA |
| $k_2$ ($\Delta h_{CO}$; kcal mol$^{-1}$ H-Bond$^{-1}$) | -1.121 | 1.484 | 0.463 |
| $k_3$ ($\Delta h_{NH}$; kcal mol$^{-1}$ H-Bond$^{-1}$) | -5.732 | 3.863 | 0.161 |
| $k_4$ ($\Delta\mu$; kcal mol$^{-1}$ Debye$^{-1}$) | 0.173 | 0.461 | 0.714 |

**Table 6** Values of the fitting parameters from SI Equation 5, and associated standard errors and P-values. NA: not applicable; * calculated via error propagation.

In a second attempt to understand our data, we turn to a model (SI Equation 6) we have previously successfully fitted to the changes in hydration free energy associated with tri- and hexa-fluorinated amino acids only[3]. This model has similar terms to those in SI Equation 5; the main difference is that the coefficient of $\Delta A$ is estimated from the *difference* in areas and in hydration free energies between $CF_4$ and $CH_4$.

$$\Delta\Delta G_{Hyd} = 0.0304\Delta A - 3.59\Delta h_{CO} - 3.02\Delta h_{NH} - 0.188\Delta\mu \tag{6}$$

SI Figure 4 illustrates the correlation between the model predictions and the hydration free energies calculated using FEP for all the amino acids under consideration here. The model largely fails for the mono- and difluorinated amino acids.

### 5.2 A successful linear, multivariate model to understand how fluorination alters hydration free energies of amino acids

This model, described by Equation 1 in the main text, includes a surface area contribution, captured by the $k_1\Delta A$ term, which is identical to that in SI Equation 5. This contribution increases linearly with the number of fluorine atoms,as shown in SI Figure 5. The model also includes two terms that reflect the contributions of carbonyl-water and amine-water hydrogen bonds. The main difference between this model and SI Equations 5 and 6 is that it includes three separate terms for water–fluorine hydrogen bonds because their strength depends on the number of fluorine atoms in each group. Our

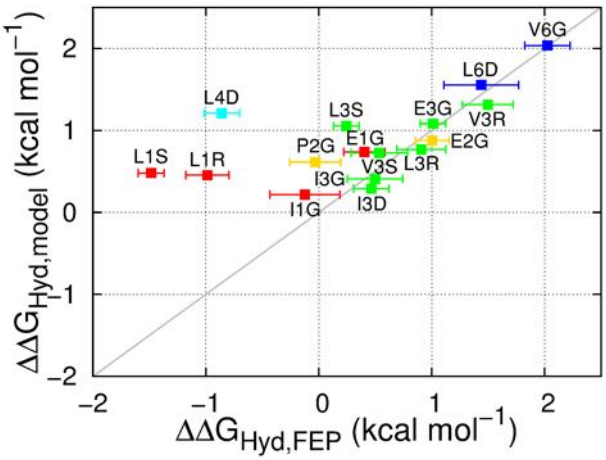



**Fig. 4** Correlation between the hydration free energy differences between fluorinated and non-fluorinated amino acids calculated with FEP ($\Delta\Delta G_{Hyd,FEP}$) or as a result of the multivariate linear model ($\Delta\Delta G_{Hyd,model}$) given by SI equation 6, originally reported in ref. 3. This model takes as input the differences in surface area, water–carbonyl hydrogen bonds, water–amine hydrogen bonds and molecular dipole moment. FEP data points are presented as mean ± standard deviation of five independent simulations. The color code corresponds to the number of fluorine atoms in the side chain: red = one; yellow = two; green = three; cyan = four; blue = six. The gray line indicates perfect correlation.

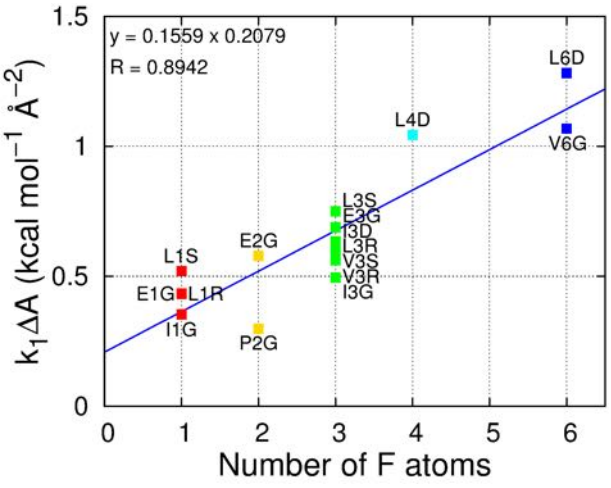



**Fig. 5** Correlation between the contribution of the surface area to the hydration free energy and the number of fluorine atoms *per* fluorinated group. The color code corresponds to the number of fluorine atoms in the side chain: red = one; yellow = two; green = three; cyan = four; blue = six. The blue line is a linear fit to the data points with the corresponding equation and regression coefficient at the top left. Related to Equation 1 in the main text.

simulations show that the donor–acceptor distance for these hydrogen bonds (SI Table 7) decreases with decreasing degree of fluorination of each alkyl group, indicating stronger interactions. Ab initio calculations, and the strong correlation between $^{19}$F NMR isotropic chemical shifts and the type of fluorine-protein interactions observed in the Protein Data Bank also suggest that hydrogen bonds to fluorinated alkyl groups are strongest for groups with low degrees of fluorination[30]. The substantially different values of the fitting parameters $k_{5-7}$ (see Table 1 in the main text), which translate the strength of water–fluorine hydrogen bonds, indicate this choice is necessary to understand interactions between fluorinated amino acids and water.

| | distance (Å) | | |
|---|---|---|---|
| | $-CH_2F$ | $-CHF_2/-CF_2-$ | $-CF_3$ |
| E1G | 2.711 | | |
| E2G | | 2.835 | |
| E3G | | | 2.951 |
| P2G | | 2.813 | |
| V3S | | | 2.966 |
| V3R | | | 2.959 |
| V6G | | | 3.005 |
| I1G | 2.740 | | |
| I3D | | | 2.925 |
| I3G | | | 2.971 |
| L1S | 2.705 | | |
| L1R | 2.701 | | |
| L4D | | 2.847 | |
| L3S | | | 2.928 |
| L3R | | | 2.939 |
| L6D | | | 2.987 |

**Table 7** Average distance between the water oxygen atom and the fluorine atom when a water–fluorine hydrogen bond is formed in a monofluoromethyl group ($-CH_2F$), a difluoromethyl/difluoromethylene group ($-CHF_2/-CF_2-$) or a trifluoromethyl group ($-CF_3$). Related to Equation 1 in the main text.

### 5.2.1 Details of the fitting procedure

The quantities used as input for fitting Equation 1 in the main text can be found in SI Table 8. The parameters $k_1$ to $k_7$ were fit in stages. The value of $k_1$ is simply the Lennard-Jones component of the hydration free energy of methane per unit area, for the reasons indicated in the main text (also, see SI Table 9). The values of $k_2$ and $k_3$, corresponding to the contributions of the hydration free energy per water-carbonyl and water-amine hydrogen bonds, respectively, are from ref. 3. The remaining parameters were optimized by fitting Equation 1 in the main text to the hydration free energies of all the amino acids using the input data in SI Table 8.

### 5.2.2 Calculating P-values

The model captures the difference in hydration free energies between the partially fluorinated amino acids and their non-fluorinated counterparts less well than for the fully fluorinated ones (Figure 4 in the main text). The cause for this discrepancy could possibly be related to an inadequacy of a linear model to represent either or both the water–fluorine hydrogen bond and the electrostatic potential terms in Equation 1 in the main text. To test the validity of using Equation 1 (see main text) to calculate the hydration free energy differences between fluorinated and non-fluorinated amino acids, we

| | ΔA, Å ($k_1$) | $\Delta h_{CO}$ ($k_2$) | $\Delta h_{NH}$ ($k_3$) | ΔΦ ($k_4$) | $h_{CH_2F}$ ($k_5$) | $h_{CHF_2}$ ($k_6$) | $h_{CF_3}$ ($k_7$) |
|---|---|---|---|---|---|---|---|
| E1G | 8.143 | -0.118 | -0.027 | 5.451 | 0.630 | 0.000 | 0.000 |
| E2G | 10.917 | -0.178 | 0.001 | 3.206 | 0.000 | 0.703 | 0.000 |
| E3G | 12.987* | -0.224* | 0.067* | 0.007 | 0.000 | 0.000 | 0.492 |
| P2G | 5.611 | -0.114 | -0.013 | 2.451 | 0.000 | 0.656 | 0.000 |
| V3S | 10.993* | -0.164* | 0.043* | 0.299 | 0.000 | 0.000 | 0.493 |
| V3R | 10.582* | -0.195* | -0.080* | 0.227 | 0.000 | 0.000 | 0.479 |
| V6G | 20.160* | -0.355* | -0.012* | 0.146 | 0.000 | 0.000 | 0.723 |
| I1G | 6.676 | -0.057 | 0.021 | 7.410 | 0.554 | 0.000 | 0.000 |
| I3D | 11.972* | -0.079* | 0.161* | 7.664 | 0.000 | 0.000 | 0.613 |
| I3G | 9.336* | -0.149* | 0.092* | 0.804 | 0.000 | 0.000 | 0.472 |
| L1S | 9.819 | -0.062 | 0.045 | 3.211 | 0.790 | 0.000 | 0.000 |
| L1R | 8.182 | -0.086 | 0.038 | 3.043 | 0.758 | 0.000 | 0.000 |
| L4D | 19.697 | -0.191 | 0.058 | 2.894 | 0.000 | 1.271 | 0.000 |
| L3S | 14.142* | -0.163* | 0.040* | 0.222 | 0.000 | 0.000 | 0.665 |
| L3R | 11.698* | -0.140* | 0.026* | 1.247 | 0.000 | 0.000 | 0.607 |
| L6D | 24.188* | -0.265* | 0.109* | 0.535 | 0.000 | 0.000 | 0.875 |

**Table 8** Input parameters associated with the coefficients $k_{1-7}$ in Equation 1 in the main text. All quantities represent the difference between fluorinated amino acids and their non-fluorinated parent, in each case averaged over 25000 configurations of the corresponding system. * Retrieved from ref. 3. ΔA is change in the amino acid surface area (calculated with a spherical probe with a radius of 1.4 Å, corresponding to that of a water molecule), $\Delta h$ is the change in the average number of the hydrogen bonds established between water and the functional group indicated as subscript, per amino acid; hydrogen bonds exist if the O···O distance <3.5 Å and the O-H···O angle is between 135° and 180°; $\Delta\Phi^{+}$ is the change in the number of water molecules, within 5.5 Å of the side chain carbon atoms, that experience a potential which is more positive than that of the waters at the $90^{th}$ percentile of the Φ probability distribution of the corresponding non-fluorinated amino acid. Related to Equation 1 in the main text.

| | $CH_4$ | $CF_4$ |
|---|---|---|
| SASA ($Å^2$) | 47.75 | 72.21 |

**Table 9** Solvent accessible surface area (SASA) of methane ($CH_4$) and tetrafluoromethane ($CF_4$). Related to Equation 1 in the main text.

have calculated P-values for the $\Delta\Delta G_{Hyd}$ of each amino acid by excluding it from the fit to Equation 1 in the main text. These P-values are presented in SI Table 10 and show that, particularly for E1G, E2G, I3D, L1S, L4D and L3S, the hydration free energy differences calculated with this type of model are not representative of those calculated with FEP. Using all amino acids to perform the fit, the P-values show an improvement (SI Table 10), where only E2G, L1S, L4D and L3S are not representative at a confidence level of 99%. Apart from E1G, E2G, I3D, L1S, L4D and L3S, all $\Delta\Delta G_{Hyd}$ values are representative to a confidence level of 90%. The emphasis should then be placed at the qualitative description of the observed trends in $\Delta\Delta G_{Hyd}$, observed in Figure 4 of the main text, which is the central message of this work.

### 5.2.3 The interplay between cavity-formation costs and gains from water-fluorine hydrogen bonds

The correlation between the energy required to form an amino acid-sized cavity in water and the energy gain from the formation of water–fluorine hydrogen bonds is found in SI Figure 6. It is clear that both the penalty associated with the surface area and the gain from hydrogen bond formation decrease with the increasing number of fluorine atoms *per* fluoromethyl group.

| | $\Delta\Delta G_{Hyd,Model}$ | P-value$_{Model}$ | C.L.$_{Model}$ | $\Delta\Delta G_{Hyd,Excl}$ | P-value$_{Excl}$ | C.L.$_{Excl}$ |
|---|---|---|---|---|---|---|
| E1G | -0.085 | 0.037 | 93 | -0.233 | 0.000 | 100 |
| E2G | 0.311 | 0.000 | 100 | 0.125 | 0.000 | 100 |
| E3G | 0.906 | 0.181 | 64 | 0.896 | 0.161 | 68 |
| P2G | -0.157 | 0.288 | 42 | -0.183 | 0.251 | 50 |
| V3S | 0.690 | 0.268 | 46 | 0.703 | 0.251 | 50 |
| V3R | 1.156 | 0.067 | 87 | 1.129 | 0.053 | 89 |
| V6G | 1.831 | 0.169 | 66 | 1.789 | 0.121 | 76 |
| I1G | -0.101 | 0.472 | 6 | -0.089 | 0.456 | 9 |
| I3D | 0.835 | 0.009 | 98 | 1.594 | 0.000 | 100 |
| I3G | 0.478 | 0.468 | 6 | 0.477 | 0.468 | 6 |
| L1S | -1.091 | 0.000 | 100 | -0.851 | 0.000 | 100 |
| L1R | -1.007 | 0.456 | 9 | -1.019 | 0.433 | 14 |
| L4D | -0.411 | 0.002 | 100 | 0.409 | 0.000 | 100 |
| L3S | 0.721 | 0.000 | 100 | 0.803 | 0.000 | 100 |
| L3R | 0.712 | 0.184 | 63 | 0.689 | 0.156 | 69 |
| L6D | 1.281 | 0.319 | 36 | 1.230 | 0.264 | 47 |

**Table 10** Difference in hydration free energy ($\Delta\Delta G_{Hyd}$, kcal mol$^{-1}$), P-value and confidence level (C.L.) calculated either from Equation 1 in the main text (Model) or as the result of a multivariate linear fit, having the same parameters as Equation 1 in the main text, where each amino acid is excluded (Excl). P-values are obtained from a one-sided z-test comparing the $\Delta\Delta G_{Hyd}$ value calculated with the model to the average $\Delta\Delta G_{Hyd}$ from the FEP simulations; confidence levels are calculated from the P-values. Related to Equation 1 in the main text.

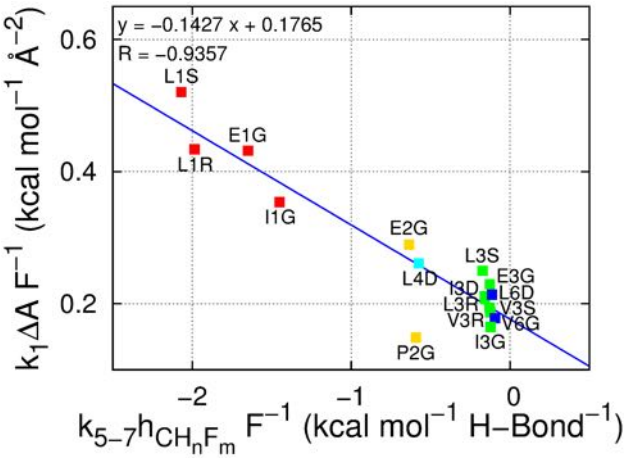



**Fig. 6** Surface area contribution to the free energy of hydration *per* fluorine atom ($k_1$ $\Delta$A F$^{-1}$) *versus* the water–fluorine hydrogen bonds contribution to the free energy of hydration *per* fluorine atom ($k_{5-7}$ h$_{CH_nF_m}$ F$^{-1}$, where $n = 0,1,2$ and $m = 1,2,3$). Contributions are calculated using Equation 1 in the main text. The color code corresponds to the number of fluorine atoms in the side chain: red = one; yellow = two; green = three; cyan = four; blue = six. The blue line is a linear fit to the data points with the corresponding equation and regression coefficient at the top left. Related to Equation 1 in the main text.

### 5.2.4 Radial distribution functions and hydration free energy

It would be useful to interpret the hydration free energy changes we have reported in terms of methyl/fluoromethyl–water radial distribution functions (RDFs). These RDFs are shown in SI Figure 7. From the RDFs, an excess free energy required to form the hydration shell of each methyl/fluoromethyl group along the $r$ reaction coordinate can be calculated as the product of the potential of mean force (PMF) required to bring a water molecule from bulk into the hydration shell and the number of water molecules in the hydration shell:

$$\Delta\Delta G_{Shell,PMF} = \sum_{r_0}^{r_{Shell}} -k_B T \ln\left[g(r)\right] g(r) V(r) \rho_{Bulk} \tag{7}$$

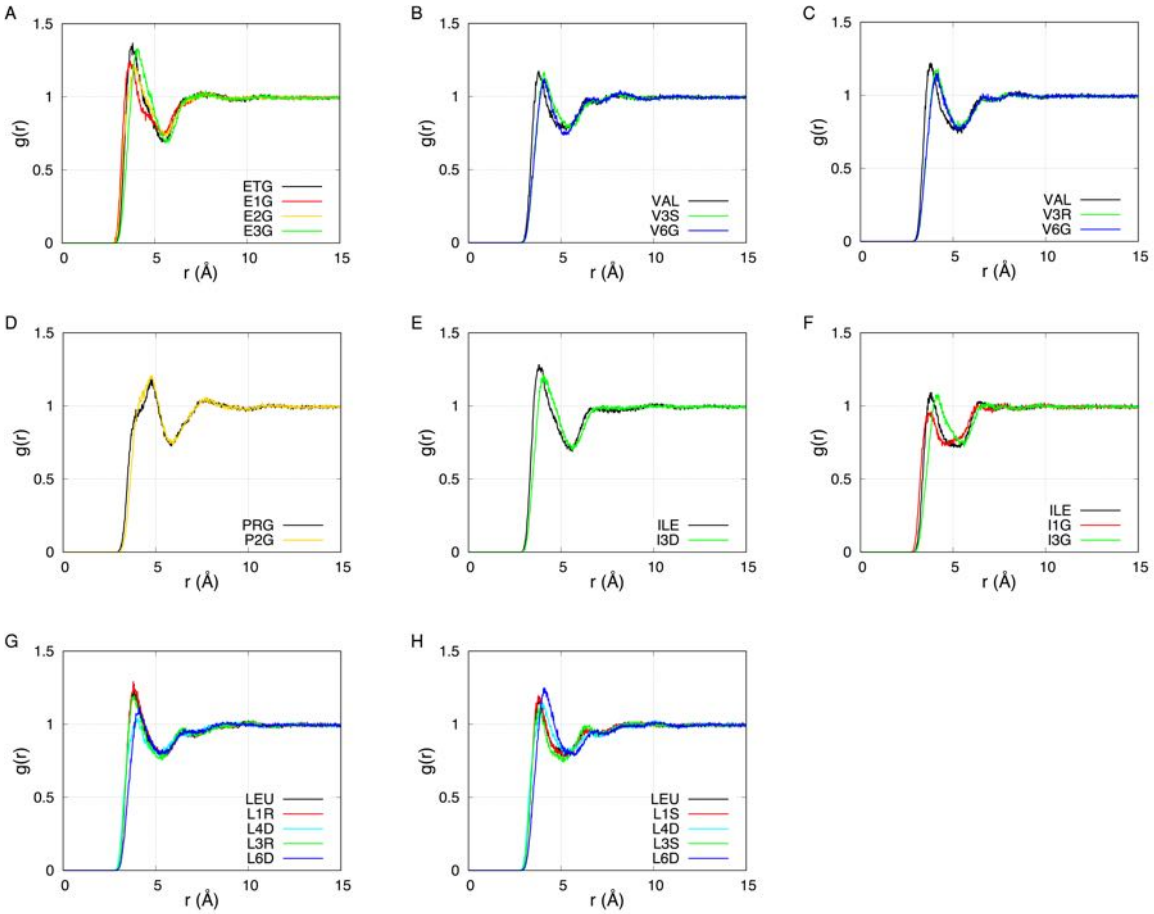



**Fig. 7** A-H) radial distribution functions (g(r)) of water oxygen atoms relative to the carbon atom in the indicated fluoromethyl group, or its equivalent methyl group in the non-fluorinated amino acid, for A) the ethylglycine series, B,C) the valine series, D) the propylglycine series, E,F) the isoleucine series and G,H) the leucine series; the color code follows the total number of fluorine atoms in the side chain: black=0, red=1, yellow=2, green=3, cyan=4, blue=6.

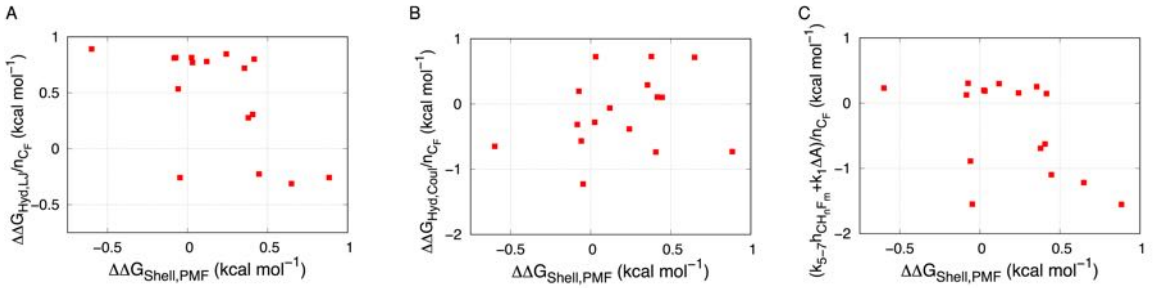



**Fig. 8** The excess free energy, $\Delta\Delta G_{Shell,PMF}$, required to form the hydration shell of each methyl/fluoromethyl group, calculated from Equation 7, *vs.* A) the Lennard-Jones component per fluorinated carbon of the free energy of hydration calculated using FEP ($\Delta\Delta G_{Hyd,LJ}/n_{C_F}$; B) the Coulombic component per fluorinated carbon of the free energy of hydration calculated using FEP ($\Delta\Delta G_{Hyd,Coul}/n_{C_F}$); C) the sum of the energetic contributions of the surface area and the fluorine–water hydrogen bonds, per fluorinated carbon, to the free energy of hydration given by Equation 1 in the main text (($k_{5-7}\ h_{CH_nF_m}+k_1\ \Delta A)/n_{C_F}$).

The term $-k_B T \ln[g(r)]$ corresponds to the PMF, with $k_B$ being the Boltzmann constant, $T$ the system temperature and $g(r)$ the radial distribution function. The term $g(r)V(r)\rho_{Bulk}$ is the particle density of water in a volume V(r) corresponding to the difference in volume of two spheres, one with radius $r$ and the other with radius $r-0.02$ (the spacing between values of $r$ at which the $g(r)$ is calculated); $\rho_{Bulk}$ is the particle density of bulk TIP4P-Ew water. The summation in Equation 7 runs over values of $r$ between $r_0$ (the C–O distance at which $g(r) > 0$) and $r_{Shell}$ (the radius of the hydration shell of either methane, 5.531Å, or tetrafluoromethane, 5.690Å). Defining $r_{shell}$ as the value for which each $g(r)$ reached its first minimum did not alter the trends described below.

We searched for correlations between the information contained in the RDFs, in the form of $\Delta\Delta G_{Shell,PMF}$, and both local and global measures of solvation. There is no correlation between $\Delta\Delta G_{Shell,PMF}$ and the Coulombic component of the hydration free energy per fluorinated carbon, $\Delta\Delta G_{Hyd,Coul}/n_{C_F}$, or the analogous Lennard-Jones component, $\Delta\Delta G_{Hyd,LJ}/n_{C_F}$ (Figure 8 A,B). $\Delta\Delta G_{Hyd,Coul}/n_{C_F}$ and $\Delta\Delta G_{Hyd,LJ}/n_{C_F}$ are global measures of hydration; these observables do not reflect only changes in local solvation around the fluorinated carbon. To isolate these local changes, we took advantage of the analytical model given by Equation 1 in the main text and summed over the contributions of the surface area and the fluorine-water hydrogen bonds. These results, shown in Figure 8 C, make clear that there is no correlation between this local measure of solvation and $\Delta\Delta G_{Shell,PMF}$.

## 6 Changes in side chain conformation induced by fluorination

To assess whether fluorination changes the occurrence of backbone-water hydrogen bonds by steric blockage, we calculated the distances between the carbon atom in a methyl or a fluoromethyl group (for example, $C^{\gamma}$ in the ethylglycine series, $C^{\delta}$ in the leucine series) and the center of mass of either the carbonyl or amine group bound to the alpha carbon. The probability distributions for ETG (SI Figure 9) show that the side chain adopts a preferred conformation relative to the amine group where the methyl-amine distance is approximately 3 Å, while simultaneously adopting two equally probable conformations relative to the carbonyl group, corresponding to distances close to 3 and 4 Å. Fluorinating this methyl group progressively increases the probability of visiting a conformation that is both farther from the amine group and closer to the carbonyl, in agreement with the decrease in the number of carbonyl-water hydrogen bonds and the increase in amine-water hydrogen bonds observed for this series of amino acids (SI Table 8).

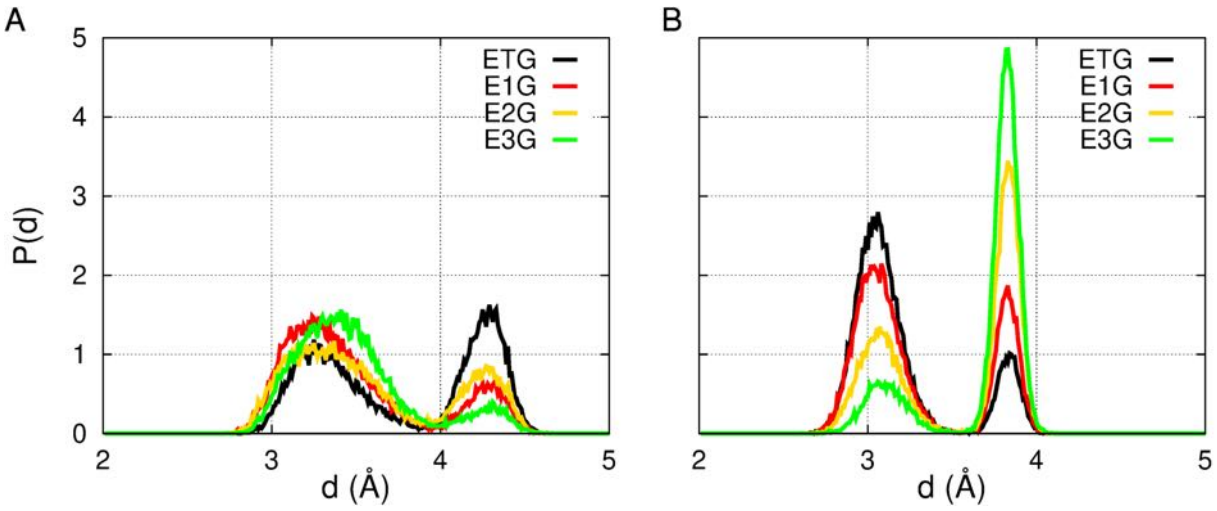



**Fig. 9** Probability density (P(d)) associated with the distances (d) between methyl or fluoromethyl groups in the side chain of ETG, E1G, E2G and E3G and the A) carbonyl group or B) amine group bound to the alpha carbon in the backbone.

## References


1 H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura and T. Head-Gordon, Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew, *J. Chem. Phys.*, 2004, **120**, 9665–9678.

2 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.

3 J. R. Robalo, S. Huhmann, B. Koksch and A. V. Verde, The Multiple Origins of the Hydrophobicity of Fluorinated Apolar Amino Acids, *Chem*, 2017, **3**, 881 – 897.

4 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, Development and testing of a general amber force field, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

5 H. Berendsen, D. van der Spoel and R. van Drunen, GROMACS: A message-passing parallel molecular dynamics implementation, *Comput. Phys. Commun.*, 1995, **91**, 43 – 56.

6 E. Lindahl, B. Hess and D. van der Spoel, GROMACS 3.0: a package for molecular simulations and trajectory analysis, *J. Mol. Model.*, 2001, **7**, 306–317.

7 D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, GROMACS: Fast, flexible, and free, *J. Comput. Chem.*, 2005, **26**, 1701–1718.

8 B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.

9 S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, *Bioinformatics*, 2013, **29**, 845–854.

10 S. Páll, M. J. Abraham, C. Kutzner, B. Hess and E. Lindahl, International Conference on Exascale Applications and Software, 2014, pp. 3–27.

11 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX*, 2015, **1**, 19–25.

12 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz,

I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J. A. Pople, *Gaussian 03, Revision E.01*, 2014.

13 J. Wang, W. Wang, P. A. Kollman and D. A. Case, Automatic atom type and bond type perception in molecular mechanical calculations, *J. Mol. Graph. Model.*, 2006, **25**, 247–260.

14 D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham III, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu and P. A. Kollman, *AMBER 14*, 2014.

15 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, LINCS: A linear constraint solver for Mol. Sim.s, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

16 J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, *J. Comput. Phys.*, 1977, **23**, 327–341.

17 U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, A smooth particle mesh Ewald method, *J. Chem. Phys.*, 1995, **103**, 8577–8593.

18 H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola and J. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

19 M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, 1989.

20 N. A. Baker, D. Sept, S. Joseph, M. J. Holst and J. A. McCammon, Electrostatics of nanosystems: Application to microtubules and the ribosome, *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 10037–10041.

21 C. H. Bennett, Efficient estimation of free energy differences from Monte Carlo data, *J. Comput. Phys.*, 1976, **22**, 245–268.

22 M. R. Shirts, E. Bair, G. Hooker and V. S. Pande, Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods, *Phys. Rev. Lett.*, 2003, **91**, 140601.

23 C. A. Gough, S. E. Debolt and P. A. Kollman, Derivation of fluorine and hydrogen atom parameters using liquid simulations, *J. Comput. Chem.*, 1992, **13**, 963–970.

24 E. Wilhelm, R. Battino and R. J. Wilcock, Low-pressure solubility of gases in liquid water, *Chem. Rev.*, 1977, **77**, 219–262.

25 A. Ben-Naim and Y. Marcus, Solvation thermodynamics of nonionic solutes, *J. Chem. Phys.*, 1984, **81**, 2016–2027.

26 M. Holst and F. Saied, Multigrid solution of the Poisson-Boltzmann equation, *J. Comput. Chem.*, 1993, **14**, 105–113.

27 M. J. Holst and F. Saied, Numerical solution of the nonlinear Poisson-Boltzmann equation: developing more robust and efficient methods, *J. Comput. Chem.*, 1995, **16**, 337–364.

28 A. Bondi, van der Waals volumes and radii, *J. Phys. Chem.*, 1964, **68**, 441–451.

29 W. Humphrey, A. Dalke and K. Schulten, VMD: visual molecular dynamics, *J. Mol. Graphics*, 1996, **14**, 33–38.

30 C. Dalvit and A. Vulpetti, Fluorine-Protein Interactions and F-19 NMR Isotropic Chemical Shifts: An Empirical Correlation with Implications for Drug Design, *ChemMedChem*, 2011, **6**, 104–114.