First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids

Matti Ropo[1,2,3*], Markus Schneider[1], Carsten Baldauf[1*],
Volker Blum[1,4*]

January 14, 2016

1. Fritz Haber Institute of the Max Planck Society, 14195 Berlin, Germany;
2. Department of Physics, Tampere University of Technology, 33720 Tampere,
Finland; 3. COMP, Department of Applied Physics, Aalto University, 00076
Aalto, Finland; 4. Department of Mechanical Engineering and Materials Science,
Duke University, Durham, NC 27708, USA.
*corresponding authors: Matti Ropo (matti.ropo@tut.fi); Carsten Baldauf
(baldauf@fhi-berlin.mpg.de); Volker Blum (volker.blum@duke.edu).

Abstract

We present a structural data set of the 20 proteinogenic amino acids
and their amino-methylated and acetylated (capped) dipeptides. Differ-
ent protonation states of the backbone (uncharged and zwitterionic) were
considered for the amino acids as well as varied side chain protonation
states. Furthermore, we studied amino acids and dipeptides in complex
with divalent cations (Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+}). The
database covers the conformational hierarchies of 280 systems in a wide
relative energy range of up to $4\,eV$ ($390\,kJ/mol$), summing up to an overall
of 45,892 stationary points on the respective potential-energy surfaces. All
systems were calculated on equal first-principles footing, applying density-
functional theory in the generalized gradient approximation corrected for
long-range van der Waals interactions. We show good agreement to avail-
able experimental data for gas-phase ion affinities. Our curated data can
be utilized, for example, for a wide comparison across chemical space of
the building blocks of life, for the parametrization of protein force fields,
and for the calculation of reference spectra for biophysical applications.

Background & Summary

Proteins are the machinery of life. We here present a first-principles study of the conformational preferences of their basic building blocks – specifically, as summarized in Figure 1: 20 proteinogenic amino acids and dipeptides, with different possible protonation states, and the conformational space changes resulting from attaching six divalent cations, i.e., Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+}. In past studies, a wide range of different approximate electronic structure methods has been applied to some of these proteinogenic amino acids – see, for example, references [1–59]. These studies have deepened our understanding of the conformational basics of individual building blocks, but a systematic comparison of properties of the different building blocks is complicated when relying on data from different sources. On the one hand this is due to the molecular models that may differ in protonation states and backbone capping. On the other, the simulations can differ in several ways:

- Different sampling strategies or methods to generate conformers may have been used. Search-dependent settings, like energy cut-offs, can also have a significant impact on the results.

- The levels of theory that have been applied range from semi-empirical to Hartree-Fock (HF) to density-functional theory (DFT) up to coupled-cluster calculations [1–59].

- Numerical settings, e.g., basis sets, can differ substantially and might lead to different results.

A further point that limits a quantitative comparison is the accessibility of the data from different studies. Energies, for example, often have to be extracted from table footnotes and/or the structural data is not always accessible in the Supplementary Information of the respective articles, sometimes even only accessible as figures in the manuscript. The data set presented here overcomes such limitations by covering a comprehensive segment of chemical space exhaustively, using a large scale computational effort. This study treats 20 proteinogenic amino acids, their dipeptides and their interactions with the divalent cations Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, and Hg^{2+} (see Figure 1 for an overview) on the same theoretical footing. The importance of peptide cation interactions may be highlighted by the fact that about 40% of all proteins bind cations [60–62]. Especially Ca^{2+} is important in a multitude of functions, ranging, for example, from blood clotting [63] to cell signaling to bone growth [64]. Such calcium mediated functions can be disturbed by the presence of alternative divalent heavy metal cations like Pb^{2+}, Cd^{2+}, and Hg^{2+} [62,65,66].

The conformations and total energies of each molecular system are calculated from first principles in the framework of density-functional theory (DFT) [67,68] using the PBE generalized-gradient exchange-correlation functional [69]. Energies are corrected for van der Waals interactions using the Tkatchenko-Scheffler formalism [70]. In this formalism, pairwise $C_6[n]/r^6$ terms are computed and

summed up for all pairs of atoms. r is the interatomic distance, a cut-off for short interatomic distances is applied, and $C_6[n]$ coefficients are obtained from the self-consistent electron density. The combined approach is referred to as "PBE+vdW" throughout this work. This level of theory is robust for potential-energy surface (PES) sampling of peptide systems [71–78]. The curated data is provided as basis for comparative studies across chemical space to reveal conformational trends and energetic preferences. It can, for example, further be used for force-field development, theoretical studies at higher levels of theory, and as a starting point for theoretical calculations of spectra for biophysical applications.

Methods

Molecular models

This study covers a total of 280 molecular systems (summarized in Figure 1). The number is the product of these chemical degrees of freedom that were considered in our study:

20 proteinogenic amino acids. In case of (de)protonatable side chains, all protomers (different protonations states) were considered as well.

2 different backbone types, either free termini (considered in uncharged or zwitterionic form) or capped (N-terminally acetylated or C-terminally amino-methylated).

7 reflecting that the respective amino acid or dipeptide was considered either in isolation or with one of six different cation additions: Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, or Hg^{2+}.

Conformational search and energy functions

For the initial scan of the PES, the empirical force field OPLS-AA [79] was employed, followed by DFT-PBE+vdW relaxations of the energy minima identified in the force field. The identified set of structures was then subjected to a further first-principles refinement step, *ab initio* replica-exchange molecular dynamics (REMD). An overview of the procedure is given in Figure 2 and the steps are described in more detail below.

Force-field based (OPLS-AA) [79] **global conformational searches (Step 1)** were performed for all dipeptides and amino acids (i) without a coordinating cation and (ii) with Ca^{2+}. These searches employed a basin hopping search strategy [80, 81] as implemented in the tool "scan", distributed with the TINKER molecular simulation package [82, 83]. We here use an in-house parallelized version of the TINKER scan utility that was first used in reference [74]. In this search strategy, input structures for relaxations are generated by projecting along normal modes starting from a local minimum. The number of search directions from a local minimum was set to 20. Conformers were accepted within

a relative energy window of 100 kcal/mol and if they differ in energy from already found minima by at least 10^{-4} kcal/mol. The search terminates when the relaxations of input structures do not result in new minima.

After that, **PBE+vdW relaxations (Step 2)** were performed with the program FHI-aims [84–86]. FHI-aims employs numeric atom-centered orbital basis sets as described in reference [84] to discretize the Kohn-Sham orbitals. Different levels of computational defaults are available, distinguished by choice of the basis set, integration grids, and the order of the multipole expansion of the electrostatic (Hartree) potential of the electron density. For the chemical elements relevant to this work, "light" settings include the so-called *tier1* basis sets and were used for initial relaxations. "Tight" settings include the larger *tier2* basis sets and ensure converged conformational energy differences at a level of few meV [84]. Unless noted otherwise, all energies discussed here are results of PBE+vdW calculations with a *tier2* basis and "tight" settings. Relativistic effects were taken into account by the so-called atomic zero-order regular approximation (atomic ZORA) [87, 88] as described in reference [84]. Previous comparisons to high-level quantum chemistry benchmark calculations at the coupled-cluster level, CCSD(T), demonstrated the reliability of this approach for polyalanine systems [72, 76], alanine, phenylalanine, and glycine containing tripeptides [76], and alanine dipeptides with Li^+ [73]. Further benchmarks at the MP2 level of theory are reported below in the section Technical Validation.

The **refinement (Step 3)** by *ab initio* REMD [89, 90] is intended to alleviate the potential effects of conformational energy landscape differences between the force field and the DFT method. In REMD, multiple molecular dynamics trajectories of the same system are independently initialized and run in a range of different temperatures. Based on a Metropolis criterion, configurations are swapped between trajectories of neighboring temperatures. Thus, the simulations can overcome barriers and provide an enhanced conformational sampling in comparison to classical molecular dynamics (MD) [90, 91]. The simulations were carried out employing a script-based REMD scheme that is provided with FHI-aims and that was first used in reference [92]. Computations were performed at the PBE+vdW level with "light" computational settings. The run time for each REMD simulation was 20 ps with an integration time step of 1 fs. The frequent exchange attempts (every 0.04 or 0.1 ps) ensure efficient sampling of the potential-energy surface as shown by Sindhikara *et al.* [93]. The velocity-rescaling approach by Bussi *et al.* [94] was used to sample the canonical distribution. Starting geometries for the replicas were taken from the lowest energy conformers resulting from the PBE+vdW relaxations in Step 2. REMD parameters for the individual systems, i.e. the number of replicas, acceptance rates for exchanges between replicas, the frequency for exchange attempts, and the temperature range, are summarized in table S1 of the Supplementary Information. Conformations were extracted from the REMD trajectories every 10th step, i.e. every 10 fs of simulation time. In order to generate a set of representative conformers, these structures were clustered using a k-means clustering algorithm [95] with a cluster radius of 0.3 Å as provided by the MMSTB package [96]. The resulting arithmetic-mean structures from each cluster were then

relaxed using PBE+vdW with "light" computational settings. The obtained conformers were again clustered and cluster representatives were relaxed with PBE+vdW ("tight" computational settings) to obtain the final conformation hierarchies. The refinement step by REMD is essential, as shown in Figure 3, which separately identifies the number of distinct conformers found in Step 2 and, subsequently, the number of additional conformers found in Step 3.

After step 2, a total of 17,381 stationary points was found for the amino acids and dipeptides in isolation and in complex with Ca^{2+}. The refinement procedure in Step 3 increases this number to a total of 21,259 structures. Initial structures for the Ba^{2+}, Cd^{2+}, Hg^{2+}, Pb^{2+} and Sr^{2+} binding amino acid and dipeptide systems were then obtained by replacing the Ca^{2+} cation in the amino acid and dipeptide structures binding a Ca^{2+} cation. These structures were subsequently relaxed with PBE+vdW employing "tight" computational settings and a tier-2 basis set. This procedure results in 24,633 further conformers with bound cations. Altogether, we thus provide information on 45,892 stationary points of the PBE+vdW PES for all systems studied in this work.

The numbers of conformers identified in the searches are also given in Table S2 of the Supplementary Information. Tables S3 and S4 provide detailed accounts of how many structures were found for which amino acid/dipeptide in isolation or with attached cations.

Data Records

The curated data, consisting of the Cartesian coordinates of 45,892 stationary point geometries of the PBE+vdW PES (the main outcome of our work) and their potential energies computed at the "tight"/tier-2 level of accuracy in the FHI-aims code, is provided as plain text files sorted in directories (see Figure 4). The PBE+vdW total energies are included since they are an integral part of the construction of our geometry data sets. Importantly, the stationary point geometries could be used as starting points to refine the total energy accuracy by higher-level methods, e.g., those discussed in "Technical Validation" below. The folder structure is hierarchic and straightforward. The naming scheme is explained in the following:

Description of the file types:

conformer.(...).xyz coordinates in standard xyz format in Å, readable by a wide range of molecule viewers, e.g. VMD, Jmol, etc.

conformer.(...).fhiaims coordinate file in FHI-aims geometry input format: for each atom of the particular system, the Cartesian coordinates are given in Å (`atom [x] [y] [z] [element]`). The electronic total energy (in eV) at the PBE+vdW level is given there as a comment.

control.in FHI-aims input file with technical parameters for the calculations. Please note that these files also include the exact specifications of the "tight" numerical settings for all included elements.

5

hierarchy_PBE+vdW_tier-2.dat in each final subfolder, contains three columns: number of the conformer, total energy (in eV, PBE+vdW, tier-2 basis set, "tight" numerical settings, computed with FHI-aims version 031011), and relative energy (in eV, relative to the respective global minimum).

The curated data is publicly available from two sources:

1. A website dedicated to this data set has been set up[1] and allows users to browse and download the data and to visualize molecular structures online.

2. From the NOMAD repository[2] the data is available via the DOI 10.17172/NO-MAD/20150526220502[3] [Data citation 1].

Technical Validation

The conformational coverage for the amino acid alanine is validated by comparing to a recent study by Maul *et al.* [12] . In that reference, 10 low energy conformers of alanine were reported, spanning an energy range of approximately 0.26 eV between the reported lowest and highest energy conformers. The level of theory used by Maul *et al.* was DFT in the generalized gradient approximation by means of the Perdew-Wang 1991 functional [97]. In our case, the force field based search step with subsequent PBE+vdW relaxations yields 5 conformers. The following *ab initio* REMD simulations increase the number of conformers to 15 within an energy range of 0.43 eV. The respective conformational energy hierarchies after global search and after REMD-refinement are shown in Figure 5A. The results of our search (with the refinement step) are in good agreement with the data from reference [12] that is also shown in Figure 5A. Structures are shown in Figure 5B. Nine of the ten conformers identified by Maul *et al.* can be confirmed. The single conformer that is missing (highlighted by an X in Figure 5A) is not a stationary point of the PBE+vdW potential energy surface. Conformers 14 and 15 are classified as saddle points by analysis of the vibrational modes.

In order to further quantify the reliability of the DFT-PBE+vdW level of theory for peptides, beyond earlier benchmark work [72, 73, 76] and especially with divalent cations, benchmark calculations were performed at the level of Møller-Plesset second-order perturbation theory (MP2) [98, 99] using the electronic structure program package ORCA [100]. Single-point energy calculations were performed for all fixed stationary-point DFT-PBE+vdW geometries in our data base for the amino acids alanine (Ala) and phenylalanine (Phe) with neutral N and C termini in isolation as well as in complex with a Ca^{2+} cation. Phe was selected to represent a "difficult" example, i.e., the interaction of the

---

cation with a larger aromatic side chain. The MP2 calculations did not include any frozen-core treatment (including semicore states is essential for Ca^{2+}) and were performed using Dunning's correlation-consistent polarized core-valence basis sets (cc-pCVnZ), with n=T/Q/5 denoting the triple-zeta, quadruple-zeta, and quintuple-zeta basis sets respectively [101]. The calculated SCF (Hartree-Fock) and MP2 correlation energies were then individually extrapolated to the complete basis set (CBS) limit as follows: For SCF energies, we used the extrapolation strategy proposed by Karton and Martin [102]:

$$E_{SCF}^n = E_{SCF}^{CBS} + A e^{-\alpha\sqrt{n}}. \tag{1}$$

A, α, and the CBS-extrapolated energy E_{SCF}^{CBS} are parameters determined from a least-squares fitting algorithm applied individually for each conformer. For the MP2 correlation energies, an extrapolation scheme proposed by Truhlar [103] was applied:

$$E_{corr}^n = E_{corr}^{CBS} + B n^{-\beta}. \tag{2}$$

Again, B, β, and the CBS-extrapolated energy E_{corr}^{CBS} are parameters determined from a least-squares fitting algorithm as before. A detailed account of all numbers is given in the Supplementary Information (Table S5). Mean absolute errors between the density-functional approximation (DFA) relative energies and the basis-set extrapolated MP2 relative energies were calculated as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\Delta E_i^{DFA} - \Delta E_i^{MP2} + c|, \tag{3}$$

where the index i runs over all N conformations of a given data set. ΔE_i in principle denotes the energy difference between conformer i and the lowest-energy conformer of the set. The adjustable parameter c is used to shift the MP2 and DFA conformational hierarchies versus one another to obtain the lowest possible MAE, rendering the reported MAE value independent of the choice of any reference structure. Figure 6A shows the corresponding obtained mean absolute errors (MAE) and maximal errors ($max_i|\Delta E_i^{DFA} - \Delta E_i^{MP2} + c|$) of different DFA calculations – performed with the FHI-aims code – with respect to benchmarks on the MP2 level obtained as described above. Within FHI-aims, the accuracy of integration grids and of the electrostatic potential was also verified by comparing "tight" and "really_tight" numerical settings, giving virtually identical results. The DFA level of theory of PBE+vdW shows a MAE well within *chemical accuracy* of $\sim 1\,\mathrm{kcal/mol} \approx 43\,\mathrm{meV}$ for both structural sets of Ala and Phe; for Phe, the maximal error is $\sim 2\,\mathrm{kcal/mol}$. We next applied a different long-range dispersion treatment, a recent many-body dispersion model based on interacting quantum harmonic oscillators denoted as MBD, [104] showing no significant improvement for the isolated amino acids. In line with Ref. [76], applying the more expensive PBE0 [105] hybrid exchange correlation functional reduces the maximum deviation for Phe to $\sim 57\,\mathrm{meV}$, i.e., $1.3\,\mathrm{kcal/mol}$. For Ala and Phe with

7

neutral end caps in complex with a Ca^{2+} cation, Figure 6B compares the same set of DFAs to MP2 benchmark energy hierarchies. However, obtaining basis-set converged total energies of the same accuracy as for the isolated peptides by straightforward CBS extrapolation proved remarkably more difficult when Ca^{2+} was involved. The reason is traced to the significant and slow-converging correlation contribution of the Ca^{2+} semicore electrons, which leads to large and conformation dependent basis set superposition errors (BSSE). This problem was verified for MP2 calculations in the FHI-aims and ORCA codes, with several different basis set prescriptions [106], and for CCSD(T) calculations. Standard DFAs, if sufficiently accurate, have a significant advantage in this respect since they are not subject to comparable numerical convergence problems. To yet arrive at reliable CBS-extrapolated MP2 conformational energy differences, we thus subjected the SCF and correlation energies of each Ca^{2+} coordinated conformation to a counterpoise correction [107,108] to minimize the effect of BSSE on the Ca^{2+} correlation energy contribution, prior to performing CBS extrapolation as described above. For the example of Ala+Ca^{2+} and assuming rigid conformers, the BSSE is estimated as:

$$E_{BSSE} = E_{BSSE}(Ala) + E_{BSSE}(Ca^{2+}) \text{ , with}$$
$$E_{BSSE}(Ala) = E^{Ala+Ca^{2+}}(Ala) - E^{Ala}(Ala) \text{ , and} \quad (4)$$
$$E_{BSSE}(Ca^{2+}) = E^{Ala+Ca^{2+}}(Ca^{2+}) - E^{Ca^{2+}}(Ca^{2+}).$$

$E^{Ala+Ca^{2+}}(Ala)$ represents the energy of Ala evaluated in the union of the basis sets on Ala and Ca^{2+}, $E^{Ala}(Ala)$ represents the energy of Ala evaluated in the basis set on Ala, *etc.* The individual BSSE errors are then subtracted from the SCF and correlation energy respectively. Phe+Ca^{2+} is treated equivalently. Complete numerical details are given in the Supplementary Information (Table S6). Following this procedure, the MAE and maximal error values of various DFAs compared to MP2 are well within $1\,kcal/mol$ for Ala+Ca^{2+}. The PBE+vdW MAE for Phe+Ca^{2+} amounts to just above $\sim 2\,kcal/mol$. The contributions from both the many-body dispersion and the hybrid PBE0 functional improve the MAE to just above $1\,kcal/mol$ at to PBE0+MBD* level of theory. The maximum errors in the energy hierarchies between individual conformers are correspondingly larger. Overall, this assessment shows that our data base of conformer geometries constitutes, e.g., an excellent starting point for more exhaustive future benchmark work of new electronic structure methods for cation-peptide systems. For example, it would be very interesting to explore how F12 approaches, which address the correlation energy convergence problem explicitly, fare for a broad range of different Ca^{2+} containing conformations of our peptides.

As a final validation, we compare the correlation of calculated gas-phase amino acid-Ca^{2+} binding energies to the binding energy hierarchy found experimentally in a study by Ho *et al.* [109]. We calculate binding energy at the PES level as

$$E_{binding} = E_{amino\ acid} + E_{cation} - E_{complex}. \quad (5)$$

Energies E denote the PBE+vdW Born-Oppenheimer potential energies, including $E_{amino\ acid}$ of the lowest-energy conformers of the isolated amino acid and $E_{complex}$ of the same amino acid in complex with a Ca^{2+} ion. Experimentally [109], the gas-phase Ca^{2+} affinities of 18 proteinogenic amino acids were determined by fragmenting Ca^{2+} complexes with a combinatoric library of tripeptides at $T \approx 330$ K, recording the mass spectrometric peak intensities of different fragmentation products. Quantitative average relative binding energies of Ca^{2+} to different amino acids were thus inferred and can be compared to our findings, albeit with several important experiment-theory differences: (i) Entropy effects [73, 75, 110] should affect the specific complexes probed experimentally but cannot be included into the calculated numbers in the exact same way, (ii) structural differences (e.g., protonation, dimerized amino acids) between the fragments recorded in experiment and the amino acids covered here, (iii) experimental Ca^{2+} affinities are not given for Asp and Glu because their gas-phase acidities, needed for data conversion, are not known. Figure 7 compares the experimentally and theoretically inferred Ca^{2+} binding affinities qualitatively. The x-axis reflects the experimental binding affinity energy hierarchy, arranging amino acids from left to right in order of decreasing binding affinity. The y axis shows calculated binding energies according to Eq. 5. Perfect correlation of the experimental and calculated hierarchies would imply a strictly monotonic decrease of calculated $E_{binding}$ values from left to right. This monotonic trend is not obeyed exactly; however, in view of the significant differences (i) and (ii) above, the qualitative agreement is quite striking. Normalized correlation coefficients between the experimental (1) and calculated (2) binding affinity data were calculated following the formula:

$$r_{12} = s_{12}/(s_1 s_2), \tag{6}$$

with s_{12} being the covariance of data sets and s_i being the standard deviations of data sets i=1,2. The result, correlation coefficients of r_{12}=0.979 or 0.909 for uncapped amino acids or dipeptides, respectively, also point to an overall remarkably good agreement. Finally, Figure 7 also gives predicted $E_{binding}$ values for protonated (overall system charge +2) and deprotonated (overall system charge +1) Asp and Glu, reflecting the significant electrostatic attraction between cations and negatively charged (deprotonated) Asp and Glu side chains. The binding energy data sets are included as Supplementary Table S5.

Usage Notes

The present data contains stationary-point geometries (mainly minima, but also saddle points since no routine normal-mode analysis was performed) on the potential energy surface of the 20 proteinogenic amino acids and dipeptides, either isolated or in complex with a divalent cation (Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, Hg^{2+}). The users of this dataset may find openbabel [111](www.openbabel.org) to be a useful tool to convert FHI-aims and xyz files to other common file formats in chemistry.

9

Author Contributions

MR performed the calculations to assemble all conformers. MR and CB curated the data. Validation calculations by DFAs and correlated methods other than PBE+vdW were carried out by MS. MR, CB, VB designed the study and wrote the data descriptor.

Competing financial interests

The authors declare no competing financial interests.
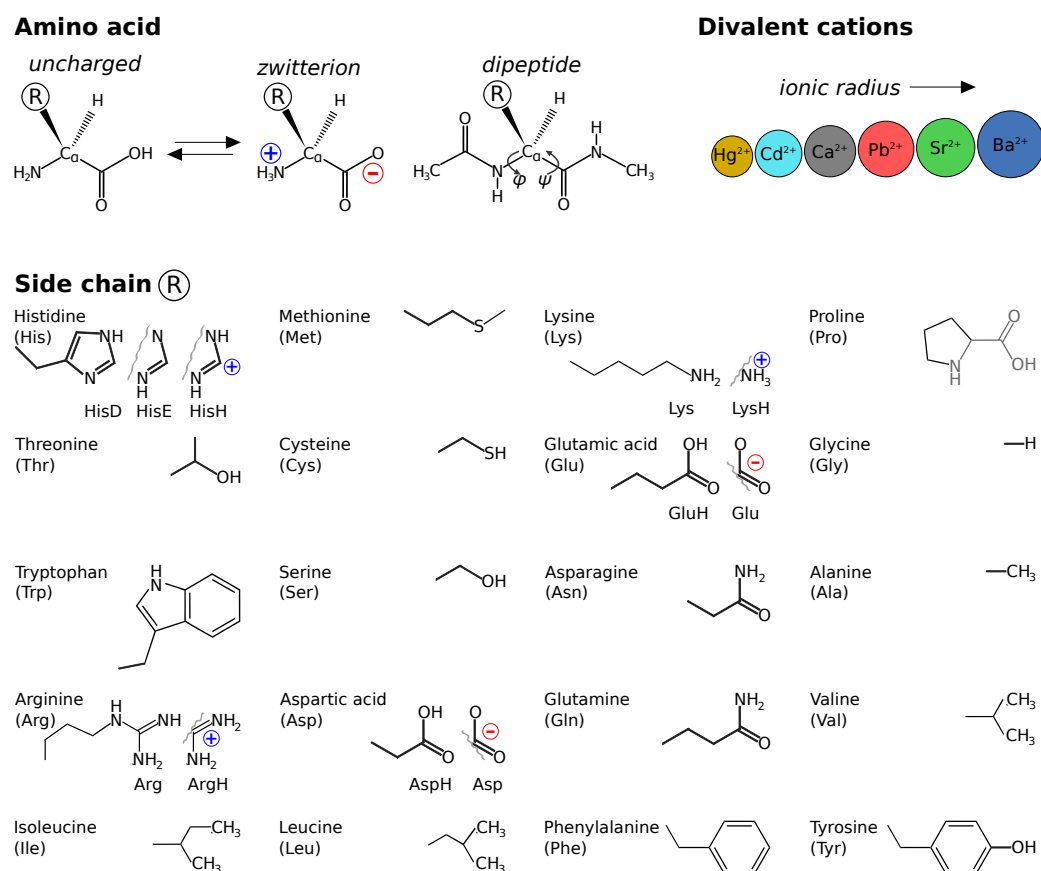
Figures and Legends

Figure 1: **Molecular systems covered in this study.** Top left and center: Schematic depiction of the backbone conformations of uncharged, zwitterionic, and dipeptide forms of the aminoacids considered in this work. Side chains are indicated by the letter **R**. Top right: Divalent ions considered for complexation with the 20 proteinogenic amino acids. Lower five rows: Side chains, including different protonation states where applicable, of the 20 proteinogenic amino acids considered in this work.
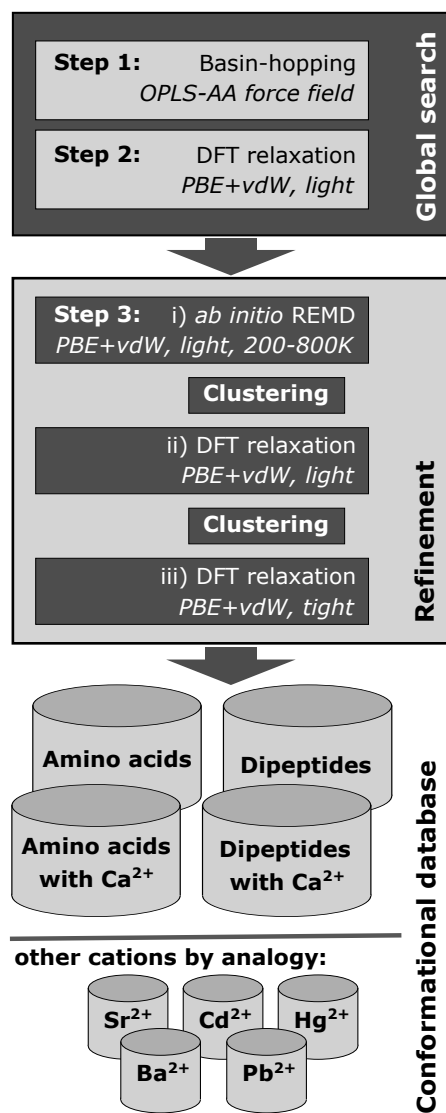
Figure 2: **Schematic representation of the workflow** employed to locate stationary points on the potential-energy surfaces of the respective molecular systems.
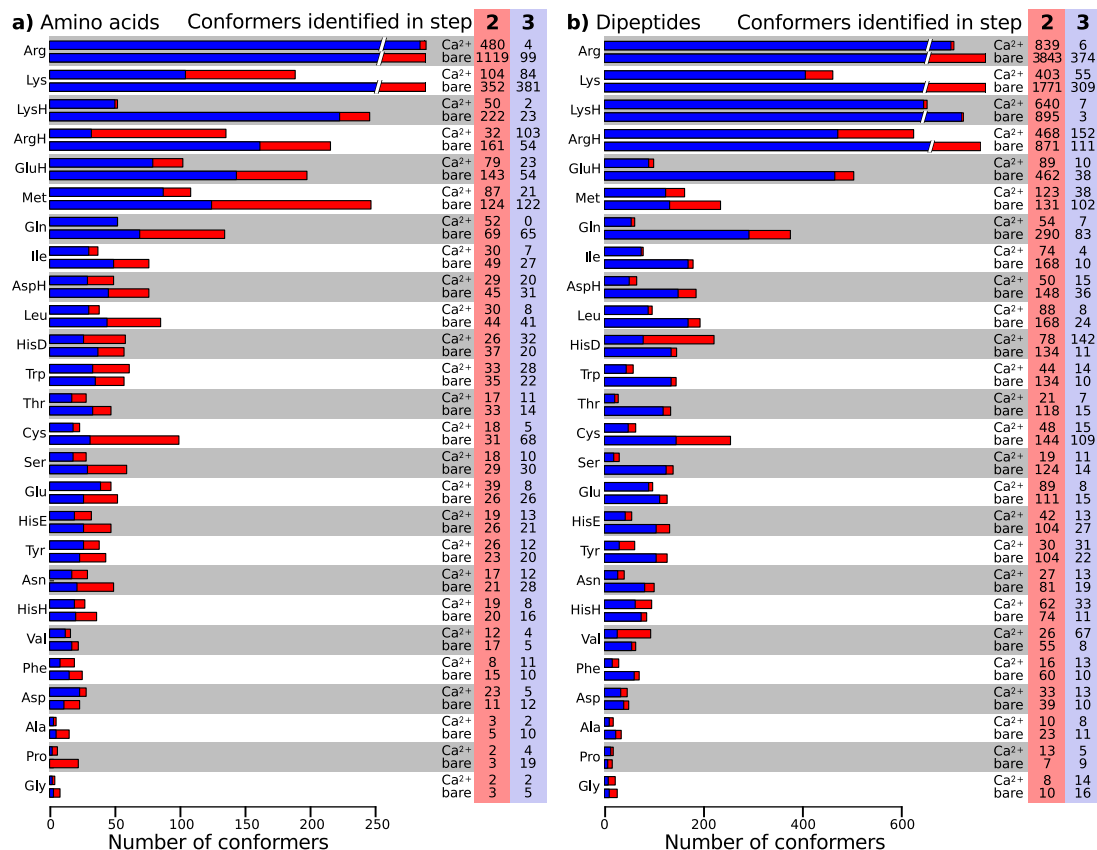
Figure 3: **Numbers of stationary points** of the PBE+vdW potential-energy surface (PES) at the "tight"/tier-2 level of accuracy that were found for the different **a)** uncapped amino acids or **b)** dipeptides in isolation ("bare") or with a Ca^{2+} cation. Blue segments of the bars and blue shaded numbers give the number of stationary points ("conformers") located in Step 2 of the search procedure detailed in Figure 2. Red bar segments and red shading highlight the number of conformers that were additionally found during Step 3 of the search. The total number of conformers found for each system is the sum of the numbers found in steps two and steps three.
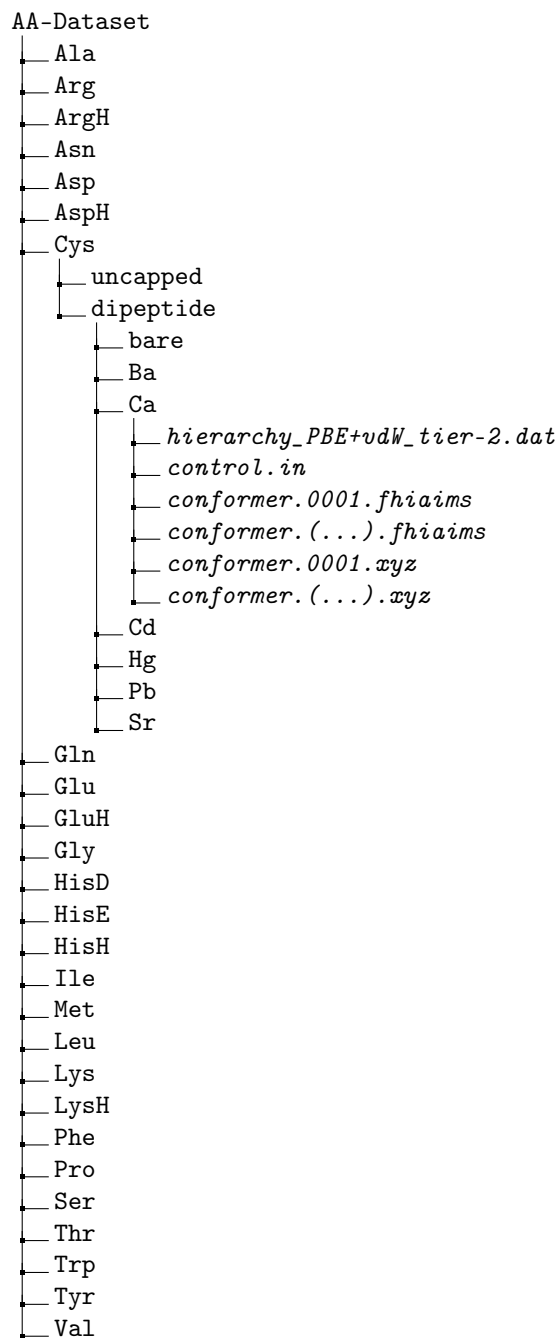
14

```
AA-Dataset
├── Ala
├── Arg
├── ArgH
├── Asn
├── Asp
├── AspH
├── Cys
│   ├── uncapped
│   └── dipeptide
│       ├── bare
│       ├── Ba
│       ├── Ca
│       │   ├── hierarchy_PBE+vdW_tier-2.dat
│       │   ├── control.in
│       │   ├── conformer.0001.fhiaims
│       │   ├── conformer.(...).fhiaims
│       │   ├── conformer.0001.xyz
│       │   └── conformer.(...).xyz
│       ├── Cd
│       ├── Hg
│       ├── Pb
│       └── Sr
├── Gln
├── Glu
├── GluH
├── Gly
├── HisD
├── HisE
├── HisH
├── Ile
├── Met
├── Leu
├── Lys
├── LysH
├── Phe
├── Pro
├── Ser
├── Thr
├── Trp
├── Tyr
└── Val
```

Figure 4: **Schematic representation of folder organization of the data.** Each folder, as exemplified for the Ca^{2+}-coordinated cysteine dipeptide, contains coordinate files in two formats (standard XYZ and FHI-aims input), the computational settings file for FHI-aims (control.in), and the energy hierarchies (PBE+vdW, "tight"/tier-2 level) per system.
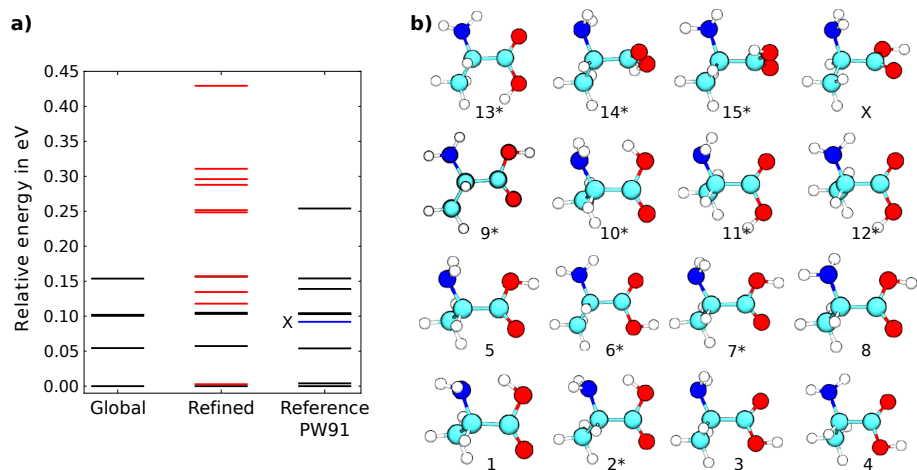
Figure 5: **Comparison of search strategies.** (**a**)) The conformational energy hierarchies for alanine after the global search and the local refinement together with the reference hierarchy at the DFT-PW91 level that was published by Maul *et al.* [12]. Conformers indicated by black lines were found in the global search, the conformers in red were located only after the local refinement step. The blue line in the reference conformational hierarchy represents a minimum not found in our search and not present at the PBE+vdW level. (**b**)) Conformations of the alanine molecule. Conformers marked with an asterisk (*) were found in the local refinement step of our search strategy. Atoms are color-coded as follows: Cyan (C), blue (N), red (O), white (H). The conformer labeled with X was found by Maul *et al.* in PW91 calculations [12] but is unstable at the PBE+vdW level.
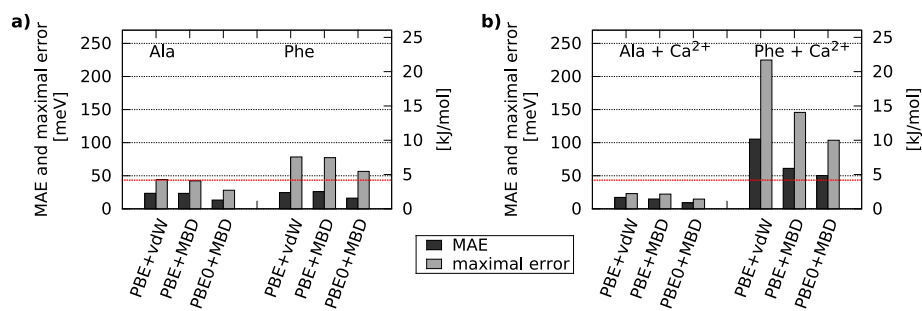
Figure 6: **Comparison of different DFAs to MP2 energies.** Mean absolute error (MAE) and maximal error (in meV) between different relative energies at the DFA (PBE+vdW, PBE+MBD*, and PBE0+MBD*) and MP2 level of theory, using structures of obtained minima on the PBE+vdW level from the database for the systems of Ala and Phe with neutral end caps, both in isolation and in complex with a Ca^{2+} cation. Computational details are given in the text. Exact numbers are summarized in Table 1.
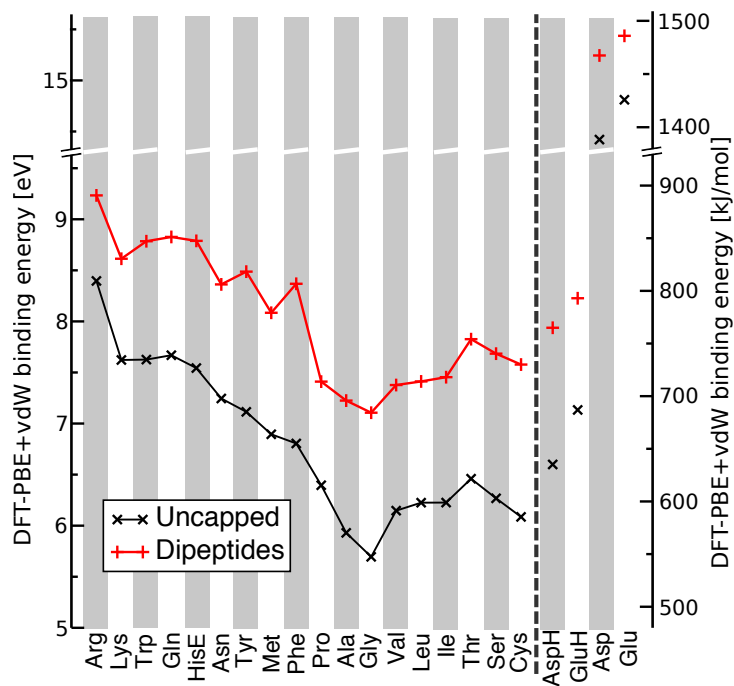
Figure 7: **Comparison of the gas-phase binding energies of Ca^{2+} to different amino acids** calculated in this work (y axis) to the experimentally inferred hierarchy of gas-phase binding energies of Ca^{2+} to different amino acids by Ho *et al.* [109] The amino acids are ordered along the x axis from the highest to lowest experimental Ca^{2+} binding energy. Protonated and deprotonated Asp and Glu are not included among the experimental data and are here shown as predictions. E_{binding} is high for deprotonated Asp and Glu since these forms of the amino acid would carry a negative charge.

Table 1: Mean absolute error (MAE) and maximal error (in meV; in parentheses: in kcal/mol) between different relative energies at the DFA (PBE+vdW, PBE+MBD*, and PBE0+MBD*) and MP2 level of theory, using structures of obtained minima on the PBE+vdW level from the database for the systems of Ala and Phe with neutral end caps, both in isolation and in complex with a Ca^{2+} cation. Computational details are given in the text.

System		MAE [meV]	Maximal error [meV]
Ala	PBE+vdW	24 (0.5)	44 (1.0)
	PBE+MBD*	23 (0.5)	44 (1.0)
	PBE0+MBD*	13 (0.3)	28 (0.6)
Phe	PBE+vdW	25 (0.6)	78 (1.8)
	PBE+MBD*	26 (0.6)	77 (1.8)
	PBE0+MBD*	16 (0.4)	57 (1.3)
Ala+Ca^{2+}	PBE+vdW	17 (0.4)	23 (0.5)
	PBE+MBD*	15 (0.3)	22 (0.5)
	PBE0+MBD*	9 (0.2)	15 (0.3)
Phe+Ca^{2+}	PBE+vdW	105 (2.4)	225 (5.2)
	PBE+MBD*	61 (1.4)	146 (3.4)
	PBE0+MBD*	50 (1.2)	104 (2.4)

Tables

Further tables are provided in a Microsoft Excel file and as tab-delimited text files as Supplementary Information to this article:

Table S1 Parameters specific to the REMD simulations of the different systems: the number of *Replicas*, the probability of *Acceptance* as well as the *Time* between exchange attempts, and the *Temperature* range of the replicas.

Table S2 Number of conformers found in the different stages (after global search and after refinement) of the search scheme for amino acids, dipeptides, and complexes thereof with Ca^{2+} cations. For the amino acids, the basin hopping search was performed starting from the non-zwitterionic as well as from the zwitterionic state. These numbers are separated by a "+" in the respective column.

Table S3 Numbers of conformers found for the amino acids (AA) and their complexes with the investigated divalent cations.

Table S4 Numbers of conformers found for the dipeptides (Dip.) and their complexes with the investigated divalent cations.

Table S5a Extrapolation of SCF energies as proposed by Karton and Martin: $E_{SCF}^n = E_{SCF}^{CBS} + A * e^{(-alpha*\sqrt{n})}$ with $n = 3, 4, 5$; A, $alpha$, E_{SCF}^{CBS} to be determined by a least squares fit; perfect fit as $\#parameters = \#datapoints = 3$; all values in eV.

Table S5b Extrapolation of MP2 correlation energies as proposed by Truhlar: $E_{corr}^n = E_{corr}^{CBS} + B * n^{-beta}$ with $n = 3, 4, 5$; B, $beta$, E_{corr}^{CBS} to be determined by a least squares fit; perfect fit as perfect fit as $\#parameters = \#datapoints = 3$; all values in eV.

Table S6 Basis set superposition errors (BSSE) for SCF and MP2 correlation energies with $n = T/Q/5$; all values in eV.

Table S7 Relative gas-phase Ca^{2+} binding energies for the amino acids from experiments by Ho *et al.* [109] and absolute binding energies in the gas phase from DFT-PBE+vdW calculations for amino acids and dipeptides.

References

[1] Wenbo Yu, Xuee Xu, Hongbao Li, Rui Pang, Kun Fang, and Zijing Lin. Extensive conformational searches of 13 representative dipeptides and an efficient method for dipeptide structure determinations based on amino acid conformers. *J. Comput. Chem.*, 30(13):2105–2121, 2009.

[2] Shyam Kishor, Suman Dhayal, Manjula Mathur, and Lavanya M. Ramaniah. Structural and energetic properties of α-amino acids: A first principles density functional study. *Mol. Phys.*, 106(19):2289–2300, 2008.

[3] Attila G. Császár and András Perczel. Ab initio characterization of building units in peptides and proteins. *Prog. Biophys. Mol. Biol.*, 71(2):243 – 309, 1999.

[4] Guy Bouchoux. Gas phase basicities of polyfunctional molecules. Part 3: Amino acids. *Mass Spectrom. Rev.*, 31(3):391–435, 2012.

[5] Chérif F. Matta and Richard F. W. Bader. Atoms-in-molecules study of the genetically encoded amino acids. II. Computational study of molecular geometries. *Proteins: Struct., Funct., Bioinf.*, 48(3):519–538, 2002.

[6] Sebastian Schlund, Robert Müller, Carsten Grassmann, and Bernd Engels. Conformational analysis of arginine in gas phase - a strategy for scanning the potential energy surface effectively. *J. Comput. Chem.*, 29(3):407–415, 2008.

[7] Attila G. Császár. On the structures of free glycine and α-alanine. *J. Mol. Struct.*, 346(0):141 – 152, 1995.

[8] Attila G. Császár. Conformers of gaseous glycine. *J. Am. Chem. Soc.*, 114(24):9568–9575, 1992.

[9] Vanessa Riffet, Gilles Frison, and Guy Bouchoux. Acid-base thermochemistry of gaseous oxygen and sulfur substituted amino acids (Ser, Thr, Cys, Met). *Phys. Chem. Chem. Phys.*, 13:18561–18580, 2011.

[10] Martin Kabelac, Pavel Hobza, and Vladimir Spirko. The ab initio assigning of the vibrational probing modes of tryptophan: Linear shifting of approximate anharmonic frequencies vs. multiplicative scaling of harmonic frequencies. *Phys. Chem. Chem. Phys.*, 11:3921–3926, 2009.

[11] R. Kaschner and D. Hohl. Density functional theory and biomolecules: A study of glycine, alanine, and their oligopeptides. *J. Phys. Chem. A*, 102(26):5111–5116, 1998.

[12] R. Maul, F. Ortmann, M. Preuss, K. Hannewald, and F. Bechstedt. DFT studies using supercells and projector-augmented waves for structure, energetics, and dynamics of glycine, alanine, and cysteine. *J. Comput. Chem.*, 28(11):1817–1833, 2007.

[13] P. Selvarengan and P. Kolandaivel. Potential energy surface study on glycine, alanine and their zwitterionic forms. *J. Mol. Struct.: THEOCHEM*, 671(1–3):77 – 86, 2004.

[14] Ming Cao, Susan Q. Newton, Julianto Pranata, and Lothar Schäfer. *J. Mol. Struct. THEOCHEM*, 332:251, 1995.

[15] Heather M. Jaeger, Henry F. Schaefer, Jean Demaison, Attila G. Császár, and Wesley D. Allen. Lowest-lying conformers of alanine: Pushing theory to ascertain precise energetics and semiexperimental re structures. *J. Chem. Theory Comput.*, 6(10):3066–3078, 2010.

[16] Michael D. Beachy, David Chasman, Robert B. Murphy, Thomas A. Halgren, and Richard A. Friesner. Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Am. Chem. Soc.*, 119(25):5908–5920, 1997.

[17] K. Y. Baek, Y. Fujimura, M. Hayashi, S. H. Lin, and S. K. Kim. Density functional theory study of conformation-dependent properties of neutral and radical cationic l-tyrosine and l-tryptophan. *J. Phys. Chem. A*, 115(34):9658–9668, 2011.

[18] Mingliang Chen and Zijing Lin. Ab initio studies of aspartic acid conformers in gas phase and in solution. *J. Chem. Phys.*, 127(15):–, 2007.

[19] Franca Maria Floris, Claudia Filippi, and Claudio Amovilli. A density functional and quantum monte carlo study of glutamic acid in vacuo and in a dielectric continuum medium. *J. Chem. Phys.*, 137(7):075102, 2012.

[20] A. L. Heaton, R. M. Moision, and P. B. Armentrout. Experimental and theoretical studies of sodium cation interactions with the acidic amino acids and their amide derivatives. *J. Phys. Chem. A*, 112(15):3319–3327, 2008.

[21] P.B. Armentrout, Amy Gabriel, and R.M. Moision. An experimental and theoretical study of alkali metal cation/methionine interactions. *Int. J. Mass Spectrom.*, 283(1-3):56 – 68, 2009.

[22] D. T. Nguyen, A. C. Scheiner, J. W. Andzelm, S. Sirois, D. R. Salahub, and A. T. Hagler. A density functional study of the glycine molecule: Comparison with post-hartree–fock calculations and experiment. *J. Comput. Chem.*, 18(13):1609–1631, 1997.

[23] Clifton Espinoza, Jan Szczepanski, Martin Vala, and Nick C. Polfer. Glycine and its hydrated complexes: A matrix isolation infrared study. *J. Phys. Chem. A*, 114(18):5919–5927, 2010.

[24] Bram Boeckx, Wouter Nelissen, and Guido Maes. Potential energy surface and matrix isolation ft-ir study of isoleucine. *J. Phys. Chem. A*, 116(12):3247–3258, 2012.

[25] David M. Close. Calculated vertical ionization energies of the common alpha-amino acids in the gas phase and in solution. *J. Phys. Chem. A*, 115(13):2900–2912, 2011.

[26] K. Y. Baek, M. Hayashi, Y. Fujimura, S. H. Lin, and S. K. Kim. Investigation of conformation-dependent properties of l-phenylalanine in neutral and radical cations by using a density functional taking into account noncovalent interactions. *J. Phys. Chem. A*, 114(28):7583–7589, 2010.

[27] Michelle A. Sahai, Tara A. K. Kehoe, Joseph C. P. Koo, David H. Setiadi, Gregory A. Chass, Bela Viskolcz, Botond Penke, Emil F. Pai, and Imre G. Csizmadia. First principle computational study on the full conformational space of l-proline diamides. *J. Phys. Chem. A*, 109(11):2660–2679, 2005.

[28] Rolf Linder, Kai Seefeld, Andreas Vavra, and Karl Kleinermanns. Gas phase infrared spectra of nonaromatic amino acids. *Chem. Phys. Lett.*, 453(1–3):1 – 6, 2008.

[29] Nidhi Vyas and Animesh K. Ojha. Investigation on transition states of [alanine + m2+] (m = ca, cu, and zn) complexes: A quantum chemical study. *Int. J. Quant. Chem.*, 112(5):1526–1536, 2012.

[30] R. J. Lavrich, D. F. Plusquellic, R. D. Suenram, G. T. Fraser, A. R. Hight Walker, and M. J. Tubergen. Experimental studies of peptide bonds: Identification of the c[sub 7][sup eq] conformation of the alanine dipeptide analog n-acetyl-alanine n[sup [prime]]-methylamide from torsion-rotation interactions. *J. Chem. Phys.*, 118(3):1253–1265, 2003.

[31] Meiling Zhang, Zhijian Huang, and Zijing Lin. Systematic ab initio studies of the conformers and conformational distribution of gas-phase tyrosine. *J. Chem. Phys.*, 122(13):134313, 2005.

[32] Supaporn Dokmaisrijan, Vannajan Sanghiran Lee, and Piyarat Nimmanpipug. The gas phase conformers and vibrational spectra of valine, leucine and isoleucine: An ab initio study. *J. Mol. Struct.: THEOCHEM*, 953(1-3):28–38, AUG 15 2010.

[33] Maria L. Ceci, Maria A. Lopez Verrilli, Sandra S. Vallcaneras, José A. Bombasaro, Ana M. Rodriguez, Botond Penke, and Ricardo D. Enriz. Exploratory conformational analysis of n-acetyl-l-tryptophan-n-methylamide. an ab initio study. *J. Mol. Struct.: THEOCHEM*, 631(1-3):277 – 290, 2003.

[34] Minglian Chen, Zhijian Huang, and Zijing Lin. *J. Mol. Struct. THEOCHEM*, 719:153, 2005.

[35] Monee Rassolian, Gregory A Chass, David H Setiadi, and Imre G Csizmadia. Asparagine—ab initio structural analyses. *J. Mol. Struct.: {THEOCHEM}*, 666–667(0):273 – 278, 2003.

[36] M.A. Zamora, H.A. Baldoni, J.A. Bombasaro, M.L. Mak, A. Perczel, O. Farkas, and R.D. Enriz. An exploratory ab initio study of the full conformational space of n-acetyl-l-cysteine-n-methylamide. *J. Mol. Struct.: {THEOCHEM}*, 540(1–3):271 – 283, 2001.

[37] Amareshwar Kumar Rai, Ce Song, and Zijing Lin. An exploration of conformational search of leucine molecule and their vibrational spectra in gas phase using ab initio methods. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 73(5):865 – 870, 2009.

[38] G. von Helden, I. Compagnon, M. N. Blom, M. Frankowski, U. Erlekam, J. Oomens, B. Brauer, R. B. Gerber, and G. Meijer. Mid-ir spectra of different conformers of phenylalanine in the gas phase. *Phys. Chem. Chem. Phys.*, 10:1248–1256, 2008.

[39] Vanessa Riffet and Guy Bouchoux. Gas-phase structures and thermochemistry of neutral histidine and its conjugated acid and base. *Phys. Chem. Chem. Phys.*, 15:6097–6106, 2013.

[40] Dorit Shemesh, Andrzej L. Sobolewski, and Wolfgang Domcke. Role of excited-state hydrogen detachment and hydrogen-transfer processes for the excited-state deactivation of an aromatic dipeptide: N-acetyl tryptophan methyl amide. *Phys. Chem. Chem. Phys.*, 12:4899–4905, 2010.

[41] Pohl Gabor, Andras Perczel, Elemer Vass, Gabor Magyarfalvi, and Gyorgy Tarczay. A matrix isolation study on ac-gly-nhme and ac-l-ala-nhme, the simplest chiral and achiral building blocks of peptides and proteins. *Phys. Chem. Chem. Phys.*, 9:4698–4708, 2007.

[42] Joost M. Bakker, Luke Mac Aleese, Gerard Meijer, and Gert von Helden. Fingerprint ir spectroscopy to probe amino acid conformations in the gas phase. *Phys. Rev. Lett.*, 91:203003, Nov 2003.

[43] Susana Blanco, M. Eugenia Sanz, Juan C. López, and José L. Alonso. Revealing the multiple structures of serine. *Proc. Natl. Acad. Sci. U.S.A.*, 104(51):20183–20188, 2007.

[44] Tamás Szidarovszky, Gábor Czakó, and Attila G. Császár. *Mol. Phys.*, 107:761, 2009.

[45] Bram Boeckx and Guido Maes. Experimental and theoretical observation of different intramolecular h-bonds in lysine conformations. *J. Phys. Chem. B*, 116(41):12441–12449, 2012.

[46] Lingbiao Meng and Zijing Lin. Comprehensive computational study of gas-phase conformations of neutral, protonated and deprotonated glutamic acids. *Computational and Theoretical Chemistry*, 976(1–3):42 – 50, 2011.

[47] R. Shankar, P. Kolandaivel, and L. Senthilkumar. Interaction studies of cysteine with li+, na+, k+, be2+, mg2+, and ca2+ metal cation complexes. *Journal of Physical Organic Chemistry*, 24(7):553–567, 2011.

[48] G. J. Fleming, P. R. McGill, and H. Idriss. Gas phase interaction of l-proline with be2+, mg2+ and ca2+ ions: a computational study. *Journal of Physical Organic Chemistry*, 20(12):1032–1042, 2007.

[49] Ching-Han Hu, Mingzuo Shen, and Henry F. Schaefer. *J. Am. Chem. Soc.*, 115:2923, 1993.

[50] V. Barone, M. Biczysko, J. Bloino, and C. Puzzarini. Characterization of the elusive conformers of glycine from state-of-the-art structural, thermodynamic, and spectroscopic computations: Theory complements experiment. *J. Chem. Theory Comput.*, 9(3):1533–1547, 2013.

[51] H. Q. Ai, Y. X. Bu, P. Li, and C. Zhang. The regulatory roles of metal ions (m+/2+ = li+, na+, k+, be2+, mg2+, and ca2+) and water molecules in stabilizing the zwitterionic form of glycine derivatives. *New J. Chem.*, 29(12):1540–1548, 2005.

[52] Carsten Baldauf and Hans-Jörg Hofmann. Ab initio mo theory–an important tool in foldamer research: Prediction of helices in oligomers of ω-amino acids. *Helvetica Chimica Acta*, 95(12):2348–2383, 2012.

[53] Michael Ramek, Anne-Marie Kelterer, and Sonja Nikolić. Ab initio and molecular mechanics conformational analysis of neutral l-proline. *Int. J. Quant. Chem.*, 65(6):1033–1045, 1997.

[54] Eszter Czinki and Attila G. Császár. Conformers of gaseous proline. *Chem. Eur. J.*, 9(4):1008–1019, 2003.

[55] Young Kee Kang. *J. Phys. Chem.*, 100:11589, 1996.

[56] Shan Xu Wang Ke-Dong, Ma Peng-Fei. Conformation effects on the molecular orbitals of serine. *Chinese Physics B*, 20(3):33102, 2011.

[57] Yongna Yuan, Matthew J. L. Mills, Paul L. A. Popelier, and Frank Jensen. Comprehensive analysis of energy minima of the 20 natural amino acids. *J. Phys. Chem. A*, 118(36):7876–7891, 2014.

[58] Amir Karton, Li-Juan Yu, ManojK. Kesharwani, and JanM.L. Martin. Heats of formation of the amino acids re-examined by means of w1-f12 and w2-f12 theories. *Theoretical Chemistry Accounts*, 133(6), 2014.

[59] Manoj K. Kesharwani, Amir Karton, and Jan M. L. Martin. Benchmark ab initio conformational energies for the proteinogenic amino acids through explicitly correlated methods. assessment of density functional methods. *J. Chem. Theory Comput.*, 2015.

[60] Richard H. Holm, Pierre Kennepohl, and Edward I. Solomon. Structural and functional aspects of metal sites in biology. *Chemical Reviews*, 96(7):2239–2314, 1996.

[61] John A. Tainer, Victoria A. Roberts, and Elizabeth D. Getzoff. Protein metal-binding sites. *Current Opinion in Biotechnology*, 3(4):378 – 387, 1992.

[62] Michael Kirberger and Jenny J. Yang. Structural differences between Pb^{2+}- and Ca^{2+}-binding sites in proteins: Implications with respect to toxicity. *Journal of Inorganic Biochemistry*, 102(10):1901–1909, 2008.

[63] Minyun Zhou, Xianchi Dong, Carsten Baldauf, Hua Chen, Yanfeng Zhou, Timothy A Springer, Xinping Luo, Chen Zhong, Frauke Gräter, and Jianping Ding. A novel calcium-binding site of von Willebrand factor A2 domain regulates its cleavage by ADAMTS13. *Blood*, 117(17):4623–4631, 2011.

[64] Ricky Cheng and Boris Zhorov. Docking of calcium ions in proteins with flexible side chains and deformable backbones. *European Biophysics Journal*, 39:825–838, 2010.

[65] Sanah Sadiq, Zena Ghazala, Arnab Chowdhury, and Dietrich Büsselberg. Metal toxicity at the synapse: Presynaptic, postsynaptic, and long-term effects. *Journal of Toxicology*, 2012.

[66] Sandeep K. Sharma, Pierre Goloubinoff, and Philipp Christen. Heavy metal ions are potent inhibitors of protein folding. *Biochemical and Biophysical Research Communications*, 372(2):341 – 345, 2008.

[67] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[68] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[69] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, Oct 1996.

[70] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.*, 102(7):073005, Feb 2009.

[71] Mariana Rossi, Volker Blum, Peter Kupser, Gert von Helden, Frauke Bierau, Kevin Pagel, Gerard Meijer, and Matthias Scheffler. Secondary structure of Ac-Ala$_n$-LysH$^+$ polyalanine peptides (n = 5,10,15) in vacuo: Helical or not? *J. Phys. Chem. Lett.*, 1(24):3465–3470, 2010.

[72] Alexandre Tkatchenko, Mariana Rossi, Volker Blum, Joel Ireta, and Matthias Scheffler. Unraveling the stability of polypeptide helices: Critical role of van der Waals interactions. *Phys. Rev. Lett.*, 106:118102, Mar 2011.

[73] Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Beate Koksch, Volker Blum, and Matthias Scheffler. How cations change peptide structure. *Chemistry - A European Journal*, 19(34):11224–11234, 2013.

[74] Sucismita Chutia, Mariana Rossi, and Volker Blum. Water adsorption at two unsolvated peptides with a protonated lysine residue: From self-solvation to solvation. *J. Phys. Chem. B*, 116(51):14788–14804, 2012.

[75] Mariana Rossi, Matthias Scheffler, and Volker Blum. Impact of vibrational entropy on the stability of unsolvated peptide helices with increasing length. *J. Phys. Chem. B*, 117(18):5574–5584, 2013.

[76] Mariana Rossi, Sucismita Chutia, Matthias Scheffler, and Volker Blum. Validation challenge of density-functional theory for peptides - example of Ac-Phe-Ala$_5$-LysH$^+$. *J. Phys. Chem. A*, 118(35):7349–7359, 2014.

[77] Franziska Schubert, Kevin Pagel, Mariana Rossi, Stephan Warnke, Mario Salwiczek, Beate Koksch, Gert von Helden, Volker Blum, Carsten Baldauf, and Matthias Scheffler. Native like helices in a specially designed [small beta] peptide in the gas phase. *Phys. Chem. Chem. Phys.*, 17:5376–5385, 2015.

[78] Franziska Schubert, Mariana Rossi, Carsten Baldauf, Kevin Pagel, Stephan Warnke, Gert von Helden, Frank Filsinger, Peter Kupser, Gerard Meijer, Mario Salwiczek, Beate Koksch, Matthias Scheffler, and Volker Blum. Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala19-Lys + H$^+$ vs. Ac-Lys-Ala19 + H$^+$ and the current reach of DFT. *Phys. Chem. Chem. Phys.*, page 10.1039/C4CP05541A, 2015.

[79] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[80] David J. Wales and Jonathan P. K. Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A*, 101(28):5111–5116, 1997.

[81] David J. Wales and Harold A. Scheraga. Global optimization of clusters, crystals, and biomolecules. *Science*, 285(5432):1368–1372, 1999.

[82] Jay W. Ponder and Frederic M. Richards. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. *J. Comput. Chem.*, 8(7):1016–1024, 1987.

[83] Pengyu Ren and Jay W. Ponder. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B*, 107(24):5933–5947, 2003.

[84] Volker Blum, Ralf Gehrke, Felix Hanke, Paula Havu, Ville Havu, Xinguo Ren, Karsten Reuter, and Matthias Scheffler. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications*, 180(11):2175 – 2196, 2009.

[85] V. Havu, V. Blum, P. Havu, and M. Scheffler. Efficient integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics*, 228(22):8367 – 8379, 2009.

[86] Xinguo Ren, Patrick Rinke, Volker Blum, Jürgen Wieferink, Alexandre Tkatchenko, Andrea Sanfilippo, Karsten Reuter, and Matthias Scheffler. Resolution-of-identity approach to Hartree-Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. *New Journal of Physics*, 14(5):053020, 2012.

[87] J.H. van Lenthe, S. Faas, and J.G. Snijders. Gradients in the ab initio scalar zeroth-order regular approximation (ZORA) approach. *Chem. Phys. Lett.*, 328(1-2):107–112, 2000.

[88] Christoph van Wullen. Molecular density functional calculations in the regular relativistic approximation: Method, application to coinage metal diatomics, hydrides, fluorides and chlorides, and comparison with first-order relativistic calculations. *J. Chem. Phys.*, 109(2):392–399, 1998.

[89] Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57:2607–2609, Nov 1986.

[90] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1-2):141 – 151, 1999.

[91] David J. Earl and Michael W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7:3910, 2005.

[92] Elizabeth C. Beret, Luca M. Ghiringhelli, and Matthias Scheffler. Free gold clusters: beyond the static, monostructure description. *Faraday Discuss.*, 152:153–167, 2011.

[93] Daniel Sindhikara, Yilin Meng, and Adrian E. Roitberg. Exchange frequency in replica exchange molecular dynamics. *J. Chem. Phys.*, 128(2):024103, 2008.

[94] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[95] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):pp. 100–108, 1979.

[96] Michael Feig, John Karanicolas, and Charles L. Brooks III. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling*, 22(5):377 – 395, 2004.

[97] John P. Perdew. *Unified Theory of Exchange and Correlation Beyond the Local Density Approximation*, chapter Electronic structure of solids '91 - proceedings of the 75. WE-Heraeus-Seminar and 21st Annual International Symposium on Electronic Structure of Solids held in Gaussig (Germany), March 11-15, 1991. Akademie Verlag, Berlin, 1991.

[98] C. Møller and M. S. Plesset. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.*, 46:618–622, 1934.

[99] M. Head-Gordon, J. A. Poplei, and M. J. Frisch. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.*, 153:503–506, 1988.

[100] F. Neese. The ORCA program system. *WIREs Comput. Mol. Sci.*, 2:73–78, 2012.

[101] D. E. Woon and T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. V. Corevalence basis sets for boron through neon. *J. Chem. Phys.*, 103:4572–4585, 1995.

[102] A. Karton and J. M. L. Martin. Comment on: "Estimating the Hartree–Fock limit from finite basis set calculations" [Jensen F (2005) Theor Chem Acc 113:267]. *Theor. Chem. Acc.*, 115:330–333, 2006.

[103] D.G. Truhlar. Basis-set extrapolation. *Chem. Phys. Lett.*, 294:45–48, 1998.

[104] A. Ambrosetti, A. M. Reilly, R. A. DiStasio, and A. Tkatchenko. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.*, 140:18A508, 2014.

[105] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110(13):6158–6170, 1999.

[106] Igor Ying Zhang, Xinguo Ren, Patrick Rinke, Volker Blum, and Matthias Scheffler. Numeric atom-centered-orbital basis sets with valence-correlation consistency from h to ar. *New Journal of Physics*, 15(12):123033, 2013.

[107] H. B. Jansen and P. Ros. Non-empirical molecular orbital calculations on the protonation of carbon monoxide. *Chem. Phys. Lett.*, 3:140–143, 1969.

[108] S.F. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.*, 19:553–566, 1970.

[109] Yen-Peng Ho, Ming-Wei Yang, Li-Ting Chen, and Yu-Chuan Yang. Relative calcium-binding strengths of amino acids determined using the kinetic method. *Rapid Communications in Mass Spectrometry*, 21(6):1083–1089, 2007.

[110] Adam Liwo, Mey Khalili, and Harold A. Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Natl. Acad. Sci. U.S.A.*, 102(7):2362–2367, 2005.

[111] Noel O'Boyle, Michael Banck, Craig James, Chris Morley, Tim Vandermeersch, and Geoffrey Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 2011.

Data Citations

Bibliographic information for the data records described in the manuscript.

1 Ropo, M., Baldauf, C., Blum, V. *NOMAD repository.* http://dx.doi.org/10.17172/NOMAD/20150526220502 (2015).