Adapting DFT+U for the Chemically Motivated Correction of Minimal Basis Set Incompleteness

Citation: Kulik, Heather J., Natasha Seelam, Brendan D. Mar, and Todd J. Martínez. "Adapting DFT+Ufor the Chemically Motivated Correction of Minimal Basis Set Incompleteness." The Journal of Physical Chemistry A 120, no. 29 (July 28, 2016): 5939–5949.

As Published: http://dx.doi.org/10.1021/acs.jpca.6b04527

Publisher: American Chemical Society

Persistent URL: http://hdl.handle.net/1721.1/110047

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Adapting DFT+U for the Chemically-Motivated Correction of Minimal Basis Set Incompleteness

Heather J. Kulik[†,*], Natasha Seelam[†], Brendan D. Mar[†] and Todd J. Martínez[§¶]

[†]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139 United States*

[§]*Department of Chemistry and PULSE Institute, Stanford University, Stanford, CA 94305*
[¶]*SLAC National Accelerator Laboratory, Menlo Park, CA 94025*

ABSTRACT: Recent algorithmic and hardware advances have enabled the application of electronic structure methods to the study of large-scale systems such as proteins with $O(10^3)$ atoms. Most such methods benefit greatly from the use of reduced basis sets to further enhance their speed, but truly minimal basis sets are well-known to suffer from incompleteness error that gives rise to incorrect descriptions of chemical bonding, preventing minimal basis set use in production calculations. We present a strategy for improving these well-known shortcomings in minimal basis sets by selectively tuning the energetics and bonding of nitrogen and oxygen atoms within proteins and small molecules to reproduce polarized-double-ζ basis-set geometries at minimal basis set cost. We borrow the well-known +U correction from the density functional theory community normally employed for self-interaction errors and demonstrate its power in the context of correcting basis set incompleteness within a formally self-interaction free Hartree-Fock framework. We tune the Hubbard U parameters for nitrogen and oxygen atoms on small molecule tautomers (e.g., cytosine), demonstrate the applicability of the approach on a number of amide-containing molecules (e.g., formamide, alanine tripeptide), and test our strategy on a 10 protein test set where anomalous proton transfer events are reduced by 90% from RHF/STO-3G to RHF/STO-3G+U, bringing the latter into quantitative agreement with RHF/6-31G* results. Although developed with the study of biological molecules in mind, this empirically-tuned U approach shows promise as an alternative strategy for correction of basis set incompleteness errors.

1. Introduction

Macromolecular structure-function relationships hold a key to addressing grand challenges in human health and energy utilization. Recent advances[1-7] in computational techniques enable fully *ab initio*, quantum chemical simulation of polypeptides[8-11]. Although first-principles methods directly treat charge transfer, polarization, and bond rearrangement needed to infer physicochemical relationships, geometry optimization and dynamic sampling of protein structures is often carried out with more restrictive semi-empirical[12-14] or non-polarizable force field[15-16] methodologies. Minimal atom-centered basis sets provide significant computational speed-up over larger basis sets, and their application would enable the greater use of electronic structure methods in protein studies. Small double-ζ basis sets have been demonstrated as a valuable approach for accelerating chemical discovery[17-18]. However, minimal basis sets are often excluded from production-level electronic structure calculations due to poor qualitative descriptions of bonding and geometry[10, 19].

In addition to well-known basis set superposition error (BSSE)[20] between separated fragments, these minimal basis sets suffer from intramolecular basis set superposition error (i-BSSE) and basis set incompleteness error (BSIE). Generally, BSSE refers to the artificial lowering of energy of a molecule in the multimolecular basis through the availability of unoccupied basis functions from another molecule. The Boys-Bernardi counterpoise scheme[21] was developed to correct for this form of BSSE, although it has not been without critique[22-27]. The intramolecular form of BSSE is observed[28-32] in large molecules, as differing regions of a molecule may borrow basis functions from each other. Several corrections to i-BSSE have been

proposed[33-37], but more fundamental corrections to the imbalanced descriptions of chemical bonding that we refer to as BSIE have been more restricted[19].

With the advent of electronic structure codes developed for graphical processing unit (GPU) architecture, fully minimal basis sets have again become advantageous to revisit due to enhanced performance of compact basis sets on GPUs[2] and surprisingly good performance of Hartree-Fock (HF) with minimal basis sets, as compared to density functional theory (DFT) or wavefunction theory with larger basis sets[38]. Some of us have previously observed[10] pathologies for the STO-3G minimal basis set during geometry optimization of a 55-protein data set with both Hartree-Fock and hybrid exchange-correlation functionals in density functional theory. Using protein health scores[39], we identified a high rate of steric clashing, i.e. unexpectedly short distances between atoms that are not bonded in a protein structure. These clashing events were traced[10] predominantly to the transfer of hydrogen atoms from nitrogen to neighboring oxygen atoms, especially along the amide backbone (Fig. 1 inset). Double-ζ basis sets are observed to greatly reduce clashing rates (Fig. 1), but for very large-scale simulation of proteins on the order of 3000 atoms on graphical-processing units[1-3, 40-41], a minimal basis set greatly reduces overhead and may be the most feasible option. Additionally, minimal basis set methods capable of producing reliable structural and bonding information could be useful for generating geometries that are good starting points for accurate, larger basis set calculations.

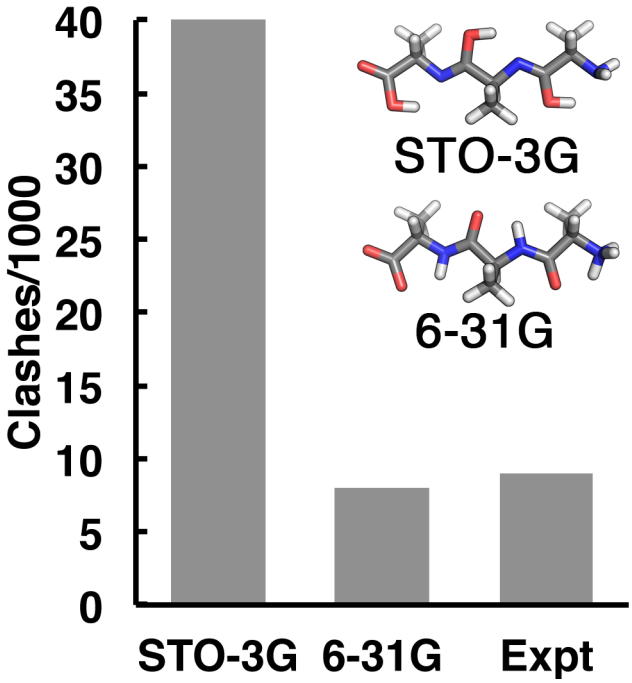


Figure 1. Clashing frequency as defined in the text for RHF geometry optimizations on a 55-protein data set with minimal (STO-3G) and double-ζ (6-31G) basis sets compared to experimental (X-ray, NMR) structures. An example of the source of these clashes is shown in the inset with imine nitrogen atoms observed on an STO-3G-optimized tri-alanine peptide.

In previous work also directed at improving the accuracy of minimal basis set calculations, Grimme and co-workers introduced the composite HF-3c method[19] which combines three distance-based corrections: i) the geometrical counterpoise correction (gCP)[37], ii) empirical van der Waals dispersion (D3)[42], and iii) short-range bonding (SRB) to improve i-BSSE- and BSIE-derived errors in MINI[43] minimal basis HF calculations. We have observed that bond lengths computed with the MINI family of basis sets are on average 0.07-0.10 Å longer than those computed with the STO-3G minimal basis set, and this particular shortcoming likely motivated the incorporation of the SRB correction. The SRB correction for atoms A and B separated by a distance of R_{AB} is:

$$E_{SRB} = -s \sum_{A}^{atoms} \sum_{A \neq B}^{atoms} (Z_A Z_B)^{3/2} \exp(-\gamma (R_{AB}^{cut}) R_{AB}) \ , \qquad (1)$$

where s = 0.03 and γ = 0.7 are global fitting parameters, Z_A and Z_B are nuclear charges, and R_{AB}^{cut} is a pre-defined distance cutoff defined in the dispersion correction. The incorporation of D3, gCP, and SRB corrections in HF-3c reduces clashing rates of 45/1000 observed in RHF/MINI[10] and 40/1000 in RHF/STO-3G[10] to 34/1000[19] over the same 55-protein data set we previously used to explore the fidelity of ab initio approaches for protein structure.[10] Thus, the number of clashes obtained with HF-3c remains high with respect to the 8/1000 RHF/6-31G value[10], despite improvement of other properties, such as average bond lengths. Re-examining the SRB correction in eq. 1, we note that the scaling with nuclear charge will favor proton transfer from nitrogen (*Z*=7) to oxygen (*Z*=8) atoms, and the modest reduction in clashing from 40-45 to 34 is likely due instead to the gCP term in the HF-3c correction. In order to achieve further improvement of minimal basis set calculations, it is useful to consider the chemical origins of the high clash scores and unexpected protonation states in minimal basis sets.

Motivated by the success of Hubbard U corrections in DFT to selectively tune frontier orbital energies and occupations of a target subshell in an approach commonly referred to as DFT+U, we investigate and validate +U corrections for treating minimal basis set incompleteness. In minimal basis sets, the observed anomalous proton transfer may be loosely interpreted as an imbalance in the relative electron or proton affinities of nitrogen and oxygen atom substituents of organic molecules. We previously observed BSIE-driven anomalous proton transfer to occur in both practical DFT and RHF calculations. In this work, we validate +U corrections on RHF minimal basis set calculations to streamline our efforts to treating basis set incompleteness, as distinct from the usual use of +U to ameliorate self-interaction error present in approximate DFT exchange-correlation functionals.

The outline of this paper is as follows. We review the theory and implementation of adding a Hubbard U term to electronic structure calculations in Section 2 and provide an overview of the details of calculations in Section 3. In Section 4, we develop, explain, and validate our approach. Finally, we provide our conclusions in Section 5.

2. Overview and Implementation of DFT+U/HF+U

Since its inception in the 1990s[44-47], the Hubbard-model ("+U") correction has been increasingly employed to approximately correct the well-known self-interaction error (SIE) of presently available DFT methods that lead to the over-delocalization of electronic subshells that should be highly localized (e.g. 3*d* or 4*f* electrons) in a method commonly known as DFT+U. Within the framework of SIE and band-gap corrections to semi-local DFT, +U corrections have also been applied to 2*p* electrons[48]. The original Hubbard model Hamiltonian was derived to describe a range of degrees of electron localization. The Hubbard U, or Coulomb repulsion of the electrons within the model Hamiltonian, corresponds to the energy required to remove an electron from one site and pair it with an electron on another site:

$$U_{nl}^{I} = IP_{nl}^{I} - EA_{nl}^{I} = E(N_{nl}^{I}+1) + E(N_{nl}^{I}-1) - 2E(N_{nl}^{I}) \ , \tag{2}$$

where *U* is the difference between the ionization potential (*IP*) and electron affinity (*EA*) for a particular atom (*I*) and subshell (*nl*) of electrons. Eqn. 2 may be recognized as a finite difference representation of the second derivative of the total energy with respect to N_{nl}^{I}, the number of electrons in the *nl* subshell:

$$U_{nl}^{I} = \frac{\partial^2 E}{\partial (N_{nl}^{I})^2} \ . \tag{3}$$

By invoking Koopmans'[49] or Janak's[50] theorem, this Hubbard U term is often expressed as a first derivative of orbital eigenvalues with respect to occupations of the *nl* subshell.

We employ a widely-adopted, simplified version of DFT+U[51] with the following functional form:

$$E^{\mathrm{HF/DFT+U}} = E^{\mathrm{HF/DFT}} + \frac{1}{2}\sum_{I,\sigma}\sum_{nl} U^{I}_{nl}[\mathrm{Tr}(\mathbf{n}^{I,\sigma}_{nl}) - \mathrm{Tr}(\mathbf{n}^{I,\sigma}_{nl}\mathbf{n}^{I,\sigma}_{nl})] \ , \tag{4}$$

where $\mathbf{n}^{I,\sigma}_{nl}$ is an occupation matrix of localized states in the *nl* subshell on atom *I*, σ is a spin index, and U^{I}_{nl} is the effective electron-electron repulsion interaction parameter that may be calculated[52-56] or, more commonly, tuned[57-59] and is specific to each atom and subshell. Although HF+U is unconventional, we have previously motivated[60] the use of a +U correction in the context of formally self-interaction free HF theory by highlighting the role of this term in altering electron localization and molecular orbital energies. A variety of definitions are available for the occupation matrices that enter into the +U energy functional. In the solid state, occupation matrices are obtained by projecting extended plane-wave-based molecular orbitals (bands) onto a localized, atomic basis set. Within a localized basis set formalism, occupations are easier to obtain. Here, it is most convenient to utilize elements of the Mulliken population matrix (**q**), which are defined as:

$$q_{\mu\nu} = \frac{1}{2}(P_{\mu\nu}S_{\nu\mu} + S_{\mu\nu}P_{\nu\mu}) \tag{5}$$

i.e., the entrywise product of the density (**P**) and overlap (**S**) matrices. If the system is closed-shell, population matrix values are reduced by a factor of two (i.e. a fully occupied orbital

corresponds to a matrix element of 1). Alternative definitions for occupation matrices, such as Löwdin populations[61], are also possible but generally will complicate analytic gradients for the Hubbard contribution to the forces. We re-express the +U correction in terms of a Mulliken population matrix $\mathbf{q}_{nl}^{I,\sigma}$, which spans all basis functions μ and ν centered on the *I*th atom and corresponding to the σ spin index and *nl* subshell:

$$E^{\mathrm{HF/DFT+U}} = E^{\mathrm{HF/DFT}} + \frac{1}{2}\sum_{I,\sigma}\sum_{nl} U_{nl}^{I}[\mathrm{Tr}(\mathbf{q}_{nl}^{I,\sigma}) - \mathrm{Tr}(\mathbf{q}_{nl}^{I,\sigma}\mathbf{q}_{nl}^{I,\sigma})] \ . \tag{6}$$

Throughout the rest of this article, we use the more commonly employed notation **n** to represent the block of the Mulliken population matrix that corresponds to oxygen or nitrogen 2*p* orbitals.

This +U correction is incorporated into the self-consistent calculation through direct modification of the potential:

$$V_{\mu\nu}^{\mathrm{HF/DFT+U}} = V_{\mu\nu}^{\mathrm{HF/DFT}} + V_{\mu\nu}^{\mathrm{U}} \equiv \frac{\partial E^{\mathrm{HF/DFT}}}{\partial P_{\mu\nu}} + \frac{\partial E^{\mathrm{U}}}{\partial P_{\mu\nu}} \ , \tag{7}$$

where the potential is added to a μν matrix element if both indices correspond to I^{th}-atom-centered *nl* subshell basis functions for which the corresponding U_{nl}^{I} parameter is nonzero. Thus, the total potential incorporates the dependence of the +U energy functional on the density matrix. The +U potential term may be further decomposed:

$$\frac{\partial E^{\mathrm{U}}}{\partial P_{\mu\nu}} = \frac{\partial E^{\mathrm{U}}}{\partial q_{\mu\nu}}\frac{\partial q_{\mu\nu}}{\partial P_{\mu\nu}} \ , \tag{8}$$

where the dependence of the +U energy on Mulliken population matrix elements (also denoted as $v_{\mu\nu}$) is explicitly:

$$\frac{\partial E^{\mathrm{U}}}{\partial q_{\mu\nu}} \equiv v_{\mu\nu} = \frac{1}{2} U^{I}_{nl} (\delta_{\mu\nu} - 2q_{\mu\nu}) \ . \tag{9}$$

Off diagonal elements of the occupation matrix thus only contribute through the derivative of the Tr(**qq**) term. The Mulliken population matrix then depends on density matrix elements as:

$$\frac{\partial q_{\mu\nu}}{\partial P_{\mu\nu}} = \frac{1}{2} (S_{\mu\nu} + S_{\nu\mu}) \, . \tag{10}$$

For the minimal basis sets employed in this work, the overlap matrix is simply the identity matrix because μ and ν must correspond to the same atom *I* for the correction to be applied and all same-subshell basis functions on an atom are orthonormal in a minimal basis set.

Nuclear gradient contributions due to the +U correction within the Mulliken population occupation matrix definition are:

$$\nabla_I E^{\mathrm{U}} = \sum_{nl \in I} \sum_{\mu\nu \in I, nl} \frac{\partial E^{\mathrm{U}^I_{nl}}}{\partial P_{\mu\nu}} \nabla_I P_{\mu\nu} + \frac{\partial E^{\mathrm{U}^I_{nl}}}{\partial S_{\mu\nu}} \nabla_I S_{\mu\nu} \ , \tag{11}$$

where the +U energy functional depends both on the density matrix (**P**) and overlap matrix (**S**) and their nuclear derivatives with respect to atom *I*. In eqn. 11, the first term is simply the DFT+U potential multiplied by the gradient of the density matrix, which is already accounted for in the self-consistent calculation. The second term may be expanded and simplified in the same fashion as was done in eqns. 9 and 10:

$$\frac{\partial E^{\mathrm{U}}}{\partial S_{\mu\nu}} = v_{\mu\nu} P_{\nu\mu} \, . \tag{12}$$

This term only needs to be added to the energy-weighted density matrix (**W**) for μ and ν elements corresponding to differing *n* and *l* values, as the diagonal blocks are already present in the energy-weighted density matrix. For cases where the overlap matrix is the identity matrix (i.e. the minimal basis sets used in this work), the overall correction to **W** vanishes.

3. Computational Details

Restricted Hartree-Fock (RHF) calculations were carried out using the TERACHEM[62] quantum chemistry package on a series of nucleobase tautomers (cytosine, thymine, and guanine), amide-bond model compounds (formamide, acetamide, N-methylacetamide, and alanine tripeptide), representative small molecules (ammonia, water), and 10 proteins. These molecules were structurally optimized with RHF/STO-3G[63], RHF/6-31G[64], RHF/6-31G*[65] and RHF/STO-3G+U method/basis set combinations, as specified throughout the text. For RHF+U calculations, the +U correction was applied to the 2*p* subshell of all oxygen and nitrogen atoms. Geometry optimizations were carried out to default thresholds of 4.5×10^{-4} hartree/bohr for the maximum gradient and 1×10^{-6} hartree for the change in self-consistent field energy between steps. For the large-scale optimizations, 10 proteins (PDB IDs: 1MZI, 1Y49, 1YJP, 2E4E, 2FXZ, 2OL9, 2ONW, 2RLJ, 3FTK, and 3FTR) were selected from a previously identified 55-protein data set[10]. This subset includes the structures that exhibited unexpected proton transfer when optimizing with STO-3G and also a control structure that did not exhibit problematic proton transfers in STO-3G optimization. The experimental structures obtained from the protein data bank[66] were protonated using the H++ webserver[67-69] at a pH of 7.0, regardless of the pH at which the proteins were experimentally solved. For NMR ensemble structures, the first structure was selected as the most representative of the ensemble for geometry optimization.

4. Results and Discussion

4.1 Tuning a +U Correction for Minimal Basis Sets

Spurious proton transfer from nitrogen to oxygen (see Fig. 1) occurs in amide backbones of even relatively large peptide molecules (100-600 atoms)[10], and we developed a test set of representative biological molecules that contain key chemical-bonding motifs observed in the tautomeric forms of these peptides. For oxygen, these species include carbonyl oxygen (=O) and hydroxyl groups bound to a carbon (OH). For nitrogen, the environments are more varied and include primary amines (R-NH_2), doubly-coordinated primary imines (R=NH), secondary amines (R_2NH), and secondary imines (R=N-R). Throughout the rest of this text, we refer to the nitrogen substituents by the shorthand NH*/N* for primary and secondary imines and NH_2/NH for primary and secondary amines, respectively. In peptides, the driving force is mixed between neutralization of charged termini and imbalance in neutral tautomer relative energies, whereas in our test cases, only the latter effect is captured. We will demonstrate that these neutral tautomers are a suitable proxy for both features of proteins in our benchmark of large protein geometry optimizations (Sec 4.3). The STO-3G[70] minimal basis set is the focus of our study even though MINI[43] basis sets have been noted on occasion to reduce basis set superposition error[71]. MINI basis sets do not appreciably reduce proton transfer events and are known to predict significantly elongated bond lengths (see Ref. [10] supporting information), which would necessitate additional corrections[19].

As a first test case, the nucleobase cytosine has three very closely spaced tautomers, referred to here by letters and the substituent oxygen and nitrogen atoms: A (=O, NH, N*, NH_2), B (OH, 2xN*, NH_2), and C (=O, 2xNH, NH*) (structures in Fig. 2). Polarized double-ζ RHF/6-

31G* calculations predict A and B to be nearly degenerate in energy, whereas the C tautomer is ~ 3 kcal/mol higher in energy (Fig. 2). These RHF/6-31G* energetics are in very close agreement with correlated quantum-chemistry results[72] (potentially due to cancellation of errors). As expected, RHF/STO-3G overstabilizes the (OH, 2xN*)-containing tautomer B with respect to the species that contain more three-coordinate nitrogen species and lower-coordinate oxygen. The minimal basis set destabilizes A and C tautomers by 15 kcal/mol with respect to a B tautomer ground state (Fig. 2). The root sum square (RSS) energetic errors are computed with respect to reference relative energies as:

$$\Delta E_{rss} = \sqrt{\sum_{j=1}^{n_{tautomers}-1} (\Delta E_{0\rightarrow j} - \Delta E_{0\rightarrow j,ref})^2} \quad , \tag{13}$$

where $\Delta E_{0\rightarrow j}$ is the relative energy for a given method/basis set between a reference (0^{th}) state and the j^{th} state, and the ref subscript refers to the set of results against which the RSS error is assessed. The RSS energetic error for RHF/STO-3G with respect to RHF/6-31G* results is nearly 20 kcal/mol.

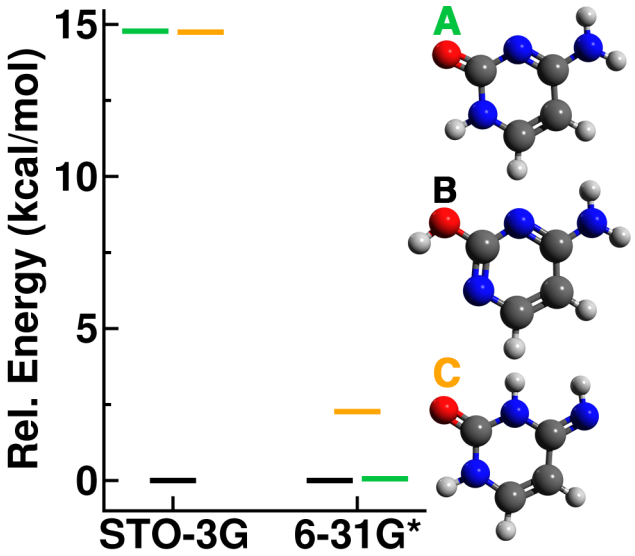


Figure 2. (left) Relative energetics of cytosine tautomers for RHF/STO-3G and RHF/6-31G* and (right) structures of the tautomers from RHF/6-31G* optimizations with tautomer labels color-coded according to symbols in the graph.

It is useful to identify whether the energetic imbalance present in RHF/STO-3G calculations is primarily derived from the description of the nitrogen and therefore only requires evaluating a U parameter on the nitrogen atoms, U_N, or whether it is also necessary to incorporate a U parameter on oxygen atoms, U_O. We investigated the RSS error dependence of the cytosine tautomers on applied positive and negative values of U_O and U_N. The justification for negative U values is to apply opposing forces to the two species in order to amplify the effect of a +U correction. A contour plot over a wide range of applied U values (U_O=[-8.75,8.75 eV] and U_N=[-8.75,8.75 eV]) reveals that a U correction is needed for both oxygen and nitrogen species to fully minimize RSS errors (Fig. 3). The U term on nitrogen appears to play the primary role, as exclusive use of U_O never lowers the RSS error below the standard STO-3G value (see also Supporting Information Tables S1-2). Large positive U_N values alone reduce the RSS error to around 10 kcal/mol, and RHF/6-31G* energetic ordering is reproduced with U_O=-6 eV and U_N=+6 eV. The chemical meaning behind these parameter choices is considered in Sec 4.2.

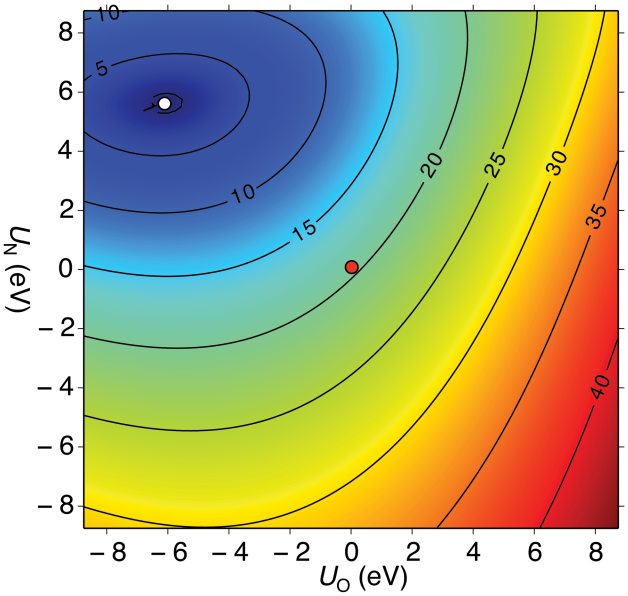


Figure 3. Contour plot of root mean square energy errors (in kcal/mol) between the A, B, and C tautomers of cytosine in RHF/STO-3G+U calculations with respect to RHF/6-31G* results. The RHF/STO-3G result is highlighted with a red circle, and the minimum-error U-pairing (U_O=-6 eV, U_N=+6 eV) is indicated with a white circle.

We now verify the transferability of these cytosine-trained U parameters to the energetics of the tautomers of formamide, the smallest model of the amide bond in protein backbones. Formamide has an iminol tautomer in which a hydrogen atom is transferred from the primary amine to the oxygen atom (see inset of Fig. 4). Minimal basis set errors are again apparent for the energetics of formamide tautomers: with RHF/STO-3G, the iminol (OH, NH*) tautomer is only 5 kcal/mol higher in energy than the amide ground state (=O, NH_2), whereas the RHF/6-31G* iminol-amide splitting is 21 kcal/mol. We compare a range of $U_O = -U_N$ values assuming that the =O/OH and NH_2/NH* tuning in this molecule would similarly require opposing tuning as was observed for cytosine. With this constraint, the $-U_O = U_N$ =+6 eV pairing previously selected for cytosine yields good agreement with the polarized-double-ζ basis set splitting at around 22 kcal/mol (Fig. 4). More importantly, the general qualitative trend is preserved (Fig. 4) that a positive U value on nitrogen atoms and negative U value on oxygen atoms increases the energetic penalty for proton transfer to the carbonyl oxygen.

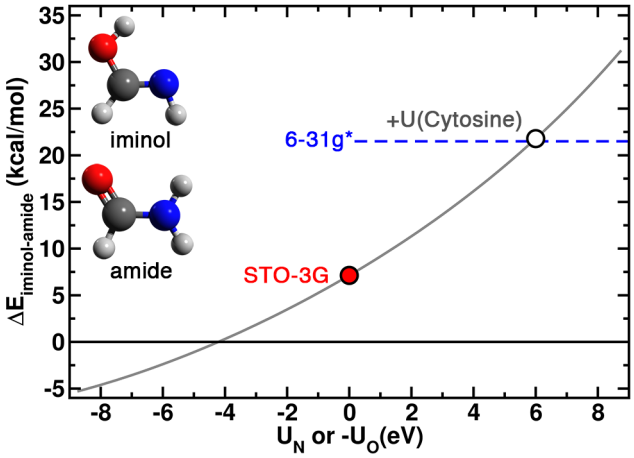


Figure 4. Plot of the relative energetics (in kcal/mol) of the amide and iminol tautomers of formamide (structures shown in inset) for RHF/STO-3G+U calculations. The applied values of U are set such that U_O=-U_N everywhere and the x-axis corresponds to the applied value of U_N. A blue dashed line indicates the relative energetics for a larger 6-31G* basis set. The standard RHF/STO-3G result is indicated with a red circle, and the result from applying the cytosine-tuned U values (U_O=-6 eV, U_N=+6 eV) is indicated by a white circle.

Having validated this approach on cytosine and formamide, we generalize it to the tautomers of the guanine and thymine nucleobases. We introduce qualitative metrics to generate a score to represent how faithfully the RHF/STO-3G and RHF/STO-3G+U (U_O = -6 eV, U_N = +6 eV) approaches reproduce the RHF/6-31G* relative tautomer ordering. These metrics include a ground state score (GS) and high-energy state score (HES) that are 0 if the method correctly identifies the ground state or high-energy state, respectively, and 1 if the method is incorrect. Additionally, the relative positioning any mid-state tautomers is defined as a mid-state ratio:

$$\mathrm{MS(basis)} = \frac{E^{\mathrm{basis}}_{\mathrm{MS(ref)}} - E^{\mathrm{basis}}_{\mathrm{GS(ref)}}}{\max(E^{\mathrm{basis}}) - \min(E^{\mathrm{basis}})} \quad , \tag{14}$$

where $E^{\mathrm{basis}}_{\mathrm{MS(ref)}}$ and $E^{\mathrm{basis}}_{\mathrm{GS(ref)}}$ are the mid-state and ground-state tautomer following the reference basis set's ordering evaluated with the current basis set and the denominator is the range of energies of all tautomers evaluated with the current basis set. For cases with multiple mid-state

tautomers, we compute each ratio separately and subscript it with a number reflecting tautomer ordering in the reference basis. We define the MS ratio score as:

$$\mathrm{MSS} = \left|\mathrm{MS(basis)\text{-}MS(ref)}\right| \ . \tag{15}$$

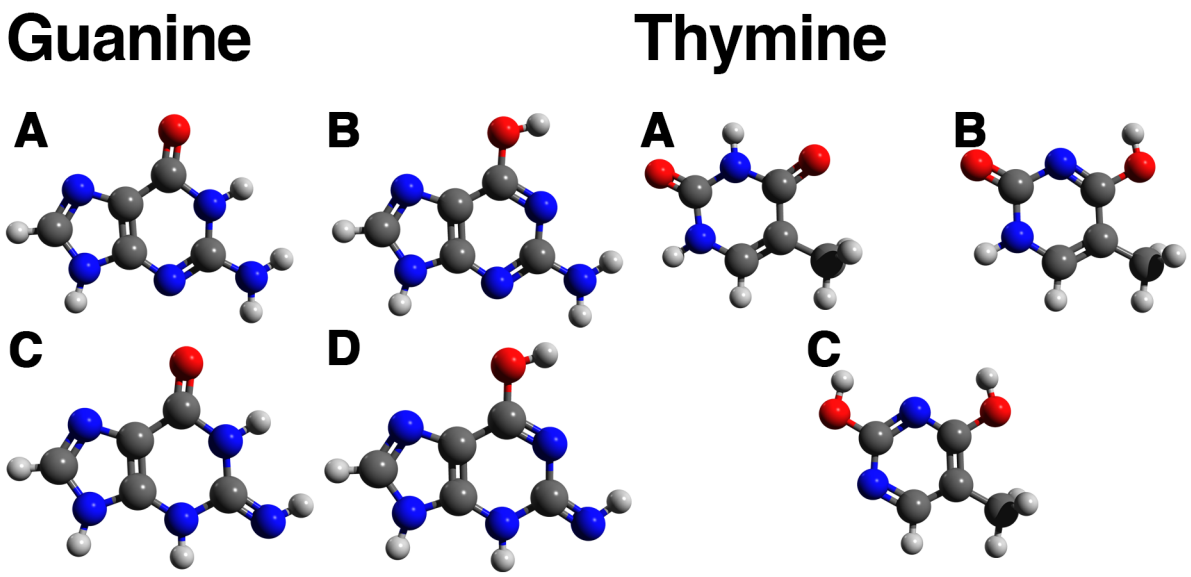
We define a total quality score (QS) as a composite of the GS, HES, and MSS with the following weighting:

$$\mathrm{QS} = 2 * \mathrm{GS} + \mathrm{HES} + \sum_i \mathrm{MS}_i\mathrm{S} \ , \tag{16}$$

where a QS close to 0 indicates maximum qualitative agreement with the reference basis. In addition to these qualitative scores, quantitative relative energetics are provided in Supporting Information Tables S3-S4 for guanine and Tables S5-S6 for thymine.

Guanine has five nitrogen atoms and one oxygen atom, making it much more nitrogen rich than typical peptides or the previous test cases. The four lowest energy tautomers are distinguished by the protonation state of the oxygen atom and three of the nitrogen atoms (Fig. 5). Minimal basis set RHF/STO-3G greatly stabilizes the hydroxyl-containing tautomers, B (OH, N*, NH_2,N*) and D (OH,N*,NH*,NH) over the carbonyl-containing tautomers, A (=O, NH, NH_2, N*) and C (=O, NH, NH*,NH). The RHF/6-31G* results reverse this ordering and stabilize the A (ΔE_{AB}=1.5 kcal/mol) and C (ΔE_{CD}=11 kcal/mol) tautomers, and RHF/STO-3G+U (U_O = -6 eV, U_N = +6 eV) calculations are in qualitative agreement (Fig. 5). The RHF/STO-3G+U results predict the correct ground state tautomer (A) and highest energy tautomer (D) but do not quantitatively reproduce mid-state tautomer (B, C) ordering. The 0.6 QS score obtained

with RHF/STO-3G+U is a marked improvement over the 2.3 QS with RHF/STO-3G due to the latter approach's incorrect ground state assignment (Table 1).



Basis	Ordering
STO-3G	B < A, C ~ D
6-31G*	A < B, C < D
STO-3G+U	A < B, C < D

Basis	Ordering
STO-3G	C < A < B
6-31G*	A << B < C
STO-3G+U	A << B < C

Figure 5. Structure and qualitative energetic ordering of guanine (left) and thymine (right) tautomers for RHF/STO-3G, RHF/6-31G*, and RHF/STO-3G+U.

Table 1. Comparison of qualitative ordering scores: ground state (GS), high-energy state (HES), mid-state ratio (MS), mid-state ratio score (MS S), and quality score (QS) for cytosine, guanine, and thymine tautomers obtained with RHF and STO-3G and STO-3G+U (U_O = -6 eV, U_N = + 6 eV) scored against a 6-31G* reference.

Basis	GS	HES	MS_1	MS_1 S	MS_2	MS_2 S	QS
				Cytosine			
STO-3G	0	1	1.0	1.0	--	--	2.0
STO-3G+U	0	0	0.1	0.0	--	--	0.0
6-31G*	0	0	0.0	0.0	--	--	0.0
				Guanine			
STO-3G	1	0	-0.3	0.3	0.3	0.0	2.3

STO-3G+U	0	0	0.3	0.3	0.3	0.3	0.6
6-31G*	0	0	0.1	0.0	0.6	0.0	0.0
				Thymine			
STO-3G	1	1	0.5	0.4	--	--	3.4
STO-3G+U	0	0	0.8	0.1	--	--	0.1
6-31G*	0	0	0.9	0.0	--	--	0.0

Although directly applying cytosine-tuned U values preserves the qualitative ordering predicted by the larger basis set (see Fig. 5), it does not greatly reduce RSS errors with respect to the RHF/STO-3G values (See Supporting Information Tables S3-4). Instead, RSS energetic error would be minimized by applying a negative U_O to the oxygen atoms and omitting any treatment of the nitrogen atoms. We can rationalize the differences between optimal U parameters for guanine and cytosine in terms of the high nitrogen abundance in guanine. The A/B and C/D tautomers of guanine represent two sets of carbonyl to carboxyl tautomers that are distinguished by either having an N^*/NH_2 configuration or an NH/NH* configuration. Comparison of ΔE_{AC} and ΔE_{BD} energetic splittings isolates the relative effect of these nitrogen configurations and reveals remarkably close agreement between STO-3G and 6-31G* basis sets (ΔE_{AC} =14.4 kcal/mol for STO-3G and 14.6 kcal/mol for 6-31G*, ΔE_{BD} = 26.1 kcal/mol for STO-3G and 24.5 kcal/mol for 6-31G*). Therefore, any U_N correction applied to STO-3G is likely to artificially shift up C and D tautomers. Nevertheless, the ground state tautomer and qualitative ordering is preserved with RHF/STO-3G+U, which is most relevant for the large scale geometry optimizations that are the focus of this method.

For the three representative low energy tautomers of thymine (Fig. 5), minimal basis sets again strongly stabilize a hydroxyl rich tautomer (C: N*, OH, N*, OH). The minimal basis set also unexpectedly stabilizes one carbonyl tautomer (A: NH, =O, NH, =O) over a hydroxylated tautomer (B: NH, =O, N*, OH), reversing RHF/6-31G* relative energies (A << B < C). In contrast, the polarized-double-ζ qualitative ordering is preserved with RHF/STO-3G+U using the cytosine-tuned U values, as indicated by a QS of 0.1 compared to the STO-3G QS of 3.4 (see Table 1). However, applying cytosine-tuned U values overstabilizes the carbonyl-rich A tautomer (Supporting Information Tables S5-S6). Thus, this single-atom-parameter approach appears suitable for reproducing qualitative, if not quantitative, energetic ordering of the larger basis set.

4.2 Origins of Chemical Specificity in +U Tuning

Although it should be possible to minimize RSS errors, as we have demonstrated in cytosine, with two parameters, the U corrections must distinguish the occupation matrices of the three differing tautomers to produce the correct energetic shifts. Now, we consider the source of the utility of our approach in correcting minimal basis set energetics. Generally, the +U correction performs two roles in energetic tuning: 1) for a fixed set of occupations, the energetic penalty is maximal for any orbital that is half full (E^U=U/8 per electron), parabolically reducing to zero for a filled or empty orbital in the occupation matrix; and 2) the potential shifts occupations and hybridization to encourage (discourage) filling of n > ½ orbitals and discourage (encourage) occupation of n < ½ orbitals for positive (negative) U values. For the systems studied here, the occupation shifts are small, and applying moderate U parameters does not substantially vary the atomic orbital occupation of a given molecular orbital. We thus alter

energetic splittings by penalizing differences in the Tr[**n**(1-**n**)] term in the energy functional, which we refer to as the fractionality of the occupations. If the difference in tautomer fractionality is a sufficient fingerprint of each functional group, then the energy correction parameters will be transferable.

We thus investigate the specificity of the RHF/STO-3G+U approach for correcting the imbalances in minimal basis sets and quantify relative effects of this correction for qualitatively distinct oxygen (=O, OH) and nitrogen (NH2, NH, N*, NH*) chemical bonding environments. By examining differences in Tr[**n**(1-**n**)] values for chemical species, we are able to interpret the U values that minimized RSS errors on cytosine tautomer energetics. The species we previously identified as chemically stable (NH_2, NH) but understabilized in the presence of oxygen in RHF/STO-3G simulations have the least-fractional occupation matrices, whereas the NH* and N* species are considerably more fractional (Fig. 6), motivating a positive U_N to destabilize the latter geometries. The same trend is apparent for formamide, although the NH_2 species exhibits slightly more fractional occupations than observed for the cytosine case. We confirm the need for a negative U_O by examining trends in oxygen Tr[**n**(1-**n**)]: occupation matrices of the hydroxyl oxygen are less fractional than the carbonyl oxygen cases for both formamide and cytosine, although the differences are less pronounced. Further, it is clear that U_N fixes A-B cytosine tautomer energetics, whereas U_O primarily stabilizes the C tautomer, which would otherwise be destabilized by the NH* substituent with high fractional occupations.

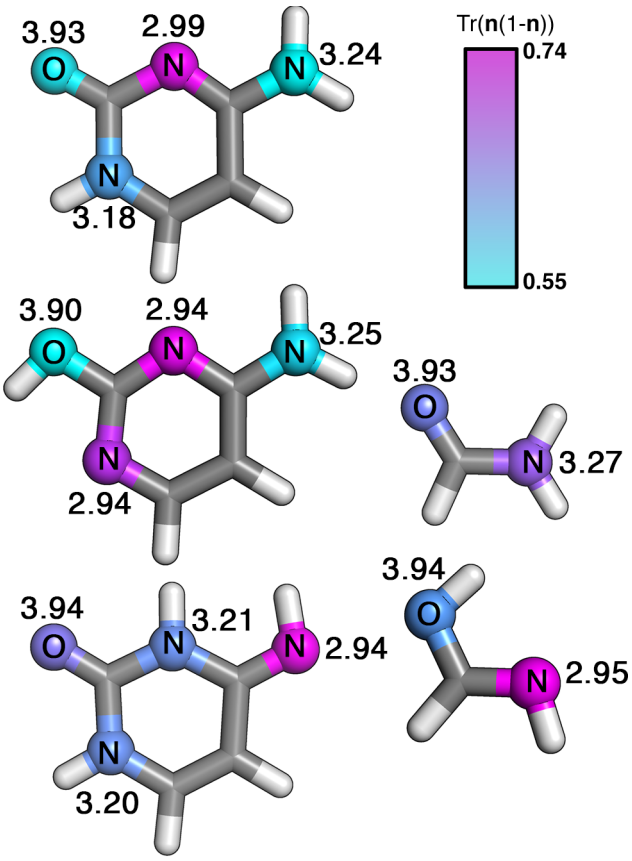


Figure 6. Occupation matrix properties [Tr(**n**) and Tr(**n**(1-**n**))] for 2*p* electrons of oxygen and nitrogen atoms in A, B, and C tautomers of cytosine (left, top to bottom) and formamide and its iminol tautomer (right, top to bottom). Nitrogen and oxygen atoms are shown as spheres and colored by Tr(**n**(1-**n**)) values from cyan (0.55) to magenta (0.74), as indicated by color bar at top right. Numbers adjacent to atoms are the sum of occupations, Tr(**n**), for each set of N and O 2*p* electrons.

We now extend comparison of nitrogen and oxygen atom fractionality across the previously discussed cytosine, guanine, thymine, and formamide to also include results from the larger amide bond models acetamide, N-methylacetamide, and a tri-alanine peptide. This broader test set of molecules preserve the trends observed for cytosine and formamide (see Fig. 6). For oxygen, the range of fractional occupation values is wide compared to the decrease from carbonyl to hydroxyl oxygen, but the trend is preserved for all compounds, and the shift at low (thymine) and high (formamide) values is consistent (Fig. 7). In the case of nitrogen, NH* and

N* fractional occupations are either very- or reasonably-well-clustered, respectively, suggesting significant promise for the applicability of this approach to a variety of systems. The NH and NH_2 distributions are broader but exhibit increasingly fractional occupations when dehydrogenation occurs. We also use this data to identify why cytosine-tuned U parameters would not be suitable for obtaining quantitative energy differences of guanine tautomers. We had observed that guanine tautomers A/B are overstabilized with respect to the C/D pair with RHF/STO-3G+U. These two sets are distinguished by NH* vs. NH_2 configurations, where we previously identified the NH* species as having the most fractional occupation matrices that are well clustered across all compounds. The particular NH_2 data points for guanine are also the least fractional that we observe across all compounds, giving a larger difference between the $+U_N$ energy contributions in these two compounds compared to more modest differences observed for other compounds. Overall, the trend in the extent of fractional occupation across different chemical substituents demonstrates promise to improve qualitative predictions of bonding in minimal basis sets on larger systems. However, these results also demonstrate the limitations for a single-parameter atom based approach to describe highly variable chemical environments in order to reproduce truly quantitative energetic orderings.

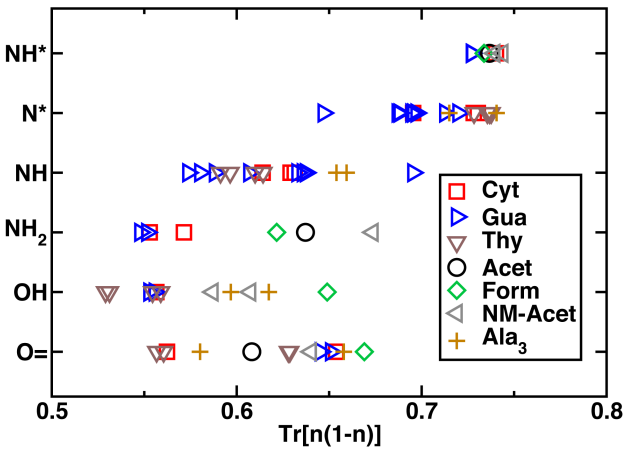


Figure 7. Trends in Tr[**n**(1-**n**)] values of the 2*p* occupation matrix from Mulliken populations for various chemical configurations of oxygen and nitrogen in several molecules (cytosine, guanine, thymine, acetamide, formamide, N-methylacetamide, and trialanine, as shown in legend). For nitrogen, N* and NH* indicates species that are doubly-coordinated versus triply-coordinated NH and NH_2.

We now reconsider how cytosine-tuned U values impact frontier orbital energies that are oxygen- or nitrogen-centered to further motivate parameter choice. Water and ammonia are employed as test molecules to separately investigate oxygen and nitrogen tuning. In both cases, the highest occupied molecular orbital (HOMO) has strong $2p_z$ character with nearly integer occupations. A molecular orbital that resembles a nearly completely occupied atomic orbital will have a maximally negative U-dependent potential of around -1/2 eV/eV of U (Fig. 8). We thus observe that a positive U on nitrogen lowers the HOMO of ammonia, increasing the Koopmans' ionization potential, whereas conversely negative U values on oxygen increase the HOMO of water, decreasing the Koopmans' ionization potential. Although it is tempting to envision the under-coordinated nitrogen atoms (N*, NH*) as N^+ or NH^+, the total occupation of the 2*p* states on these atoms (Tr[**n**]) shows subtle differences (2.94 vs. 3.25 for NH_2), and there is no discernible difference between carbonyl and hydroxyl oxygen atom values. Although populations and partial charges are highly sensitive to the partitioning scheme employed[73], we highlight the observations from the trace of the Mulliken-population-based occupation matrix because we directly manipulate these quantities when applying a +U correction. Thus, the parameter choice may be loosely interpreted as raising or lowering the effective hybridization and filling preference of orbitals corresponding to the substituent atoms to which the correction is applied.

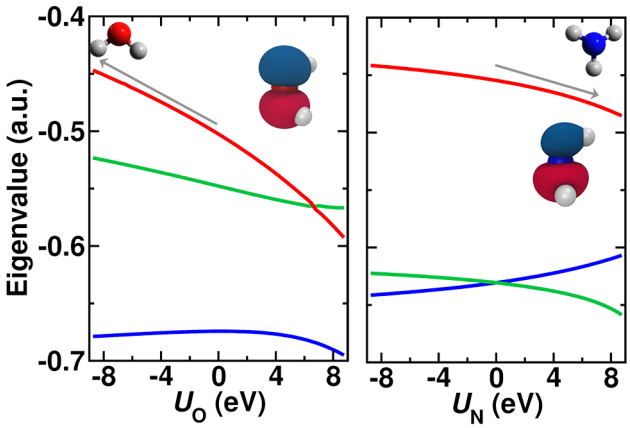


Figure 8. Dependence of frontier orbital energies (the HOMO isosurface is shown in inset and corresponds to the red curve) on applied $U_{(O,N)}$ values for water (left) and ammonia (right). The direction of molecular orbital energy shifts with U signs suggested from cytosine fitting are indicated with gray arrows.

4.3 Validating RHF/STO-3G+U on a Peptide Test Set

Finally, we validate the RHF/STO-3G+U approach with cytosine-tuned U values (U_O = -6 eV, U_N = 6 eV) on a set of 10 peptides selected from a larger 55-protein data set[10]. We include nine peptides (PDB IDs: 1MZI, 1Y49, 1YJP, 2E4E, 2FXZ, 2OL9, 2ONW, 3FTK, and 3FTR) that were previously observed[10] to exhibit the largest percentage of spurious proton transfer (PT) events from the original data set. The tenth peptide (PDB ID: 2RLJ) had comparable structure with STO-3G and 6-31G basis sets and is therefore included as a control structure to confirm STO-3G+U preserves good STO-3G behavior. The experimental stick and cartoon structures of the ten proteins are shown in Figure 9 and are colored by the number of PT events per residue at the RHF/STO-3G level (red is highest and blue or black are lowest). The largest number of proton transfer events occurs in linear, unstructured peptides (1YJP, 2OL9, 2ONW, 3FTK, 3FTR), although we do observe significant proton transfer in a number of helical- (1MZI, 2FXZ) and turn- (1Y49, 2E4E) containing peptides. Review of the primary sequence of the nine HT-abundant peptides (Table 2), demonstrates a high occurrence of the amide-sidechain-containing

asparagine in the test set (14.5%) vs. the human genome[74] (HG, 3.1%). Although the effect is not as pronounced for glutamine, it is still well-represented (4.8% vs. 4.7% in HG[74]).

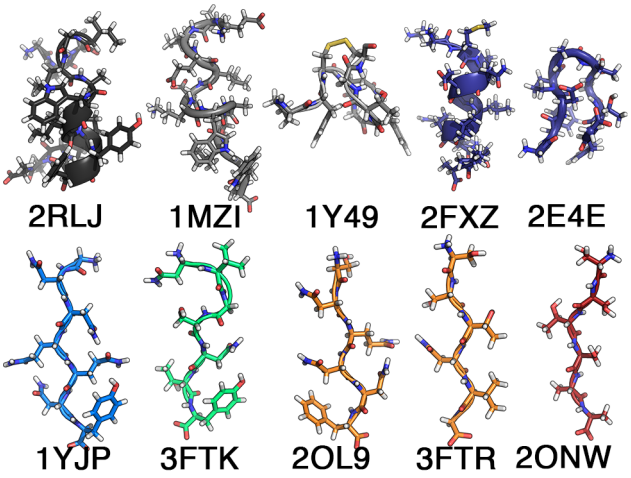


Figure 9. Stick and cartoon representation of experimental peptide structure test set. The peptides are ordered and backbones colored according to the ratio of proton transfer events observed in small basis set optimizations to the total number of residues in the peptide (lowest: black, highest: red).

Table 2. Summary of proton transfer (PT) events in protein data set for RHF/6-31G* (lg.), RHF/STO-3G (sm.) and RHF/STO-3G with U_O=-6 eV and U_N=+6 eV (+U).

PDB ID	Sequence	Res	At.	PT lg.	PT sm.	PT +U
1MZI	ELLELDKWASLWN	13	226	0	4	0
1Y49	PFCNAFTGC	9	122	0	3	0
1YJP	GNNQQNY	7	107	0	4	0
2E4E	GYDPATGTFG	10	129	1	4	1
2FXZ	KMVNEALVRQGLA	13	209	1	5	1
2OL9	SNQNNF	6	93	0	5	0
2ONW	SSTSAA	6	70	0	6	0
2RLJ	GAAIGLAWIPYFGPAA	16	224	1	1	1
3FTK	NVGSNTY	7	100	0	5	1
3FTR	SSTNVG	6	76	0	5	0

A few of the proteins in this test set have been the focus of previous studies with either HF or DFT. The HF-3c study[19] on these 10 proteins demonstrated higher clashing rates (42) than

average (34) across the larger data set, and the authors noted backbone bending for the linear peptides (2ONW, 3FTK, 3FTR) and changes in helical properties in 2RLJ. Liu et al.[75] carried out a comprehensive study of the conformational landscape of the 3FTR peptide, confirming a strong preference of minimal basis set calculations for anomalous protonation states that could be counteracted partially by inclusion of implicit solvent even with a minimal basis but required more basis functions to fully stabilize the normal protonation states. Periodic, crystalline models of packed 1YJP protein have been studied recently[76] with semi-local DFT functionals, and this peptide was observed to consistently prefer the zwitterionic charge state. However, earlier work[10, 77] has highlighted challenges associated with the general application of semi-local DFT to large models of proteins due to unphysical closing of the HOMO-LUMO gap.

We carried out gas phase geometry optimizations of each of the 10 proteins starting from experimental structures, and coordinates for all optimized proteins are provided in the Supporting Information. The protonation approach we employ[68] assumes the presence of a solvent environment, but some protonation states that are stable in solvent may be unstable in isolation in the gas phase[78-80]. Thus, we compare RHF/STO-3G and RHF/STO-3G+U (U_O = -6 eV, U_N = 6 eV) optimized structures directly to RHF/6-31G* results. In RHF/STO-3G optimizations, 40 PT events occur across the 9 proteins and 1 proton transfer event is observed in the control protein. For two proteins (2ONW, 2OL9), there is nearly a 1:1 ratio between protein residue count and proton transfer events. Consistent with previous results[10], RHF/6-31G* geometry optimizations yield a single PT event in each of the 2E4E, 2FXZ, and 2RLJ proteins, for a total of two proton transfer events in the nine "high" PT proteins and one in the control protein. In the case of 2E4E and 2RLJ, the proton transfer that occurs leads to the neutralization of the zwitterionic N and C termini that are expected to be relatively unstable in the gas phase. In

the case of 2FXZ, Glu5 abstracts a hydrogen atom from the N terminus of Lys1. With RHF/STO-3G+U we observe nearly comparable PT event counts to the larger basis with a total of 4 PT events observed across all 10 proteins. This result is encouraging since our approach had only thus far been benchmarked on small molecules. For the three proteins that exhibit proton transfer at RHF/6-31G*, the same proton transfers are observed with RHF/STO-3G+U. In one protein (3FTK), there is an N-to-C terminus proton transfer with RHF/STO-3G+U that does not occur in the larger basis set geometry optimization. For comparison, we computed the effect of empirical dispersion[42] and gCP[37] corrections on the relative energetics of the STO-3G+U- and STO-3G-optimized 2ONW structure that has also previously been studied with the HF-3c[19] method. In both the cytosine tautomers used for the initial +U tuning and the 2ONW peptide structures, differences in dispersion corrections are small, suggesting that the +U correction is compatible with and complementary to dispersion corrections (Supporting Information Tables S7-S8). The BSSE-focused gCP correction reduces errors in STO-3G tautomer energetics by up to 25%, which is insufficient to reproduce qualitative ordering. Re-optimization of the +U correction parameters is likely beneficial if both gCP and +U corrections are used simultaneously.

A careful examination of the protein structures optimized with RHF/STO-3G reveals two types of structural anomalies in addition to amide backbone proton transfer (Fig. 10). With RHF/STO-3G, proton transfer occurs between sidechains in disagreement with known pKa values for those sidechains, including dehydrogenation of asparagine and glutamine primary amines. Even in cases where no proton transfer occurs, there is a pyramidalization of the amide nitrogen on RHF/STO-3G Asn and Gln sidechains that is absent when a +U correction is applied or a larger basis set is employed (Fig. 10). Amide bonds are formally a resonance between a

neutral N/O species in which the nitrogen is sp^3 hybridized and a charge-separated N^+/O^- species with a C=N double bond. The non-planarity of Asn and Gln sidechains suggests charge separation via delocalization and resonance is problematic for minimal basis sets to describe, even in formally neutral molecular fragments. In the standard STO-3G calculations, frequent backbone cyclization also occurs when carbonyl and α-carbon atoms form covalent bonds to dehydrogenated backbone nitrogen atoms. We have not applied any corrections to the carbon atoms here, but this cyclization is still absent from RHF/STO-3G+U optimizations, in agreement with the larger basis RHF/6-31G* results.

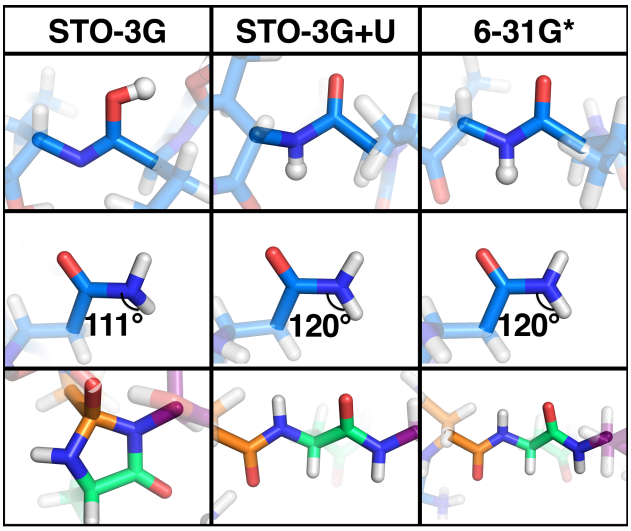


Figure 10. Prototypical optimized protein structure features in an example protein (PDB ID: 1YJP) obtained with RHF/STO-3G, RHF/STO-3G+U (U_O=-6 eV, U_N=6 eV), and RHF/6-31G*: amide backbone proton transfer (top), asparagine sidechain pyramidalization (middle), and backbone cyclization (bottom) are shown.

We note that the choice of parameterization for the RHF/STO-3G+U method of U_O=-6 eV, U_N=6 eV was based not just on energetics but also molecular orbital energy analysis and fractionality comparisons between differing chemical motifs. Furthermore, we find that parameters optimally tuned for trialanine peptide energetics are comparable to those obtained from the cytosine tautomer energetic error minimization (Supporting Information Table S9). The

more restrictive nature of same s- and p- exponent STO-3G minimal basis likely demands more of the +U correction than an alternatively more flexible basis set. Nevertheless, STO-3G was still chosen over MINI[S] due to the former's improved ability to produce geometries consistent with larger basis sets. Future efforts will be focused toward simultaneous minimal-basis optimization with +U-parameter reoptimization.

5. Conclusions

We have introduced a correction that reduces imbalances in minimal basis set descriptions of nitrogen and oxygen chemical bonding configurations observed in biological molecules. We have demonstrated the applicability of U_O and U_N parameters tuned to a single molecule, cytosine, to reproduce qualitative energetic ordering in a number of small molecules. These same parameters also prevent spurious proton transfer that is normally observed in minimal basis set geometry optimizations of proteins even when self-interaction-free Hartree-Fock is employed. We anticipate that such a set of U values is transferable to other applications where hydroxyl oxygen atoms and imines are overstabilized with respect to carbonyl oxygen and amines. Other minimal basis sets should have comparable, if not quantitatively identical, "ideal" U values because they exhibit comparable levels of spurious proton transfer[10]. However, we note that changing the basis set more significantly (e.g. to a minimal double-ζ basis set) or employing self-interaction-contaminated DFT will likely require more significant tuning of U parameters. We motivated the use of this correction in the context of Hartree-Fock by demonstrating how the approach tunes molecular orbital energies of Hubbard atoms and distinguishes chemically unique substituents through relative fractionality of the local occupation matrix. Such an approach may be straightforwardly applied to treat other known imbalances in electronic structure methods and basis sets in order to fix qualitative energetic ordering.

ASSOCIATED CONTENT

Supporting Information Available: Coordinates of RHF/STO-3G, RHF/STO-3G+U, and RHF/6-31G* optimized proteins; quantitative relative energetics of small molecules with U tuning; alanine-tuned parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*email: hjkulik@mit.edu phone: 617-253-4584

Notes

The authors declare the following competing financial interest(s): T.J.M. is a cofounder of PetaChem, LLC.

ACKNOWLEDGMENT

This work was supported by the Department of Defense (Office of the Director of Defense Research and Engineering) through a National Security Science and Engineering Faculty Fellowship (to T.J.M). H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, which supported both H.J.K and N.S..

REFERENCES

1. Ufimtsev, I. S.; Martínez, T. J., Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation. *J. Chem. Theory Comput.* **2008,** *4*, 222-231.
2. Ufimtsev, I. S.; Martínez, T. J., Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **2009,** *5*, 2619-2628.
3. Ufimtsev, I. S.; Martínez, T. J., Quantum Chemistry on Graphical Processing Units. 2. Direct Self-Consistent-Field Implementation. *J. Chem. Theory Comput.* **2009,** *5*, 1004-1015.

4. Guerra, C. F.; Snijders, J.; Te Velde, G.; Baerends, E., Towards an order-N DFT method. *Theor. Chem. Acc.* **1998,** *99*, 391-403.
5. Challacombe, M.; Schwegler, E., Linear scaling computation of the Fock matrix. *J. Chem. Phys.* **1997,** *106*, 5526-5536.
6. Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M., Fragment molecular orbital method: an approximate computational method for large molecules. *Chem. Phys. Lett.* **1999,** *313*, 701-706.
7. Merz Jr, K. M., Using quantum mechanical approaches to study biological systems. *Acc. Chem. Res.* **2014,** *47*, 2804-2811.
8. Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K., Fragment molecular orbital method: application to polypeptides. *Chem. Phys. Lett.* **2000,** *318*, 614-618.
9. Cole, D.; Skylaris, C.-K.; Rajendra, E.; Venkitaraman, A.; Payne, M., Protein-protein interactions from linear-scaling first-principles quantum-mechanical calculations. *Europhys. Lett.* **2010,** *91*, 37004.
10. Kulik, H. J.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J., Ab Initio Quantum Chemistry for Protein Structures. *J. Phys. Chem. B* **2012,** *116*, 12501-12509.
11. Cole, D. J.; O'Regan, D. D.; Payne, M. C., Ligand discrimination in myoglobin from linear-scaling DFT+ U. *J. Phys. Chem. Lett.* **2012,** *3*, 1448-1452.
12. Gaus, M.; Cui, Q.; Elstner, M., DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J. Chem. Theory. Comput.* **2011,** *7*, 931-948.
13. Stewart, J. J., Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013,** *19*, 1-32.
14. Dral, P. O.; Wu, X.; Spoerkel, L.; Koslowski, A.; Thiel, W., Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Benchmarks for Ground-State Properties. *J. Chem. Theory Comput.* **2016,** *12*, 1097-1120.
15. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995,** *117*, 5179-5197.
16. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct., Funct., Bioinf.* **2006,** *65*, 712-725.
17. Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J., Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **2014,** *6*, 1044-1048.
18. Xie, L.; Zhao, Q.; Jensen, K. F.; Kulik, H. J., Direct Observation of Early-Stage Quantum Dot Growth Mechanisms with High-Temperature Ab Initio Molecular Dynamics. *J. Phys. Chem. C* **2016,** *120*, 2472-2483.
19. Sure, R.; Grimme, S., Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013,** *34*, 1672-1685.
20. Liu, B.; McLean, A., Accurate calculation of the attractive interaction of two ground state helium atoms. *J. Chem. Phys.* **1973,** *59*, 4557-4558.
21. Boys, S. F.; Bernardi, F., The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970,** *19*, 553-566.

22. Mayer, I.; Turi, L., An analytical investigation into the bsse problem. *J. Mol. Struct.: THEOCHEM* **1991,** *227*, 43-65.
23. Cook, D.; Sordo, J.; Sordo, T., Some comments on the counterpoise correction for the basis set superposition error at the correlated level. *Int. J. Quantum Chem.* **1993,** *48*, 375-384.
24. Gutowski, M.; Chal/asiński, G., Critical evaluation of some computational approaches to the problem of basis set superposition error. *J. Chem. Phys.* **1993,** *98*, 5540-5554.
25. Van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G.; van Lenthe, J. H., State of the art in counterpoise theory. *Chem. Rev.* **1994,** *94*, 1873-1885.
26. Mentel, Ł.; Baerends, E., Can the counterpoise correction for basis set superposition effect be justified? *J. Chem. Theory Comput.* **2013,** *10*, 252-267.
27. Kristensen, K.; Ettenhuber, P.; Eriksen, J. J.; Jensen, F.; Jørgensen, P., The same number of optimized parameters scheme for determining intermolecular interaction energies. *J. Chem. Phys.* **2015,** *142*, 114116.
28. Moran, D.; Simmonett, A. C.; Leach, F. E.; Allen, W. D.; Schleyer, P. v. R.; Schaefer, H. F., Popular theoretical methods predict benzene and arenes to be nonplanar. *J. Am. Chem. Soc.* **2006,** *128*, 9342-9343.
29. van Mourik, T.; Karamertzanis, P. G.; Price, S. L., Molecular conformations and relative stabilities can be as demanding of the electronic structure method as intermolecular calculations. *J. Phys. Chem. A* **2006,** *110*, 8-12.
30. Holroyd, L. F.; van Mourik, T., Insufficient description of dispersion in B3LYP and large basis set superposition errors in MP2 calculations can hide peptide conformers. *Chem. Phys. Lett.* **2007,** *442*, 42-46.
31. Valdés, H.; Klusák, V.; Pitoňák, M.; Exner, O.; Starý, I.; Hobza, P.; Rulíšek, L., Evaluation of the intramolecular basis set superposition error in the calculations of larger molecules:[n] helicenes and Phe‐Gly‐Phe tripeptide. *J. Comput. Chem.* **2008,** *29*, 861-870.
32. Balabin, R. M., Communications: Intramolecular basis set superposition error as a measure of basis set incompleteness: Can one reach the basis set limit without extrapolation? *J. Chem. Phys.* **2010,** *132*, 211103.
33. Mayer, I., Bond orders and valences from ab initio wave functions. *Int. J. Quantum Chem.* **1986,** *29*, 477-483.
34. Galano, A.; Alvarez–Idaboy, J. R., A new approach to counterpoise correction to BSSE. *J. Comput. Chem.* **2006,** *27*, 1203-1210.
35. Jensen, F., Describing anions by density functional theory: fractional electron affinity. *J. Chem. Theory Comput.* **2010,** *6*, 2726-2735.
36. Faver, J. C.; Zheng, Z.; Merz Jr, K. M., Model for the fast estimation of basis set superposition error in biomolecular systems. *J. Chem. Phys.* **2011,** *135*, 144110.
37. Kruse, H.; Grimme, S., A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems. *J. Chem. Phys.* **2012,** *136*, 154101.
38. Davidson, E. R.; Feller, D., Basis set selection for molecular calculations. *Chem. Rev.* **1986,** *86*, 681-696.
39. Chen, V. B.; Arendall III, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: All-Atom Structure Validation for Macromolecular Crystallography. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010,** *66*, 12-21.

40. Isborn, C. M.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J., Excited-State Electronic Structure with Configuration Interaction Singles and Tamm-Dancoff Time-Dependent Density Functional Theory on Graphical Processing Units. *J. Chem. Theory Comput.* **2011,** *7*, 1814-1823.
41. Ufimtsev, I. S.; Luehr, N.; Martinez, T. J., Charge Transfer and Polarization in Solvated Proteins from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2011,** *2*, 1789-1793.
42. Grimme, S., Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006,** *27*, 1787-1799.
43. Huzinaga, S.; Andzelm, J.; Klobukowski, M.; Radzio-Andzelm, E.; Sakai, Y.; Tatewaki, H., *Gaussian Basis Sets for Molecular Calculations*. Elsevier: Amsterdam, 1984.
44. Anisimov, V. I.; Gunnarsson, O., Density-Functional Calculation Of Effective Coulomb Interactions In Metals. *Phys. Rev. B* **1991,** *43*, 7570-7574.
45. Anisimov, V. I.; Zaanen, J.; Andersen, O. K., Band Theory And Mott Insulators - Hubbard-U Instead Of Stoner-I. *Phys. Rev. B* **1991,** *44*, 943-954.
46. Liechtenstein, A. I.; Anisimov, V. I.; Zaanen, J., Density-Functional Theory And Strong-Interactions - Orbital Ordering In Mott-Hubbard Insulators. *Phys. Rev. B* **1995,** *52*, R5467-R5470.
47. Anisimov, V. I.; Aryasetiawan, F.; Lichtenstein, A. I., First-principles calculations of the electronic structure and spectra of strongly correlated systems: The LDA+U method. *J. Phys.: Condens. Matter* **1997,** *9*, 767-808.
48. Jiang, L.; Levchenko, S.; Rappe, A., Rigorous Definition of Oxidation States of Ions in Solids. *Phys. Rev. Lett.* **2012,** *108*, 166403.
49. Koopmans, T., Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms. *Physica* **1934,** *1*, 104-113.
50. Janak, J. F., Proof that dE/dni=epsilon in density-functional theory. *Phys. Rev. B* **1978,** *18*, 7165-7168.
51. Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, A. P., Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Phys. Rev. B* **1998,** *57*, 1505-1509.
52. Cococcioni, M.; de Gironcoli, S., Linear response approach to the calculation of the effective interaction parameters in the LDA+U method. *Phys. Rev. B* **2005,** *71*, 035105.
53. Kulik, H. J.; Cococcioni, M.; Scherlis, D. A.; Marzari, N., Density functional theory in transition-metal chemistry: A self-consistent Hubbard U approach. *Phys. Rev. Lett.* **2006,** *97*, 103001.
54. Kulik, H. J.; Marzari, N., Systematic study of first-row transition-metal diatomic molecules: A self-consistent DFT plus U approach. *J. Chem. Phys.* **2010,** *133*, 114103.
55. Kulik, H. J.; Marzari, N., Accurate potential energy surfaces with a DFT+U(R) approach. *J. Chem. Phys.* **2011,** *135*, 194105.
56. Mosey, N.; Carter, E., Ab initio evaluation of Coulomb and exchange parameters for DFT+U calculations. *Phys. Rev. B* **2007,** *76*, 155123.
57. Wang, L.; Maxisch, T.; Ceder, G., Oxidation energies of transition metal oxides within the GGA+U framework. *Phys. Rev. B* **2006,** *73*, 195107.
58. Loschen, C.; Carrasco, J.; Neyman, K.; Illas, F., First-principles LDA+U and GGA+U study of cerium oxides: Dependence on the effective U parameter. *Phys. Rev. B* **2007,** *75*, 035115.
59. Huang, M.; Fabris, S., CO Adsorption and Oxidation on Ceria Surfaces from DFT+U Calculations. *J. Phys. Chem. C* **2008,** *112*, 8643-8648.

60. Kulik, H. J., Perspective: Treating electron over-delocalization with the DFT+U method. *J. Chem. Phys.* **2015,** *142*, 240901.

61. Löwdin, P. O., On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *J. Chem. Phys.* **1950,** *18*, 365-375.

62. Petachem. http://www.petachem.com. (accessed May 4, 2016).

63. Hehre, W. J.; Stewart, R. F.; Pople, J. A., Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *J. Chem. Phys.* **1969,** *51*, 2657-2664.

64. Ditchfield, R.; Hehre, W. J.; Pople, J. A., Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1971,** *54*, 724-728.

65. Hariharan, P. C.; Pople, J. A., The influence of polarization functions on molecular orbital hydrogenation energies. *Theoret. Chim. Acta* **1973,** *28*, 213-222.

66. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28*, 235-242.

67. H++ webserver. http://biophysics.cs.vt.edu/H++ (accessed May 4, 2016).

68. Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A., H++: A Server for Estimating pKas and Adding Missing Hydrogens to Macromolecules. *Nucleic Acids Res.* **2005,** *33*, W368-W371.

69. Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V., H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012,** *40*, W537-W541.

70. Hehre, W. J.; Stewart, R. F.; Pople, J. A., Self-Consistent Molecular Orbital Methods. 1. Use of Gaussian expansions of Slater-type atomic orbitals. *J. Chem. Phys.* **1969,** *51*, 2657-2664.

71. Remko, M.; Scheiner, S., Abinitio Investigation of Interactions between Models of Local-Anesthetics and Receptor - Complexes Involving Amine, Phosphate, Amide, Na+, K+, Ca-2+, and Cl-. *J. Pharm. Sci.* **1988,** *77*, 304-308.

72. Fogarasi, G., Relative Stabilities of Three Low-Energy Tautomers of Cytosine: A Coupled Cluster Electron Correlation Study. *J. Phys. Chem. A* **2002,** *106*, 1381-1390.

73. Fonseca Guerra, C.; Handgraaf, J.-W.; Baerends, E. J.; Bickelhaupt, F. M., Voronoi deformation density (VDD) charges: Assessment of the Mulliken, Bader, Hirshfeld, Weinhold, and VDD methods for charge analysis. *J. Comput. Chem.* **2004,** *25*, 189-210.

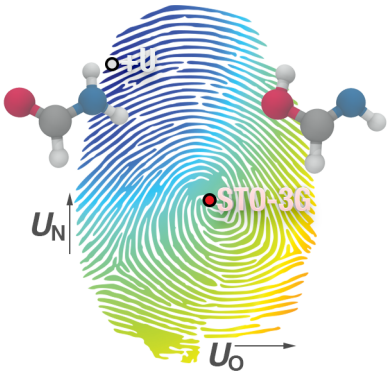
74. Echols, N.; Harrison, P.; Balasubramanian, S.; Luscombe, N. M.; Bertone, P.; Zhang, Z. L.; Gerstein, M., Comprehensive Analysis of Amino Acid and Nucleotide Composition in Eukaryotic Genomes, Comparing Genes and Pseudogenes. *Nucleic Acids Res.* **2002,** *30*, 2515-2523.

75. Liu, F.; Luehr, N.; Kulik, H. J.; Martínez, T. J., Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. *J. Chem. Theory Comput.* **2015,** *11*, 3131-3144.

76. Nochebuena, J.; Ireta, J., On cooperative effects and aggregation of GNNQQNY and NNQQNY peptides. *J. Chem. Phys.* **2015,** *143*, 135103.

77. Rudberg, E., Difficulties in applying pure Kohn–Sham density functional theory electronic structure methods to protein molecules. *J. Phys.: Condens. Matter* **2012,** *24*, 072202.

78. Rodgers, M. T.; Campbell, S.; Marzluff, E. M.; Beauchamp, J. L., Site-Specific Protonation Directs Low-Energy Dissociation Pathways of Dinucleotides in the Gas-Phase. *Int. J. Mass Spectrom. Ion Processes* **1995,** *148*, 1-23.
79. Cerda, B. A.; Wesdemiotis, C., Zwitterionic Vs. Charge-Solvated Structures in the Binding of Arginine to Alkali Metal Ions in the Gas Phase. *Analyst* **2000,** *125*, 657-660.
80. Patriksson, A.; Marklund, E.; Van Der Spoel, D., Protein Structures under Electrospray Conditions. *Biochemistry* **2007,** *46*, 933-945.


U_N
U_O