scientific data

OPEN

DATA DESCRIPTOR

Better force fields start with better data: A data set of cation dipeptide interactions

Xiaojuan Hu ✉, Maja-Olivia Lenz-Himmer & Carsten Baldauf ✉

We present a data set from a first-principles study of amino-methylated and acetylated (capped) dipeptides of the 20 proteinogenic amino acids – including alternative possible side chain protonation states and their interactions with selected divalent cations (Ca^{2+}, Mg^{2+} and Ba^{2+}). The data covers 21,909 stationary points on the respective potential-energy surfaces in a wide relative energy range of up to 4 eV (390 kJ/mol). Relevant properties of interest, like partial charges, were derived for the conformers. The motivation was to provide a solid data basis for force field parameterization and further applications like machine learning or benchmarking. In particular the process of creating all this data on the same first-principles footing, i.e. density-functional theory calculations employing the generalized gradient approximation with a van der Waals correction, makes this data suitable for first principles data-driven force field development. To make the data accessible across domain borders and to machines, we formalized the metadata in an ontology.

Background & Summary

Metal cations are essential to life: one third of the proteins in the human body require metal cofactors[1,2]. By shaping the structure of proteins, cations affect biological processes like molecular recognition or enzyme activity. Understanding the structure, dynamics, and function of metalloproteins is in the ongoing focus of many researchers, we summarize a few examples that involve simulation approaches: Tamames *et al.* analyzed zinc coordination spheres in a data set from the Protein Data Bank and complemented with DFT-B3LYP calculations[3]. Sala *et al.* investigated folding of Pyrococcus furiosus rubredoxin (PfRd), which includes an iron ion, with classical molecular dynamics (MD) simulations[4]. A calcium binding site in the blood protein von Willebrand Factor (VWF) regulates force-triggered unfolding for cleavage and therewith its activity in primary hemostasis, as illustrated by classical force-probe MD simulations[5]. Gogoi *et al.* investigated protein-metal ion binding affinities by analysing MD simulations of 49 different cation-protein complexes[6]. Metal cations can alter peptide structure by interacting with backbones and thereby enforcing non-Ramachandran geometries[7]. Cations can, by repulsion or attraction, also substantially reduce the conformational flexibility of functional sidechains[8,9].

MD simulations of biomolecules typically rely on additive force fields, where distinct terms describe bonded and non-bonded interactions based on empirically derived parameters. Studies have shown that the accuracy of force fields is especially limited when describing interactions involving ionic species[10–13]. In particular non-bonded interactions are critical, but of course the effect that nearby located cations exert on bonds is almost impossible to grasp by the combination of bonded and non-bonded interactions in a general-purpose force field. Modeling of electrostatic interactions via pairwise Coulomb potentials is based on assigning partial charges to atoms[14]. Partial charges are derived by: (i) fitting to experimental data (GROMOS and OPLS prior 2005), e.g. by fitting partial charges to reproduce hydration free enthalpies[15,16], (ii) deriving partial charges from QM calculations (Amber and Charmm)[17,18], or the combination of the two strategies (OPLS after 2005)[19].

The reliability of a force field also depends on the physics behind the formulation. The failures of established biomolecular force fields when describing cation-peptide systems may result from a central underlying assumption – modeling atoms by fixed point charges and neglecting charge transfer and polarization effects, while both are crucial to ionic systems[20–23]. Introducing more physics to the model appears a promising route to improve force fields: The inclusion of electronic polarization and charge transfer plays a central role in the next generations of biomolecular force fields[24–26]. However, including additional terms leads to force fields with way more

Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, 14195, Berlin, Germany. ✉e-mail: xhu@fhi.mpg.de; baldauf@fhi.mpg.de

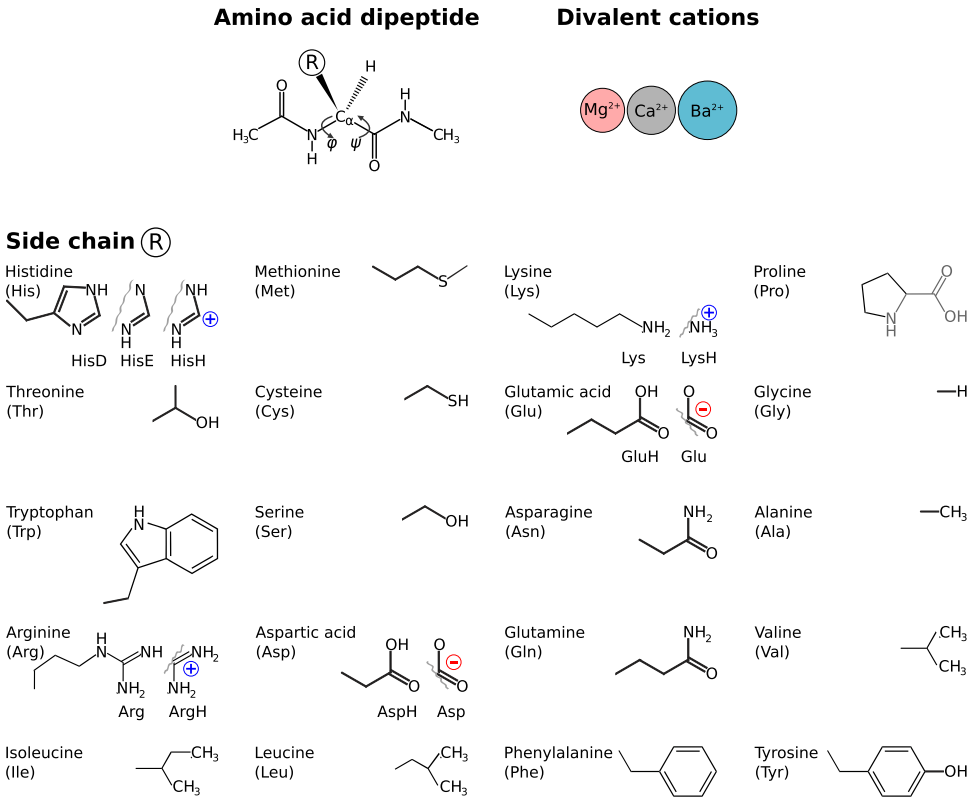


Fig. 1 The molecular systems in this study are dipeptides of the 19 proteinogenic amino acids that differ in the side chain **R** and the proteinogenic imino acid proline. Where applicable, different protonation states were considered.

parameters, which makes parameterization more challenging[27,28], in particular in the absence of high-resolution experimental data of less stable conformations, i.e. higher-energy structures[29]. To summarize, we see three main challenges:

- The availability of sufficiently-accurate electronic-structure data as well as choosing the "right ways" to derive e.g. partial charges from it.
- Designing the formulation of next-generation force fields that also include, for example, charge transfer and polarization.
- Finding sets of parameters (force fields) for such potentials in the absence of experimental data at sufficient spatial and time resolution.

Thorough studies have deepened our understanding of the conformational basics of individual building blocks, e.g.[30–41]. However, these studies are highly diverse with regards to the approximations made to model and to search the potential energy surfaces (PES) of the respective molecular systems; furthermore, the data is often not available. The availability of uniform and comprehensive computational data at an appropriately accurate level of theory has the potential to substantially increase the predictive power of force fields[42]. In order to provide such amino acid data sets for force field development on consistent computational footing, we extend previous work[43] by focusing on dipeptides as models of amino acid building blocks in polypeptide chains in complex with the divalent cations Mg^{2+}, Ca^{2+}, and Ba^{2+}, which play prominent roles in physiology: Mg^{2+} takes structural, catalytic, and regulatory roles[44] regulating ion channels, mitochondrial function, and cell's pH and volume[45]. Ca^{2+} levels regulate muscle contraction, hormone secretion, metabolism, ion transport, division, *etc*[46]. Mg^{2+} and Ca^{2+} may compete for the same binding sites[47]. Ba^{2+} can cause cardiac irregularities and affect the nervous system presumably by blocking potassium channels[48].

Combining these 3 cations with the proteinogenic amino acids in all meaningful side chain protonation states results in a data set that covers a wide range of molecular systems, see Fig. 1.

For the 21,909 stationary points, properties relevant to force field development were computed, details can be found in the Methods section. Making the data FAIR[49,50] – as in findable, accessible, interoperable, and reusable – is a challenge. In particular as we want to make the data available also to experts from other domains of science or to autonomous agents. To that end, we make the data freely available and also provide ontologies. An ontology defines a common vocabulary of basic concepts in a domain and relations among them[51]. The benefit is primarily that these definitions are machine-readable. This allows for interoperability between resources and databases as well as data interpretation across data collections. Through developed ontological representation of the data set, it can be connected to upper level concepts and thereby made machine-usable, which in turn

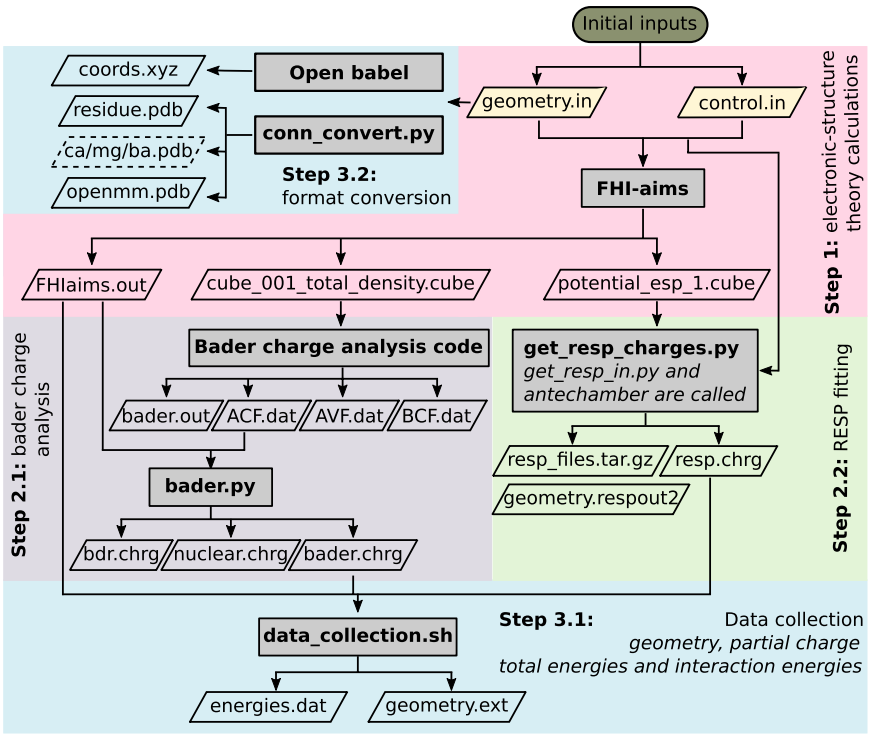


Fig. 2 Schematic representation of the workflow employed to derive properties of each conformer. Calculation steps were displayed in boxes with different background colour. Gray boxes indicate tools employed in each step. Parallelograms represent input and output files in each step. Links to custom codes are listed in Section *Code availability*.

enables automatic access and querying of the data. Ultimately, researchers can share their data with experts from other domains as well as making data available to machine intelligence.

Methods

Figure 1 summarizes the molecular systems in this study. Including the protonation states, we have to consider 26 dipeptides in 4 complexation states (bare, Ca^{2+}, Mg^{2+}, Ba^{2+}) which results in the 104 systems for which our structure searches identified 21,909 stationary points. For each of these stationary points, not only structure and energy are provided, but also further properties relevant to force field development, namely: van der Waals energies, interaction energies as well as electron densities and derived properties like the electrostatic potential, diverse partial charge models, and effective atomic volumes. By that, our dipeptide-cation data set allows one to explicitly assess subtle, but important, effects of local changes in the electrostatic environment due to peptide-cation interaction.

Sampling method. A hierarchical structure search that is described in detail in reference[43] was employed to locate stationary points on the potential energy surfaces of the 104 molecular systems. The initial global conformational searches of all dipeptides with/without Ca^{2+} were performed by a basin hopping search strategy[52,53] using the OPLS-AA force field[16]. Secondly, a refinement using density-functional theory calculations was performed. All electronic-structure calculations were performed with the all-electron, full-potential code FHI-aims utilizing numeric atom-centered basis functions[54–56]. The PBE generalized-gradient exchange-correlation functional[57] augmented by Tkatchenko's and Scheffler's pairwise van der Waals correction[58] was employed, and is referred to as PBE+vdW throughout this work. Stationary points that resulted from the FF-based pre-sampling were subjected to DFT-PBE+vdW relaxations with `light` settings. Next, a local first-principles based sampling step by *ab initio* replica-exchange molecular dynamics (REMD)[59,60] employing DFT-PBE+vdW with `light` settings, was applied to the identified set of structures. Conformers were extracted every 10 steps from REMD trajectories and clustered with a *k*-means clustering algorithm[61]. Obtained conformers went through relaxation with PBE+vdW (`light` computational settings), clustering and further relaxation with PBE+vdW (`tight` computational settings) to obtain the final conformational hierarchies. Initial structures of Mg^{2+} and Ba^{2+} binding dipeptides were obtained by substituting Ca^{2+} cation in dipeptide binding a Ca^{2+} cation. Subsequently, those were put into the procedure from *ab initio* REMD simulations to relaxation with PBE+vdW (`light` computational settings) to obtain final conformers as described before. These structures were further relaxed by PBE+vdW with `tight` computational settings.

Property calculations. Property calculations were performed on all structures obtained by the sampling method described above. This includes also high energy conformers. Figure 2 shows the processes involved in the property calculations; the individual steps are described in detail below. From the PBE+vdW DFT calculations with `tight` computational settings using FHI-aims, we collect in **Step 1** total energies, vdW energies, interaction energies, electron densities, electrostatic potential, Hirshfeld partial charges[62], and effective atomic volumes. Based on the effective atomic volumes V^{eff} per atom we provide, the effective vdW radii (R^0_{eff}) and the polarizability (α^0_{eff}) of an atom in a molecule can be calculated as follows[58,63]:

$$R^0_{\text{eff}} = R^0_{\text{free}} \left(\frac{V^{\text{eff}}}{V^{\text{free}}} \right)^{1/3} \tag{1}$$

$$\alpha^0_{\text{eff}} = \alpha^0_{\text{free}} \left(\frac{V^{\text{eff}}}{V^{\text{free}}} \right) \tag{2}$$

$$\frac{V_i^{\text{eff}}}{V_i^{\text{free}}} = \frac{\int r^3 \omega_i(\vec{r}) n(\vec{r}) d^3\vec{r}}{\int r^3 n_i^{\text{free}}(\vec{r}) d^3\vec{r}} \tag{3}$$

in which, R^0_{free} and α^0_{free} are the vdW radii of reference free-atom and static dipole polarizability (which can be taken from either experimental data or high-level quantum chemical calculations), respectively. V^{free} is the volume of the free atom *in vacuo*, r^3 is the cube of the distance from the nucleus of atom i, $\omega_i(\vec{r})$ is the Hirshfeld atomic partitioning weight for atom i, $n(\vec{r})$ is the total electron density, and $n_i^{\text{free}}(\vec{r})$ is the electron density of the free atom i.

The basic property resulting from a DFT calculation is the electron density, which – for each entry in our data set – was stored on a discrete grid of points with a spacing of 0.05 Å in a rectangular volume, which spans the whole molecule plus 14 Bohr (7.4 Å) beyond the outermost nuclei. The electrostatic potential exerted by a molecule on its environment may be used to derive partial charges. To that end, for each entry in the data set, five molecular surfaces were created by increasing the van der Waals radii of all atoms in the molecule (molecule with cation) by factors between 1.4 and 2.0. Points on these surfaces were represented in a cubic grid of each 35 grid points in x, y, and z direction. For these points, the electrostatic potential was evaluated. For biomolecular force fields, atomic partial charges are a crucial ingredient for computing the pairwise Coulomb term of the non-bonded interactions. We provide three types of partial charges:

- Hirshfeld atomic charges, computed by FHI-aims, were derived based on the Hirshfeld partitioning scheme[58,62]. The Hirshfeld atomic charge q_i of atom i is given by

$$q_i = Z_i - \int n_i(\vec{r}) d^3\vec{r} \tag{4}$$

where Z_i refers to the corresponding atomic number, and $n_i(\vec{r})$ is the associated electron density associated with atom i.

$$n_i(\vec{r}) = \omega_i(\vec{r}) n(\vec{r}) \tag{5}$$

where $n(\vec{r})$ denotes the total electron density, $\omega_i(\vec{r})$ is the Hirshfeld atomic partitioning weight for atom i. $\omega_i(\vec{r})$ is given by

$$\omega_i(\vec{r}) = \frac{n_i^{\text{free}}(\vec{r})}{\sum_A^{Allatoms} n_A^{\text{free}}(\vec{r})} \tag{6}$$

- Bader charges were being computed in **Step 2.1** using the Bader Charge Analysis tools[64–66] provided by the Henkelman group based on the electron density cube file produced in Step 1. The atoms in molecules (AIM) partitioning method uses what is called zero flux surfaces to distribute electron density among the atoms. Such zero flux surface is a two-dimensional surface on which the charge density is a minimum perpendicular to the surface. In molecular systems, the charge density typically reaches a minimum somewhere between pairs of neighboring nuclei. This can be seen as the natural place to separate atoms from each other. These borders between atoms define the electron density region associated with a given atom, from which the partial charges are being calculated.
- In **Step 2.2**, RESP partial charges[67–69] were computed using Antechamber[70] from the AmberTools package[71]. A two-stage restrained electrostatic potential (RESP) fitting procedure[67] was employed as implemented in Antechamber.

In the final **Steps 3.1 and 3.2**, data was collected and files converted to established formats. Geometry information is provided in three formats: the FHI-aims input format, the xyz format generated by Open Babel[72],

and PDB files that are readable by the CHARMM-GUI portal[73] and the openMM7 package[74]. Connectivity and atom type information – needed for the PDB format – was gathered based on atomic distances by the Python script `conn_convert.py`. Furthermore, energies and partial charges were tabulated for convenient usage. Interaction energies E_{inter} between cation and dipeptide were calculated as follows:

$$E_{\text{inter}} = E_{\text{complex}} - E_{\text{dipeptide}} - E_{\text{cation}} \tag{7}$$

where $E_{complex}$ corresponds to the potential energy of the dipeptide-cation complex, $E_{dipeptide}$ is the potential energy of the dipeptide alone fixed in the cation bound conformation, and E_{cation} is the potential energy of the isolated cation.

Further data and properties can be extracted from the raw and normalized data[75] that is available from the NOMAD Repository and Archive[76]. The data set was deposited as populated ontology in OWL format[77] in the EDMOND repository of the Max Planck Society. The construction of the ontology is described in the following subsection.

Ontology construction. Ontology construction is an iterative process involving many steps from defining common vocabularies, identifying the most important concepts and their relations to modelling such concepts in a semantically correct and still useful and applicable way. It can be used to enrich, annotate, and link data that is then called *linked data* and usually expressed in a semantic triple format consisting of *subject*, *predicate*, and *object*[78]. The main components of an ontology are classes, properties, individuals and axioms. Classes are the focus of most ontologies and are descriptions of concepts in a domain and represent a specific set of individuals. "Ala" is a class in the Amino Acid domain, thus each single Ala conformer in our data set is an individual of class "Ala". Properties describe features and attributes of classes and individuals. Properties can connect classes and individuals. For example, *hasProperty* can connect classes "Ala" and "Charge" as a property. Axioms are statements that all together define what is the truth in a given domain. In this work, the ontology builder Protégé[79] and the python package Owlready2[80] were employed to build ontologies in the OWL2 Web Ontology Language (http://www.w3.org/TR/owl2-overview) which is based on RDF – the Resource Description Framework (http://www.w3.org/TR/rdf-primer). Subjects and predicates are named using Internationalized Resource Identifiers (IRIs) (https://tools.ietf.org/html/rfc3987), while the object position can be filled by an IRI or a literal value (e.g. string or number). Ontologies created in this work have been tested with the OWL reasoner FACT++[81].

Data Records

Raw data and normalized data of the DFT calculations for this amino acid dipeptide data set is available from the NOMAD repository (http://nomad-repository.eu) via the https://doi.org/10.17172/NOMAD/2021.02.10-1[75]. The NOMAD Archive contains all raw input, output, and property calculation files for download, while the NOMAD Repository contains normalized data, i.e. a digest of the DFT calculations. Data in the NOMAD Repository and Archive is provided on the basis of the Creative Commons Attribution 3.0 License (CC BY 3.0) as it is stated in the NOMAD terms (https://nomad-lab.eu/terms).

The extracted data in form of a populated ontology in OWL format is available download via the https://doi.org/10.17617/3.5q10.17617/3.5q[77] under the Creative Commons Attribution 4.0 license (CC BY 4.0). In the following two subsections, we briefly introduce the data and the concept of the provided ontology.

DFT data set. The distribution of the 21,909 stationary points of the amino acid dipeptide (plus cation) systems over the different amino acid building blocks is summarized in Fig. 3. This data is in particular intended for training energy functions in machine learning approaches in the context of force field development and parameterization. Consequently, it consists not only of geometries with total energies for preferred low-energy conformers. Instead, DFT-PBE+vdW calculations also included high-energy conformers. The data we provide is particularly focused on parameterizing non-bonded interactions: The above-mentioned cation-peptide interaction energies were already used to tune force fields parameters of non-bonded interactions[26,82]. The comparison to DFT-based vdW energies computed with the Tkatchenko-Scheffler formalism[58] is useful to evaluate or adjust the non-bonded Lennard-Jones parameters ε and σ. Importantly, due to the spread over high and low energy conformations, diverse substructures and environments (due to cation binding), a range of partial charge values is sampled that informs about polarization and charge transfer. To that end, the electronic structure is simplified into partial charge models, based on Hirshfeld partitioning or Bader AIM analysis of the electron density. The electron density, in combination with the nuclear charges, also defines the electrostatic potential (ESP) around the molecule, which can be used to derive force field parameters related to electrostatic interaction[83]. The electron density has been used before to derive environment-specific force fields[84]. Electron densities for a large set of molecules have been used to predict partial charges based on machine learning[85,86], to that end, an average over similar substructures in different molecules was used.

The data is first of all made available as a set of files. The different files, their content, and which programs to read or write them are given in Table 1. A direct way to access the data is to download the compressed archive[75] and browse the folder structure that is given in Fig. 4 or download from the same source the normalized data in json-files.

This way of representing data however limits the automated access to the data by artificial agents or by researchers from other domain, as the metadata to the data is somewhat hidden. In order to alleviate this, the next section details the ontology which we developed in order to provide an extensible, machine-interpretable and machine-usable model for the automated access and post-processing of the data set.

Ontology. AAMI (Amino Acid Meta-Info) is an ontology created "bottom-up" to specifically represent the meta-information of this amino acid-cation data set in a machine-understandable and machine-processable way.

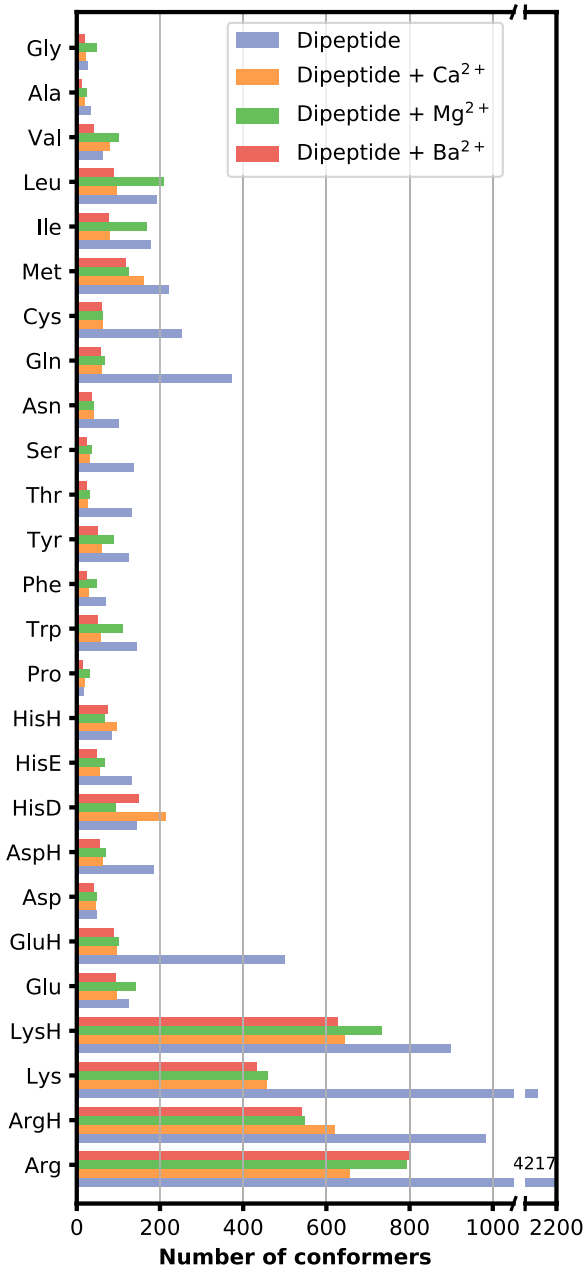


Fig. 3 Numbers of stationary points of each molecular system covered in this study.

AAMI does not only contain metadata of properties, it also covers processes of analysis, such as inputs, outputs, and tools in each process and their roles, which further makes data interpretable and understandable. Two existing ontologies were re-used in AAMI: the European Materials Modelling Ontology (EMMO) (https://emmc.info/emmo-info), which provides a representational framework for materials modelling and characterization knowledge, and the Amino Acid Ontology (http://bioportal.bioontology.org/ontologies/AMINO-ACID), which provides structured knowledge of amino acids and their properties. By reusing existing terms in EMMO and Amino Acid Ontology rather than creating the ontology from scratch, terms in AAMI were connected to upper level concepts and can be potentially linked to further ontologies. Moreover, users are able to take advantage of data and annotations that are already used in those ontologies and can by that also rely on concepts that were already agreed upon in a bigger community. The primary aim of AAMI is to make our data set FAIR (Findable, Accessible, Interoperable, and Reusable)[49], in particular accessible, interoperable and reusable. The elements of AAMI can be found in Fig. 5. In the AAMI ecosystem, we created:

File name	Description	Code/Format
FHI-aims Input Files		
geometry.in	Cartesian coordinates of the complexes	FHI-aims
control.in	Input file with technical parameters for electronic structure calculations	FHI-aims
FHI-aims Output Files		
FHIaims.out	Main output the electronic structure calculations, contains: total energy, vdW energy and effective atomic volume etc.	FHI-aims
cube_001_total_density.cube.bz2	Cube file representation of the electron density (bzip2 compressed)	FHI-aims
potential_esp_1.cube.bz2	Cube file representation of the electrostatic potential (bzip2 compressed)	FHI-aims
hirsh.chrg	Hirshfeld charges	Self-made
Geometries		
coords.xyz	Coordinate file	*xyz* format
residue.pdb	Coordinate file	CHARMM
[cation].pdb	Separate coordinate file for each of the cations Ca, Ba, Mg	CHARMM
openmm.pdb	Coordinate file	OpenMM
Bader AIM calculations		
ACF.dat, AVF.dat, BCF.dat, bader.out	Information of Bader charge analysis	Bader
nuclear.chrg, bdr.chrg	Information of Bader charge analysis	Self-made
bader.chrg	Bader charges	Self-made
RESP calculations		
geometry.respout2, resp_files.tar.gz	RESP charge information	Antechamber
resp.chrg	RESP charges	Self-made
Aggregated output		
geometry.ext	Collection of coordinate and charge information	Self-made
energies.dat	Collection of total energy and interaction energy	Self-made

Table 1. List and description of file types in the data set.

1. The cluster structure ontology (CSO) represents concepts and relations for structure description of non-periodic systems, EMMO was imported, and 351 classes and 2053 axioms were created.
2. The cluster property ontology (CPO) describes properties of non-periodic systems. CSO was imported, and 450 classes and 2984 axioms were created.
3. The force field ontology (FFO) represents concepts in force fields, e.g. atom type and atom class. Amino acid ontology and CPO were imported, and 563 classes and 4453 axioms were created.
4. AAMI represents concepts and relations in the amino acids-cation data set. FFO was imported, and 787 classes and 5466 axioms were created.
5. The different instances of AAMI-D-* are knowledge graphs created from the data set in this study. Such graph is build by populating AAMI with the data for an amino acid, e.g. ALA, ARG, *etc*., from this data set. The asterisk represents the name of the corresponding amino acid.

Partial high level class organization and some of the classes and relations of AAMI are shown in Fig. 6 to give an overview of the organization of the ontology and how terms from each ontology are related to each other.

The primary use of AAMI is to annotate database records. However, since ontologies were developed with the OWL2 Web Ontology Language, which represents data by sets of subject-predicate-object statements, so-called *triples*, the underlying computational logic enables automatic inference and querying over data repositories. In principle, any question framed in the respective mathematical logic can be answered in a finite number of steps. However, such reasoning capabilities are currently limited to description logic. Data query can be done with the ontology and linked data query language, SPARQL (https://www.w3.org/TR/sparql11-query). A user can query for sub-classes, relations between classes, functional annotation, and so on. Stardog Studio (https://www.stardog.com/studio) can be used as a *triple store* and employed to perform the SPARQL queries. A tutorial of SPARQL query language using Stardog Studio can be found in the following link: https://www.stardog.com/tutorials/sparql/. We provide two sample queries in this work to guide users to build their own queries.

Before any queries, a set of namespace prefixes were declared to abbreviate IRIs, e.g. the knowledge graph of alanine dipeptide was imported as an example under the `PREFIX ala`.

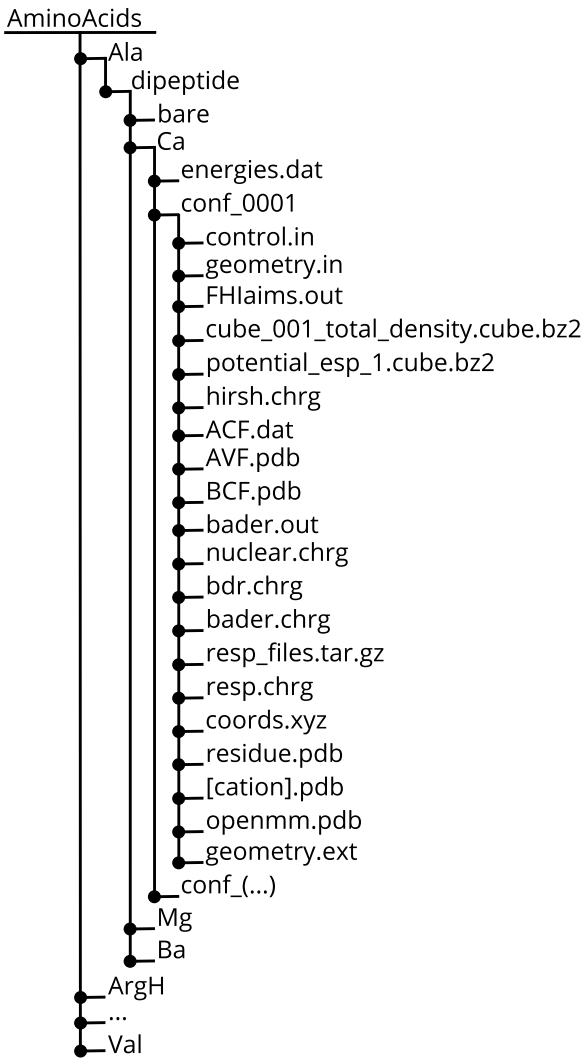


Fig. 4 Schematic representation of the folder structure of the data. Each folder, as exemplified for the Ca^{2+}-coordinated cysteine dipeptide, contains multiple properties per system.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX cso: <http://www.semanticweb.org/ClusterStructure.owl#>
PREFIX cpo: <http://www.semanticweb.org/ClusterProperty.owl#>
PREFIX ffo: <http://www.semanticweb.org/ForceField.owl#>
PREFIX aami: <http://www.semanticweb.org/AAMI.owl#>
PREFIX ala: <http://www.semanticweb.org/AAMI-D-Ala-Dipeptide.owl#>
PREFIX hasProperty:<http://emmo.info/emmo/middle/properties#
    EMMO_e1097637_70d2_4895_973f_2396f04fa204>
PREFIX hasSymbolData:<http://emmo.info/emmo/middle/perceptual#
    EMMO_23b579e1_8088_45b5_9975_064014026c42>
```

The main query form in SPARQL is a `SELECT` query. A `SELECT` query has two main components: a list of selected variables and a `WHERE` clause for specifying the graph patterns to match. For example, according to the graph shown in Fig. 6, we can query for Bader charges of atoms which have atom type of "1" in Amber10 with a `SELECT` query as follows:

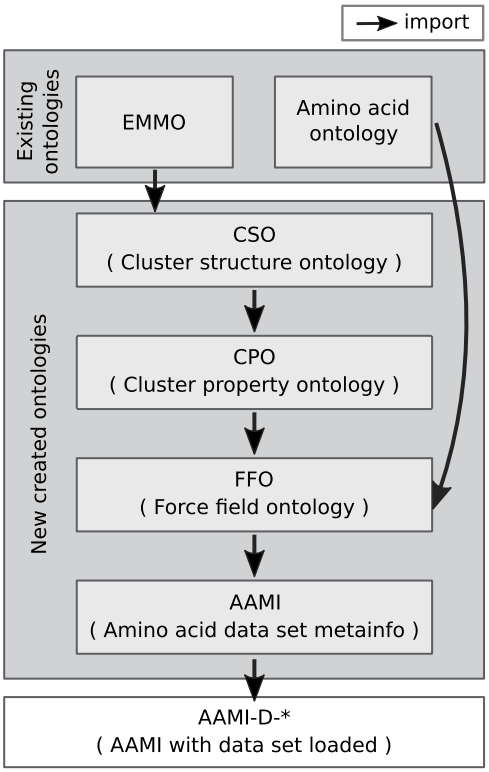


Fig. 5 Hierarchy of the ontologies linked to amino acid-cation meta-info (AAMI). Details of the ontologies and relations among them are described in Section *Ontology*.

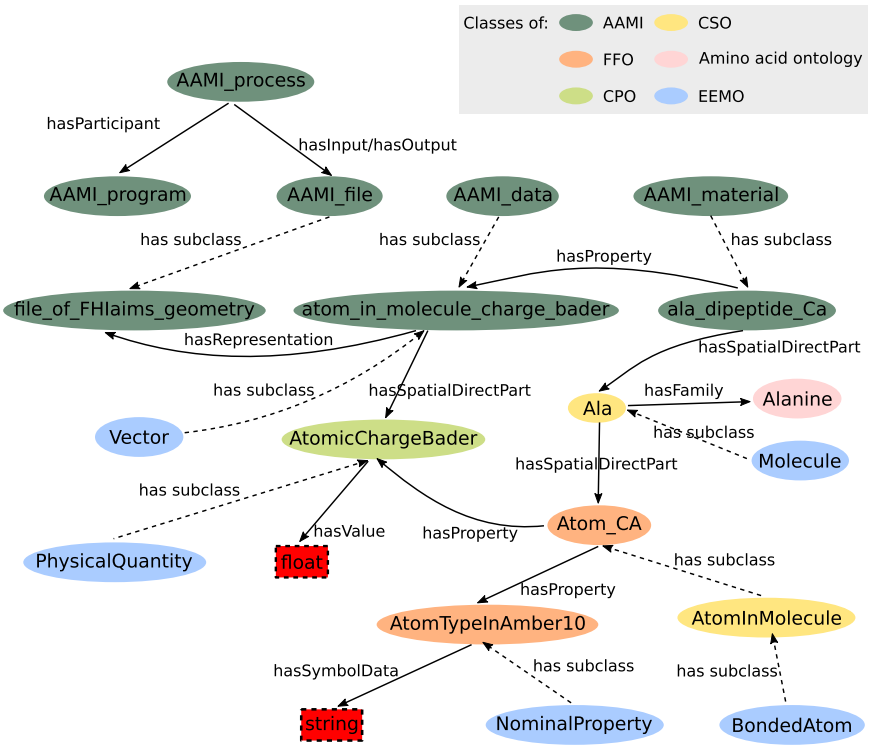


Fig. 6 Partial high-level class structure of AAMI ontology. Ovals represent classes, where classes from different ontologies are color coded. Rectangles represent literals. Solid lines are properties and dotted lines represent the relation of '*has subclass*'.

```
[...]
SELECT ?atom ?n
WHERE {
   ?atom hasProperty: ?atomtype.
   ?atomtype a ffo:AtomTypeInAmber10.
   ?atomtype hasSymbolData: "1"^^xsd:string.
   ?atom hasProperty: ?badercharge.
   ?badercharge a cpo:AtomicChargeBader.
   ?badercharge cpo:hasValue ?n
}
```

The resulting list shows all atoms of type "1" in Amber10, *i.e.* hydrogen atoms bound to a peptide bond nitrogen, and their Bader charges:

```
ala#Atom_HN_11_alaD_Ca_conf_0017 0.450512
ala#Atom_HN_11_alaD_Ca_conf_0018 0.486539
ala#Atom_HN_11_alaD_Ca_conf_0014 0.450169
ala#Atom_HN_11_alaD_Ca_conf_0012 0.484383
ala#Atom_HN_11_alaD_Ca_conf_0002 0.442222
ala#Atom_HN_11_alaD_Ca_conf_0006 0.452150
...
```

Another useful query is `DESCRIBE`, which returns all the outgoing edges of a node. `DESCRIBE` is most useful when we don't know much about the ontology and want to quickly see the terms used in the triples. For example, we can query "describe individuals which belong to class Atom_C" with `DESCRIBE` query within the alanine dipeptide knowledge graph:

```
[...]
DESCRIBE ?atom
WHERE {
   ?atom a ffo:Atom_C
}
```

In the following, we display part of the output of the query, from which we can see that an individual "Atom_C_9_alaD_Ca_conf_0017" belongs to class "Atom_C" and has properties of "AtomicChargeBader_1.35427", "position9" and so on.

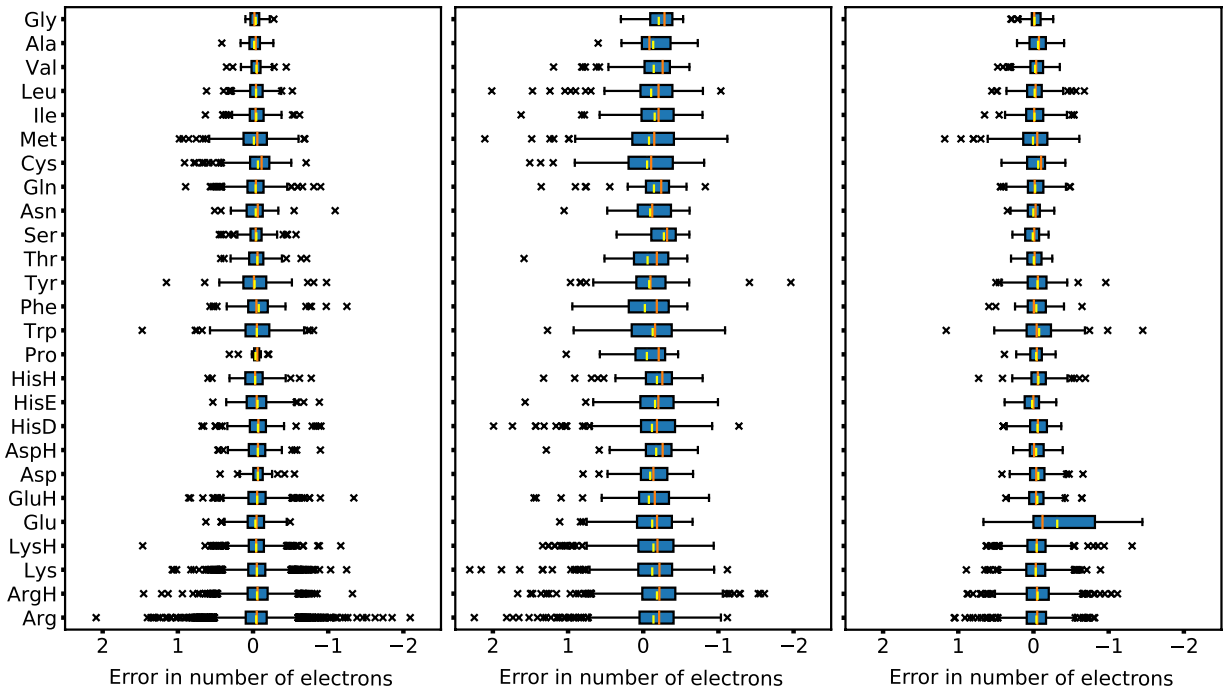
```
@prefix ffo: <http://www.semanticweb.org/ForceField.owl#> .
@prefix ala: <http://www.semanticweb.org/AAMI-D-Ala-Dipeptide.owl#> .
@prefix hasProperty: <http://emmo.info/emmo/middle/properties#
    EMMO_e1097637_70d2_4895_973f_2396f04fa204> .

{
   <ala#Atom_C_9_alaD_Ca_conf_0017> a owl:NamedIndividual , ffo:Atom_C ;
     <hasProperty> <ala#AtomicChargeBader_1.35427> , <ala#EffectivePolarizability_9
         .839967844590108> , <ala#position9>, ...
}
```

With tools like Stardog Studio, the results of such query can be written out in various file formats for further usage, e.g. `XML`, `JSON-LD` for triples output or `CSV` for tabular output.

Technical Validation

The reliability of the DFT-PBE + vdW level of theory for amino acids and amino acids binding divalent cations was evaluated before[43]. In this reference, single-point energy calculations were performed on all structures of alanine (Ala) and phenylalanine (Phe) amino acids in isolation, as well as binding with a Ca^{2+} cation employing Møller-Plesset second-order perturbation theory (MP2)[87,88]. For the structures of the amino acids Ala and Phe without cation bound, mean absolute errors (MAE) within chemical accuracy (1 kcal/mol) were estimated for PBE + vdW. A different long-range dispersion method, the many-body dispersion model (PBE + MBD)[89], didn't show significant improvements for isolated amino acids. Also the usage of a hybrid exchange-correlation functional, PBE0 (PBE0 + MBD)[89], did not significantly improve the MAEs. However, the maximum error of Phe was reduced from 2 kcal/mol to 1.3 kcal/mol. MAEs were slightly higher with PBE + vdW when Ca^{2+} was involved. They reached 1 kcal/mol and 2 kcal/mol for Ala + Ca^{2+} and Phe + Ca^{2+}, respectively. Employing both, many-body dispersion and the hybrid functional PBE0, improved the MAE to about 1 kcal/mol. In a manuscript on histidine-zinc interactions[11], DLPNO-CCSD(T)[90,91] was employed to benchmark several DFAs as well as the wave function-based MP2 method. The evaluated systems are (a) negatively charged acetylhistidine (AcH) with and without a Zn^{2+} cation, and (b) neutral AcH with and without a Zn^{2+} cation. The results showed that PBE+vdW gave an acceptable accuracy. In conclusion, PBE+vdW appears to be a valid starting point for studies on cation-peptide systems.

The validation of the sampling method can be elucidated by the work in ref. [92]. A genetic algorithm was employed to do the sampling of the low-energy segment in the conformational space of seven dipeptides:



(a) **Dipeptide bare** (b) **Dipeptide+Ca^{2+}** (c) **Dipeptide+Mg^{2+}**

Fig. 7 Error in numbers of electrons from Bader analysis of Dipeptide (**a**) bare, (**b**) with Ca^{2+} and (**c**) with Mg^{2+}. The upper and lower lines of the rectangles mark the 75% and 25% percentiles of the distribution, the orange and yellow horizontal lines in the box indicate the median (50% percentile) and mean value, and the upper and lower lines of the "error bars" depict the 99% and 1% percentiles. Crosses represent the outliers.

Glycine (Gly), Alanine (Ala), Phenylalanine (Phe), Valine (Val), Tryptophan (Trp), Leucine (Leu), Isoleucine (Ile). Conformers from our previous data set[43] were used as reference points and both studies agree in their overall structure findings.

The potential usage of our data set has been confirmed in ref. [26]. In this work, our data set was used to assess the accuracy of existing FFs by their abilities to reproduce quantum mechanical (QM) interaction energies of Ca^{2+}-dipeptide. By relating the parameter space to conformational space, the utility of our data set as a reference for future optimization of polarizable force fields is illustrated.

An assessment of the reliability of Bader charge analysis of bare dipeptides as well as dipeptide-Ca^{2+} and dipeptide-Mg^{2+} complexes is shown in Fig. 7. The number of electrons from Bader charge analysis yielded high errors in some structures of dipeptide-Ca^{2+}, reaching 2 electrons. This error apparently results from too wide grid spacing at regions of rapid density change (near "heavy" cores) when writing the electron density to cube files, the input for the Bader analysis code. Changes in electron density are particularly large close to the cations in the investigated clusters, so in principle grid spacings adjusted to the respective systems would be required. Overall, however, the mean errors of each amino acid are around 0. The errors of dipeptide-Mg^{2+} have the same trend, but are smaller than the errors of dipeptide-Ca^{2+} due to the smaller radius of Mg^{2+}. Ba^{2+} is much heavier than Ca^{2+} and Mg^{2+}, the rise in density close to the atomic center is much steeper. To analyze the Bader charges of dipeptide-Ba^{2+} complexes, a much smaller grid spacing is needed. However, this will result in electron density cube files that are impractically large for an overview study of this extend. So in this work, we did not present the electron density and Bader charges of dipeptide-Ba^{2+} complexes.

Usage Notes

Attention, the download of the whole archive of raw data is about 1.5 TB in size (compressed). Structures in this data set are stationary-point geometries, most of them can be expected to be minima, yet there are certainly also saddle points. All files in the NOMAD repository can be downloaded through `curl` based on upload and entry IDs (variables: `upload_id` and `entry_id` below). The command below downloads all files in one calculation:

```
curl "http://repository.nomad-coe.eu/app/api/raw/calc/upload_id/entry_id/*" -o download.zip
```

The metadata for the DFT calculations can in part be browsed at the NOMAD Archive page (https://www.nomad-coe.eu/the-project/nomad-archive/archive-meta-info). There are numerous tools to perform SPARQL queries, e.g. Stardog Studio (https://www.stardog.com/studio), Protégé[79], RDFLib (https://github.com/RDFLib/rdflib), Apache Jena (https://jena.apache.org), and so on. The licenses of Protégé, RDFLib, and Apache Jena are

BSD 2-Clause, BSD 3-Clause and Apache License 2.0, respectively; using Stardog Studio requires for a license from the developers.

Code availability

All custom codes used in this study have been uploaded to Github[93].

Received: 15 August 2021; Accepted: 18 March 2022;
Published online: 17 June 2022

References

1. Permyakov, E. *Metalloproteomics*, 2 (John Wiley & Sons, 2009).
2. Bertini, G. *et al. Biological inorganic chemistry: structure and reactivity* (University Science Books, 2007).
3. Tamames, B., Sousa, S. F., Tamames, J., Fernandes, P. A. & Ramos, M. J. Analysis of zinc-ligand bond lengths in metalloproteins: trends and patterns. *Proteins: Structure, Function, and Bioinformatics* **69**, 466–475 (2007).
4. Sala, D., Giachetti, A. & Rosato, A. Molecular dynamics simulations of metalloproteins: A folding study of rubredoxin from Pyrococcus furiosus. *AIMS Biophys* **5**, 77–96 (2018).
5. Zhou, M. *et al.* A novel calcium-binding site of von Willebrand factor A2 domain regulates its cleavage by ADAMTS13. *Blood* **117**, 4623–4631 (2011).
6. Gogoi, P., Chandravanshi, M., Mandal, S. K., Srivastava, A. & Kanaujia, S. P. Heterogeneous behavior of metalloproteins toward metal ion binding and selectivity: insights from molecular dynamics studies. *Journal of Biomolecular Structure and Dynamics* **34**, 1470–1485 (2016).
7. Baldauf, C. *et al.* How cations change peptide structure. *Chemistry–A European Journal* **19**, 11224–11234 (2013).
8. De, S., Musil, F., Ingram, T., Baldauf, C. & Ceriotti, M. Mapping and classifying molecules from a high-throughput structural database. *Journal of Cheminformatics* **9**, 1–14 (2017).
9. Ropo, M., Blum, V. & Baldauf, C. Trends for isolated amino acids and dipeptides: Conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead. *Scientific Reports* **6**, 1–11 (2016).
10. Vitalini, F., Mey, A. S., Noé, F. & Keller, B. G. Dynamic properties of force fields. *The Journal of Chemical Physics* **142**, 02B611_1 (2015).
11. Schneider, M. & Baldauf, C. Relative energetics of acetyl-histidine protomers with and without Zn^{2+} and a benchmark of energy methods. *arXiv preprint arXiv:1810.10596* (2018).
12. Maksimov, D., Baldauf, C. & Rossi, M. The conformational space of a flexible amino acid at metallic surfaces. *International Journal of Quantum Chemistry* **121**, e26369 (2021).
13. Marianski, M., Supady, A., Ingram, T., Schneider, M. & Baldauf, C. Assessing the accuracy of across-the-scale methods for predicting carbohydrate conformational energies for the examples of glucose and α-maltose. *Journal of Chemical Theory and Computation* **12**, 6157–6168 (2016).
14. Wang, J. & Kollman, P. A. Automatic parameterization of force field by systematic search and genetic algorithms. *Journal of Computational Chemistry* **22**, 1219–1228 (2001).
15. Oostenbrink, C., Villa, A., Mark, A. E. & Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry* **25**, 1656–1676 (2004).
16. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **118**, 11225–11236 (1996).
17. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **21**, 1049–1074 (2000).
18. Riniker, S. Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview. *Journal of Chemical Information and Modeling* **58**, 565–578 (2018).
19. Shivakumar, D., Harder, E., Damm, W., Friesner, R. A. & Sherman, W. Improving the prediction of absolute solvation free energies using the next generation opls force field. *Journal of chemical theory and computation* **8**, 2553–2558 (2012).
20. Allen, T. W., Andersen, O. S. & Roux, B. Energetics of ion conduction through the gramicidin channel. *Proceedings of the National Academy of Sciences* **101**, 117–122 (2004).
21. Roca, M. *et al.* Theoretical modeling of enzyme catalytic power: analysis of "cratic" and electrostatic factors in catechol O-methyltransferase. *Journal of the American Chemical Society* **125**, 7726–7737 (2003).
22. Zeng, J., Jia, X., Zhang, J. Z. & Mei, Y. The F130L mutation in streptavidin reduces its binding affinity to biotin through electronic polarization effect. *Journal of Computational Chemistry* **34**, 2677–2686 (2013).
23. Li, Y. L., Mei, Y., Zhang, D. W., Xie, D. Q. & Zhang, J. Z. Structure and dynamics of a dizinc metalloprotein: effect of charge transfer and polarization. *The Journal of Physical Chemistry B* **115**, 10154–10162 (2011).
24. Xie, W., Pu, J. & Gao, J. A coupled polarization-matrix inversion and iteration approach for accelerating the dipole convergence in a polarizable potential function. *The Journal of Physical Chemistry A* **113**, 2109–2116 (2009).
25. Ngo, V. *et al.* Quantum effects in cation interactions with first and second coordination shell ligands in metalloproteins. *Journal of Chemical Theory and Computation* **11**, 4992–5001 (2015).
26. Amin, K. S. *et al.* Benchmarking polarizable and non-polarizable force fields for Ca^{2+}–peptides against a comprehensive QM dataset. *The Journal of Chemical Physics* **153**, 144102 (2020).
27. Liang, G., Fox, P. C. & Bowen, J. P. Parameter analysis and refinement toolkit system and its application in MM3 parameterization for phosphine and its derivatives. *Journal of Computational Chemistry* **17**, 940–953 (1996).
28. Faller, R., Schmitz, H., Biermann, O. & Müller-Plathe, F. Automatic parameterization of force fields for liquids by simplex optimization. *Journal of Computational Chemistry* **20**, 1009–1017 (1999).
29. Cisneros, G. A., Karttunen, M., Ren, P. & Sagui, C. Classical electrostatics for biomolecular simulations. *Chemical Reviews* **114**, 779–814 (2014).
30. Rezac, J., Bm, D., Gutten, O. & Rulisek, L. Toward accurate conformational energies of smaller peptides and medium-sized macrocycles: MPCONF196 benchmark energy data set. *Journal of Chemical Theory and Computation* **14**, 1254–1266 (2018).
31. Jurečka, P., Šponer, J., Černý, J. & Hobza, P. Benchmark database of accurate (MP2 and CCSD (T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical Chemistry Chemical Physics* **8**, 1985–1993 (2006).
32. Goerigk, L. *et al.* A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Physical Chemistry Chemical Physics* **19**, 32184–32215 (2017).
33. Dohm, S., Hansen, A., Steinmetz, M., Grimme, S. & Checinski, M. P. Comprehensive thermochemical benchmark set of realistic closed-shell metal organic reactions. *Journal of Chemical Theory and Computation* **14**, 2596–2608 (2018).
34. Yu, W. *et al.* Extensive conformational searches of 13 representative dipeptides and an efficient method for dipeptide structure determinations based on amino acid conformers. *Journal of Computational Chemistry* **30**, 2105–2121 (2009).

35. Kishor, S., Dhayal, S., Mathur, M. & Ramaniah, L. M. Structural and energetic properties of α-amino acids: A first principles density functional study. *Molecular Physics* **106**, 2289–2300 (2008).
36. Selvarengan, P. & Kolandaivel, P. Potential energy surface study on glycine, alanine and their zwitterionic forms. *Journal of Molecular Structure: THEOCHEM* **671**, 77–86 (2004).
37. Császár, A. G. & Perczel, A. Ab initio characterization of building units in peptides and proteins. *Progress in Biophysics and Molecular Biology* **71**, 243–309 (1999).
38. Schlund, S., Müller, R., Grassmann, C. & Engels, B. Conformational analysis of arginine in gas phase–A strategy for scanning the potential energy surface effectively. *Journal of Computational Chemistry* **29**, 407–415 (2008).
39. Riffet, V., Frison, G. & Bouchoux, G. Acid–base thermochemistry of gaseous oxygen and sulfur substituted amino acids (Ser, Thr, Cys, Met). *Physical Chemistry Chemical Physics* **13**, 18561–18580 (2011).
40. Baek, K., Fujimura, Y., Hayashi, M., Lin, S. & Kim, S. Density functional theory study of conformation-dependent properties of neutral and radical cationic L-tyrosine and L-tryptophan. *The Journal of Physical Chemistry A* **115**, 9658–9668 (2011).
41. Floris, F. M., Filippi, C. & Amovilli, C. A density functional and quantum Monte Carlo study of glutamic acid in vacuo and in a dielectric continuum medium. *The Journal of Chemical Physics* **137**, 075102 (2012).
42. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **8**, 3192–3203 (2017).
43. Ropo, M., Schneider, M., Baldauf, C. & Blum, V. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Scientific Data* **3**, 1–13 (2016).
44. Huang, H., Li, D. & Cowan, J. Biostructural chemistry of magnesium. regulation of mithramycin-DNA interactions by Mg^{2+} coordination. *Biochimie* **77**, 729–738 (1995).
45. Romani, A. M. Cellular magnesium homeostasis. *Archives of biochemistry and biophysics* **512**, 1–23 (2011).
46. Forsen, S. & Kordel, J. Calcium in biological systems (1994).
47. Grauffel, C., Dudev, T. & Lim, C. Why cellular di/triphosphates preferably bind Mg^{2+} and not Ca^{2+}. *Journal of Chemical Theory and Computation* **15**, 6992–7003 (2019).
48. Mahmoud, W. E. Functionalized ME-capped CdSe quantum dots based luminescence probe for detection of Ba^{2+} ions. *Sensors and Actuators B: Chemical* **164**, 76–81 (2012).
49. Wilkinson, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 1–9 (2016).
50. Wittenburg, P., Lautenschlager, M., Thiemann, H., Baldauf, C. & Trilsbeek, P. FAIR practices in Europe. *Data Intelligence* **2**, 257–263 (2020).
51. Noy, N. F., *et al*. Ontology development 101: A guide to creating your first ontology (2001).
52. Wales, D. J. & Doye, J. P. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A* **101**, 5111–5116 (1997).
53. Wales, D. J. & Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **285**, 1368–1372 (1999).
54. Blum, V. *et al*. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications* **180**, 2175–2196 (2009).
55. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient O (N) integration for all-electron electronic structure calculation using numeric basis functions. *Journal of Computational Physics* **228**, 8367–8379 (2009).
56. Ren, X. *et al*. 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**, 053020 (2012).
57. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865 (1996).
58. Tkatchenko, A. & Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Physical Review Letters* **102**, 073005 (2009).
59. Swendsen, R. H. & Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters* **57**, 2607 (1986).
60. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **314**, 141–151 (1999).
61. Wong, M. A. & Hartigan, J. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**, 100–108 (1979).
62. Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoretica Chimica Acta* **44**, 129–138 (1977).
63. DiStasio, R. A., Gobre, V. V. & Tkatchenko, A. Many-body van der Waals interactions in molecules and condensed matter. *Journal of Physics: Condensed Matter* **26**, 213202 (2014).
64. Henkelman, G., Arnaldsson, A. & Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Computational Materials Science* **36**, 354–360 (2006).
65. Sanville, E., Kenny, S. D., Smith, R. & Henkelman, G. Improved grid-based algorithm for Bader charge allocation. *Journal of Computational Chemistry* **28**, 899–908 (2007).
66. Yu, M. & Trinkle, D. R. Accurate and efficient algorithm for Bader charge integration. *The Journal of Chemical Physics* **134**, 064111 (2011).
67. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993).
68. Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **5**, 129–145 (1984).
69. Fox, T. & Kollman, P. A. Application of the RESP methodology in the parametrization of organic solvents. *The Journal of Physical Chemistry B* **102**, 8070–8079 (1998).
70. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Antechamber: an accessory software package for molecular mechanical calculations. *J. Am. Chem. Soc* **222**, U403 (2001).
71. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**, 198–210 (2013).
72. O'Boyle, N. M. *et al*. Open Babel: An open Chemical toolbox. *Journal of Cheminformatics* **3**, 1–14 (2011).
73. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**, 1859–1865 (2008).
74. Eastman, P. *et al*. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology* **13**, e1005659 (2017).
75. Hu, X. & Baldauf, C. Cation-coordinated conformers of 20 proteinogenic amino acids with different protonation states. *NOMAD* https://doi.org/10.17172/NOMAD/2021.02.10-1 (2021).
76. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials* **2**, 036001 (2019).
77. Hu, X., Lenz-Himmer, M. O. & Baldauf, C. The ontology representation for a data set of cation-coordinated conformers of 20 proteinogenic amino acids with different protonation states. *EDMOND* https://doi.org/10.17617/3.5q (2021).
78. Al-Aswadi, F. N., Chan, H. Y. & Gan, K. H. Automatic ontology construction from text: a review from shallow to deep learning trend. *Artificial Intelligence Review* **53**, 3901–3928 (2020).
79. Musen, M. A. The protégé project: a look back and a look forward. *AI Matters* **1**, 4–12 (2015).

80. Lamy, J.-B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine* **80**, 11–28 (2017).
81. Tsarkov, D. & Horrocks, I. FaCT++ description logic reasoner: System description. In *International Joint Conference on Automated Reasoning*, 292–297 (Springer, 2006).
82. Wang, J. *et al*. Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *The Journal of Physical Chemistry B* **116**, 7088–7101 (2012).
83. Li, Y. *et al*. Machine learning force field parameters from ab initio data. *Journal of Chemical Theory and Computation* **13**, 4492–4503 (2017).
84. Cole, D. J., Vilseck, J. Z., Tirado-Rives, J., Payne, M. C. & Jorgensen, W. L. Biomolecular force field parameterization via atoms-in-molecule electron density partitioning. *Journal of Chemical Theory and Computation* **12**, 2312–2323 (2016).
85. Rai, B. K. & Bakken, G. A. Fast and accurate generation of ab initio quality atomic charges using nonparametric statistical regression. *Journal of Computational Chemistry* **34**, 1661–1671 (2013).
86. Bleiziffer, P., Schaller, K. & Riniker, S. Machine learning of partial charges derived from high-quality quantum-mechanical calculations. *Journal of Chemical Information and Modeling* **58**, 579–590 (2018).
87. Møller, C. & Plesset, M. S. Note on an approximation treatment for many-electron systems. *Physical Review* **46**, 618 (1934).
88. Head-Gordon, M., Pople, J. A. & Frisch, M. J. MP2 energy evaluation by direct methods. *Chemical Physics Letters* **153**, 503–506 (1988).
89. Ambrosetti, A., Reilly, A. M., DiStasio, R. A. Jr & Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *The Journal of Chemical Physics* **140**, 18A508 (2014).
90. Riplinger, C. & Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *The Journal of Chemical Physics* **138**, 034106 (2013).
91. Riplinger, C., Sandhoefer, B., Hansen, A. & Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *The Journal of Chemical Physics* **139**, 134101 (2013).
92. Supady, A., Blum, V. & Baldauf, C. First-principles molecular structure search with a genetic algorithm. *Journal of Chemical Information and Modeling* **55**, 2338–2348 (2015).
93. Hu, X. XiaojuanHu/AA_property_calculation: First release of AA_property_calculation. *Zenodo* https://doi.org/10.5281/zenodo.5672781 (2021).

Acknowledgements

X.H. is grateful for a doctoral fellowship by the China Scholarship Council. All authors acknowledge funding by the Federal Ministry of Education and Research of Germany for the project STREAM ("Semantische Repräsentation, Vernetzung und Kuratierung von qualitätsgesicherten Materialdaten", ID: 16QK11C).

Author contributions

X.H. performed the calculations of all conformers, curated the data, constructed the ontology, and wrote the manuscript. M.L. helped with the construction of ontology and contributed to the manuscript. C.B. designed the study, curated the data, and wrote the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.H. or C.B.