De *et al. J Cheminform (2017) 9:6*
DOI 10.1186/s13321-017-0192-4

Journal of Cheminformatics

RESEARCH ARTICLE

Open Access

Mapping and classifying molecules from a high-throughput structural database

Sandip De[1,2*], Felix Musil[1,2], Teresa Ingram[3], Carsten Baldauf[3] and Michele Ceriotti[1,2]

Abstract

High-throughput computational materials design promises to greatly accelerate the process of discovering new materials and compounds, and of optimizing their properties. The large databases of structures and properties that result from computational searches, as well as the agglomeration of data of heterogeneous provenance leads to considerable challenges when it comes to navigating the database, representing its structure at a glance, understanding structure–property relations, eliminating duplicates and identifying inconsistencies. Here we present a case study, based on a data set of conformers of amino acids and dipeptides, of how machine-learning techniques can help addressing these issues. We will exploit a recently-developed strategy to define a metric between structures, and use it as the basis of both clustering and dimensionality reduction techniques—showing how these can help reveal structure–property relations, identify outliers and inconsistent structures, and rationalise how perturbations (e.g. binding of ions to the molecule) affect the stability of different conformers.

Computational materials design promises to greatly accelerate the discovery of materials and molecules with novel, optimized or custom-tailored properties. With this goal in mind, several community efforts have emerged over the past few years [1–8] that aim at generating, and/or storing large amounts of simulation data in publicly available databases [9–15]. The development of these repositories of structural data, and of associated materials properties (e.g. formation energy, band gap, polarizability, ...) poses considerable challenges, from the points of view of guaranteeing consistency, accuracy and reliability of the stored information, as well as that of extracting intuitive insight onto the behavior of a given class of materials and of data-mining in search of compounds that exhibit the desired properties or that are somehow interesting or unexpected.

In order to automate these tasks—which is necessary to unlock the full potential of computational materials databases that can easily contain millions of distinct structures—a number of different machine-learning algorithms have been developed, or adapted to the specific requirements of this field [16–25]. A fundamental ingredient in all of these approaches is a concise mathematical representation of a molecular or crystalline structure, that can take the form of fingerprints (low-dimensional representation of the structure of the atoms) or more abstract measures of the (dis)similarity between elements in the database, such as distance or kernel functions.

In the present manuscript we will present a demonstration of how a very general approach to quantify structural dissimilarity [26] can be combined with non-linear dimensionality reduction and clustering techniques to address the challenges of navigating a database of molecular conformers, checking its internal consistency and rationalising structure–property relations. Even though we will focus in particular on a energy/structure data set of amino acid and dipeptide conformers obtained by an ab initio structure search [15, 27], many of the observations we will infer are general, and provide insight on the application of machine-learning techniques to the analysis of molecular and materials databases generated by high-throughput computations.

*Correspondence: sandip.de@epfl.ch
[2] Laboratory of Computational Science and Modelling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
Full list of author information is available at the end of the article

A toolbox for database analysis

Automatic analysis of atomistic structures obtained from large databases of materials and molecules requires a combination of different techniques (Fig. 1). A representation of structures in terms of "fingerprints", distances or kernels serves as the input of unsupervised-learning techniques (clustering, dimensionality reduction, ...) that greatly simplify the verification of the database for internal consistency, and the identification of organising principles and structure/property relations. Although we will not discuss this aspect explicitly here, molecular representations can also be used to directly predict properties using supervised learning techniques such as kernel-ridge regression or neural networks. In this section we will describe a specific combination of descriptors and unsupervised-learning algorithms, but we will also briefly summarize some of the alternative approaches that could be used to substitute different components of our tool chain.

Fingerprints and structural similarity

The most crucial and basic element in any structural analysis algorithm is to introduce a metric to measure (dis)similarity between two atomic configurations. Many options are available, with different levels of complexity and generality, starting from the commonly used root mean square (RMS) distance. In order to deal with symmetry operations or condensed phase structures, several "fingerprint" frameworks have been developed [8, 28–40], that assign a unique vector of order parameters to each molecular or crystalline configuration: a metric can then be easily built by taking some norm of the difference between fingerprint vectors. Any of these distances could be taken as the basis of the classification and mapping algorithms that we will describe in what follows.

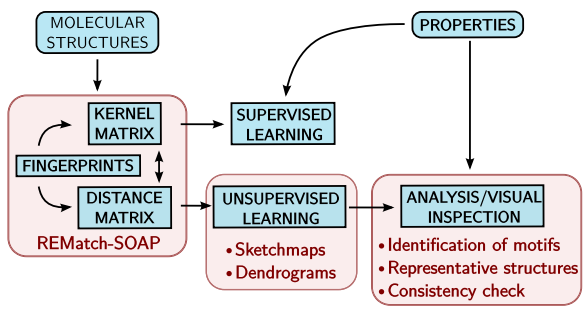

Fig. 1 A flowchart summarizing the different ingredients that enter a semi-automated workflow for the representation and analysis of a database of atomistic structures. We highlight with a *blue background* the general components, and with a *red background* the methods and applications we explore in the present work

In this paper we will use instead a very flexible framework (REMatch-SOAP) that is based on the definition of an environment similarity matrix $C_{ij}(A,B)$, which contains the complete information on the pair-wise similarity of the environment of each of the atoms within the molecules A and B. In our framework, the similarity between two local environments $\mathcal{X}_i^A$ and $\mathcal{X}_j^B$ is computed using the SOAP kernel [29]

$$C_{ij}(A,B) = k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right). \tag{1}$$

The REMatch kernel is then defined as the following weighted combination of the elements of $\mathbf{C}(A,B)$

$$\begin{aligned} \hat{K}^\gamma(A,B) &= \mathrm{Tr}\mathbf{P}^\gamma\mathbf{C}(A,B), \\ \mathbf{P}^\gamma &= \underset{\mathbf{P}\in\mathscr{U}(N,N)}{\mathrm{argmin}} \sum_{ij} P_{ij}\left(1 - C_{ij} + \gamma \ln P_{ij}\right), \\ C_{ij}(A,B) &= k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right), \end{aligned} \tag{2}$$

where the optimal combination is obtained by searching over the doubly stochastic matrices $\mathscr{U}(N,N)$ the one that minimizes the discrepancy between matching pairs of environments subject to a regularization based on the matrix information-entropy $E(\mathbf{P}) = -\sum_{ij} P_{ij} \ln P_{ij}$ [41]. Once a kernel between two configurations has been defined, it is possible to introduce a kernel distance

$$D(A,B) = \sqrt{\hat{K}^\gamma(A,A) + \hat{K}^\gamma(B,B) - 2\hat{K}^\gamma(A,B)}, \tag{3}$$

that we will use as the metric for representing and clustering structures from a database.

As discussed in Ref. [26], the choice of a SOAP kernel as the definition of an environment similarity provides at the same time great generality—it can be seamlessly applied to both molecules and solids—and elbowroom for fine-tuning—ranging from setting an appropriate cutoff distance to circumscribe an environment to the introduction of an alchemical similarity kernel that translates the notion that different chemical species can behave similarly with respect to the properties of interest.

Mapping the structural landscape of a database

The dissimilarity between the N atomic configurations in a database contains a large amount of information on the structural relations between the database items. However, this information is not readily interpretable, as it is encoded as a N^2 matrix of numbers. Several methods are available to process dissimilarity information into a form that can be understood more intuitively. A first approach involves building a low-dimensional "map", where each point corresponds to one of the structures in the database

and where the (Euclidean) distances between points represents the information on the pairwise dissimilarity matrix.

Several methods have been proposed over the years to solve this dimensionality reduction problem, starting from principal component analysis [42] and the equivalent linear multi-dimensional scaling [43], and proceeding to non-linear generalizations of the idea, such as ISOMAP [44], diffusion maps [45], kernel PCA [46]. In this manuscript, we will use sketchmap [22–24], a method in which one iteratively optimizes the objective function

$$S^2 = \sum_{ij} \left[F\left[D(X_i, X_j)\right] - f\left[d(x_i, x_j)\right]\right]^2, \tag{4}$$

that measures the mismatch of the dissimilarity between atomic configurations $D(X_i, X_j)$ with the dissimilarity (typically just the Euclidean distance) between the corresponding low-dimensional projections $\{x_i\}$. The procedure is very similar to multi-dimensional scaling, except for the appearance of the transformations F and f, which are non-linear sigmoid functions of the form:

$$F(r) = 1 - \left(1 + (2^{a/b} - 1)(r/\sigma)^a\right)^{-b/a}. \tag{5}$$

The non-linear transformation focuses the optimization of Eq. (4) on the most significant distances (typically those of the order of σ), and disregards local distortions (e.g. induced by thermal fluctuations or by incomplete convergence of a geometry optimization) and the relation between completely disconnected portions of configuration landscape. The maps that we report in this work will be labeled synthetically using the notation σ `-A_B-a_b`, where A and B denote the exponents used for the high-dimensional function F, a and b denote the exponents for the low-dimensional function f, and σ the threshold for the switching function. The choice of these parameters of the sigmoid functions are discussed in detail elsewhere [24]. In practice A, B, a and b have relatively small effect on the projection, and can be optimized and kept fixed for systems belonging to the same family. Since the structures we consider here are minimum-energy configurations, and there are no thermal fluctuations that should be filtered out, we set $A = a = 1$ (so that at short range the algorithm will still try to represent distances faithfully) and set the long-range exponents to $B = b = 4$. The parameter σ is the one to which sketchmap is most sensitive, and needs to be tuned for each system separately. To automate the process of building sketchmaps of large amount of subsets of the database, we have used a simple heuristic procedure for determining the value of σ automatically. Following the prescriptions in Ref. [24], we first compute the histogram of distances in the dissimilarity matrix of each molecular set, and detect the dissimilarity value (D_{max}) corresponding to the peak value of the histogram. We then set the value of σ to $0.8D_{max}$.

Hierarchical clustering representation

As we will demonstrate below, sketchmap provides a remarkably informative two-dimensional representation of structures in a data set, making it possible to identify groups of similar configurations, outliers, as well as to investigate structure–property relations. An alternative approach to navigate a set of structures based on the dissimilarity matrix is to use clustering algorithms, that identify groups of objects having similar properties to hint at the presence of recurring motifs underlying the behavior of the system.

A considerable number of clustering algorithms have been developed over the last few decades [17, 47, 48], including connectivity models [49] (i.e. hierarchical clustering), centroid models [50–52] (i.e. k-means algorithm) and density based models [16, 53, 54].

Clustering models based on connectivity information such as hierarchical (or agglomerative) clustering [49] are particularly suited for this purpose and we will focus only on this type of clustering in this paper. Starting from each configurations as its own cluster, the hierarchical clustering algorithm iteratively aggregates clusters together based on some assessment of their dissimilarity. Dissimilarity between two individual structures can be obviously measured by their distance $D(A, B)$. The distance between two *clusters*, however, can be defined in many different ways. In our study, we will use in particular the RMS dissimilarity between the pair of members of the two clusters. The linkage distance Δ between two clusters $\mathbb{X} = \{X_i\}$ and $\mathbb{Y} = \{Y_i\}$ is then defined as:

$$\Delta(\mathbb{X}, \mathbb{Y}) = \sqrt{\frac{1}{N_{\mathbb{X}} N_{\mathbb{Y}}} \sum_{X \in \mathbb{X}, Y \in \mathbb{Y}} D^2(X, Y)}, \tag{6}$$

where $N_{\mathbb{X}}$ and $N_{\mathbb{Y}}$ are the total number of configurations within each cluster. $D(X, Y)$ is the dissimilarity between the two configurations, that in our case was computed based on the REMatch-SOAP kernel. The complexity of this type of clustering in terms of the number of structures N is relatively cheap ($\mathcal{O}(N^2 log(N))$) compared with dimensionality reduction algorithms like sketchmap ($\mathcal{O}(N^3)$). Both procedures can be greatly accelerated through out of sample embedding. A subset of the configurations is selected (e.g. by farthest point sampling, with the possibility of weighting based on density information [24]) for either dimensionality reduction or hierarchical clustering and then is used as a reference for the projection/clustering of the other structures.

The results of a hierarchical clustering procedure can be represented in a "dendrogram" plot, that conveys visually the sequence of agglomerative clustering operations and the linkage distance at each step. The lowest level of the dendrogram is composed of single-structure clusters, so that the x axis corresponds to individual configurations sorted according to the clustering procedure. Each merge operation is represented by a line joining the two underlying clusters, with the y position of the line representing the linkage distance for that pair, as defined by Eq. (6). In this kind of representation, at the bottom of the dendrogram, each structure can be thought of as an individual cluster containing only one item. Clusters are then merged iteratively, selecting at each step the pair of clusters that are closest to each other. This operation is repeated until all the clusters collapse into one single group that encompasses all the structures in the database, thus completing the dendrogram. To avoid overcrowding the bottom of the plot, one can hide the part that corresponds to very small linkage distances, while still graphically visualising the size of the clusters by drawing bars that encompass the associated structures. Since the "leaves" of this dendrogram correspond to individual configurations, it is possible to complement the dendrogram with color-coded bar plots that represent the value of different properties of each structure, thereby giving a clear picture of the relation between structural clustering and the different properties.

In order to understand the basic motifs of a particular cluster $\mathbb{X}$, it is very useful to select one of its structures that is as representative as possible of the entire subset. In case where stability estimates are available, such structure may be the lowest-energy structure in the cluster. For a definition that is based purely on conformational or configurational information, the most representative structure $RS(\mathbb{X})$ could be defined as the item having the minimum mean square dissimilarity with respect to all other members of $\mathbb{X}$, i.e.

$$RS(\mathbb{X}) = \underset{X_1 \in \mathbb{X}}{\operatorname{argmin}} \left[\frac{1}{N_{\mathbb{X}}} \sum_{X_2 \in \mathbb{X}} D^2(X_1, X_2) \right]. \tag{7}$$

Representative structures can be defined at each level of the hierarchy, and can therefore be very useful in navigating the database, and understanding what are its most crucial structural features. The spread of the cluster around $RS(\mathbb{X})$,

$$\sigma_D(\mathbb{X}) = \sqrt{\frac{1}{N_{\mathbb{X}}} \sum_{X \in \mathbb{X}} D^2(X, RS(\mathbb{X}))}, \tag{8}$$

can be used to quantify the range of structural landscape that is covered by the cluster.

Another important aspect of database analysis is outlier detection [55–60]. An outlier configuration is defined as a configuration which is different from most of the configurations in the database. Outlier configurations are very important as they are likely to have unique structural motif in the whole database and are thus interesting for structure prediction applications. They also could represent chemical changes or indicate inconsistent configurations which are likely to be "errors" in the database.

In the following sections, we will present examples of how these different analyses can be applied to different subsets of structures taken from a database of amino acid and dipeptide conformers.

Analysis of a database

This work is based on a first-principles derived structure/energy data set with conformers of twenty proteinogenic amino acids and dipeptides, as well as their interactions with a series of divalent cations [27] (Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+}, Hg^{2+}). The potential-energy surfaces (PES) of 280 systems were explored 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 [15]. The underlying energetics were calculated by applying density-functional theory (DFT) in the generalized gradient approximation corrected for long-range van der Waals interactions [61–63] (PBE + vdW). A number of theory-theory and theory-experiment comparisons have shown the applicability of the method to amino acid and peptide systems [15, 64–69]. The generation of this dataset involved significant manual intervention, and one would expect it to be an easy starting point for studying materials and molecules across chemical space [70]. Nevertheless, we will demonstrate that, even for such heavily curated data, automated techniques are needed to extract unbiased and hypothesis-free trends and to check for internal consistency.

In this study we focus on the amino acid lysine (in short Lys) and investigate basic structural motifs of three forms, see Fig. 2. Furthermore, the machine learning techniques introduced in this work are used to detect the impact of perturbations (here Ca^{2+} cations) on the structural properties of the unperturbed systems. Finally, we demonstrate how the approach can also be applied to discover inconsistencies and outliers in the database. Hierarchical classifications and sketchmap projections for all the proteinogenic amino acids in the database are given in the Supporting Information.

Finding the dominant features of a structual landscape

Lysine dipeptide

We take as our first example a subset of the database containing 2080 conformers of lysine dipeptide. As discussed

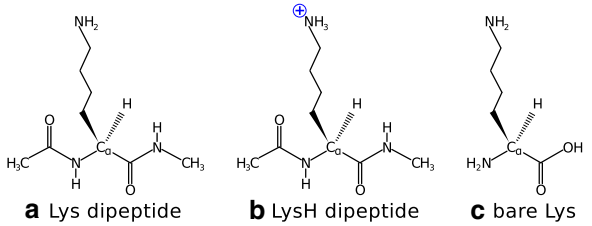

Fig. 2 The lysine building block was studied in three forms: **a** uncharged dipeptide, **b** protonated dipeptide, and **c** uncapped and uncharged amino acid

in the previous section, we start by constructing the (dis) similarity matrix using the SOAP-REMatch kernel. In Fig. 3 the dendrogram plot as well as sketchmaps have been shown along with five properties, energy and four dihedral angles, using the same color scales in both the sketchmap and dendrogram representations. In the sketchmap each circular 'disk' represents a conformer. Whereas in the case of the dendrogram plot, structures are represented by vertical lines at the bottom of the plot. The strong correlation between energy and conformational parameters on one side, and clustering and position on the map on the other, testifies how the the REMatch-SOAP kernel induces a meaningful classification of the structures in this dataset.

While both clustering and sketchmap show clearly that the dataset is composed of groups of structurally-related conformers, the agnostic nature of the underlying metric does not disclose immediately the structural features that most transparently differentiate between different clusters. Comparing the representative structures from the main clusters allowed us to quickly identify candidate structural motifs that could be used to rationalize the layout of the conformational landscape. By color-coding the dendrogram and the sketchmaps according to these indicators one can readily highlight the key correlations. When considering existing literature on the stability of oligopeptides, the two structural parameters that are most often considered as the key coordinates to navigate the conformational landscape are the Ramachandran dihedral angles ϕ and ψ, that determine the structure of the backbone around the side chain bearing Cα atom [71] under the assumption of peptide bonds being solely in *trans* conformation. While fine-grained clusters are homogeneous with respect to the ϕ and ψ angles, it is clear that for the present systems the clear-cut branching at the top of the dendrogram is determined by some other order parameter. An analysis of the representative structures for the two main clusters (a) and (b) shows that the two molecules differ by the isomerization of the N-terminal peptide bond. Further splitting of these two clusters, i.e. (a) into clusters (d) and (e), and (b) into (f) and (g), depends on the isomerization of the C-terminal peptide bond. We can confirm this attribution of the main features of the dataset by color-coding the map and the dendrogram following the dihedral angles ω_1 and ω_2. The four main clusters are largely homogeneous with respect to peptide bond isomerization, and are then further subdivided based on ϕ and ψ. This observation deserves some further comment. Peptide bonds in naturally-occurring proteins are believed to almost exclusively exist in *trans* conformation with the exception of prolyl peptide bonds where a smaller energy difference to *trans* increases the chance for *cis* conformers [72, 73]. This view is supported by the analysis of protein structures deposited in the protein databases where *cis* conformations are found for about 5% of the prolyl peptide bonds, but less than 0.1% for the others [74]. X-ray crystallographic structure represent however merely frozen snapshots of structural dynamics. The ab initio structure search protocol, instead, does consider the peptide bond torsions as variable and intentionally allowed simulations to overcome the isomerization barrier. Consequently, the dataset contains representatives of all four combinations of *cis* and *trans* conformers. Since these transitions are strongly bimodal, and reflect in significant changes of the favorable side chain conformations, they constitute the most significant feature to classify the conformers. As expected, the most stable conformers are largely in a *trans–trans* conformation. However, the large parts of conformational space that is occupied by conformers with 1 or 2 *cis* peptide bonds suggests that *cis* isomers might play a role in the dynamics of peptides and proteins. Consequently, an analysis only focused on the Ramachandran dihedrals, ϕ and ψ, would have missed one of the main features of the structural landscape that is critical to characterize the relation between structure and energetics. One could then proceed further with the analysis, focusing for instance on small clusters containing low-energy structures such as that represented by the conformer (c). All the structure in this group are *trans–trans* isomers, that in addition have $\phi \approx -90°$ and $\psi \approx 90°$, allowing for the formation of a H-bond between the side chain N_3 and H_1, and a favorable arrangement of the N_2 donating a H-bond to the carbonyl O_1 as shown in Fig. 3. Having access to the combined information on energetics, and on the grouping of structures with similar geometry makes it easier to rationalize the energy ordering of the structures, without having to separately juxtapose all the low-lying conformers but focusing on a few representative configurations.

Protonated lysine dipeptide

As the second example we considered a dataset containing 897 conformers of gas phase protonated lysine dipeptide.

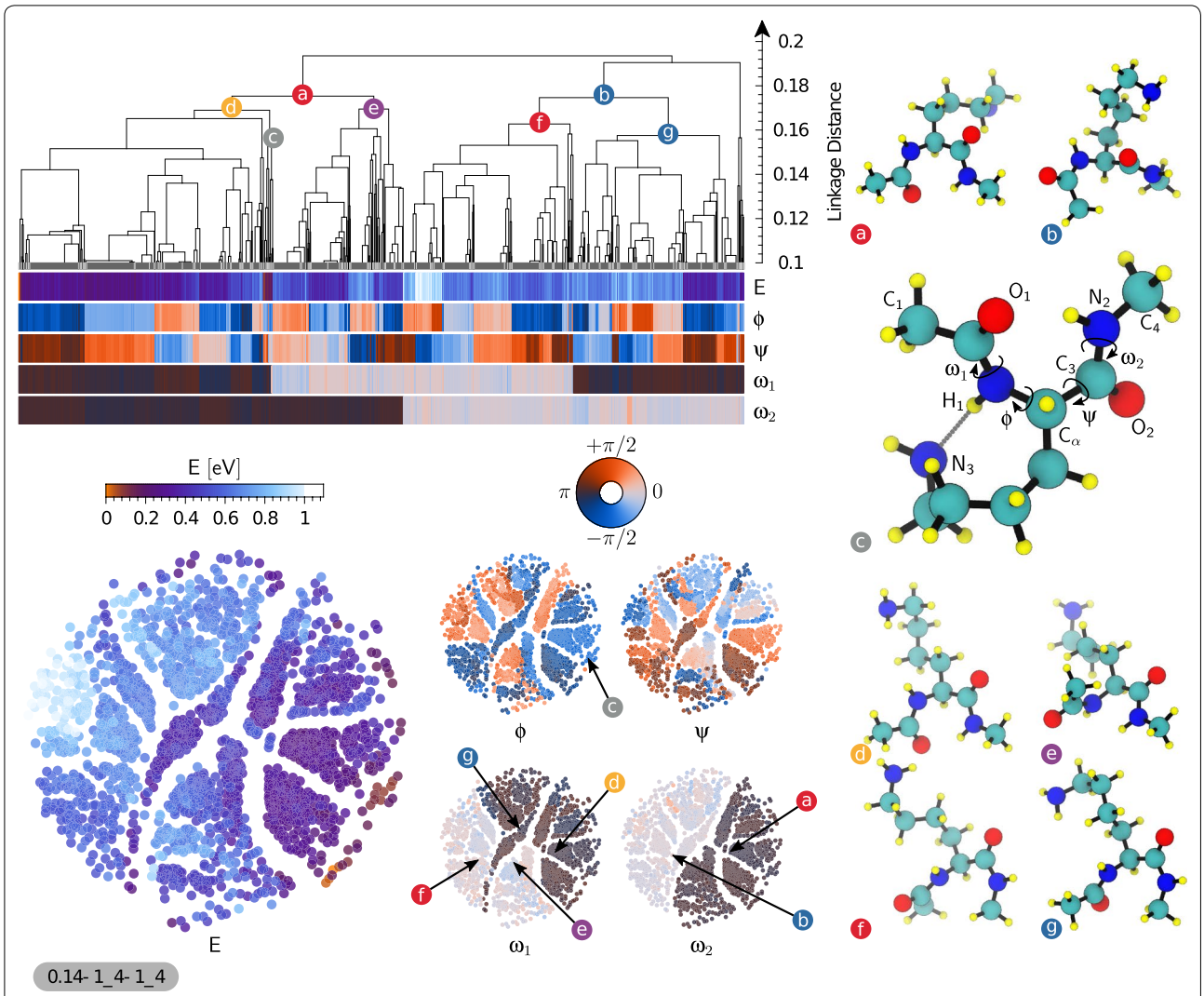

Fig. 3 Representation of the similarity matrix corresponding to the lysine dipeptide dataset using the agglomerative clustering algorithm (*top*) and the sketchmap algorithm (*bottom*, projection parameters shown following the scheme σ–*A_B*–*a_b*). A few representative structures (see Eq. (7)) of interesting clusters are shown (*right*) and their corresponding position on the sketchmaps and dendrogram representation is *highlighted*. The five sketchmaps are colored according to the conformational energy and the backbone dihedral angles ϕ, ψ, ω_1 and ω_2. The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using *color bars* below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as *thick gray bars* separated by *light-gray lines*. Clusters composed of only one structure are drawn as a *black line* reaching the bottom of the dendrogram. The main structural motifs of this set of structures are governed by the peptide bond dihedral angles ω_1 and ω_2. The two main clusters **a**, **b** are showing a global correlation with the angle ω_2 while the angle ω_1 splits them into two well correlated sub-clusters (**d**)–(**g**) respectively. The cluster **c** is highlighted as an example containing 'outlier' structures of low conformational energy

We follow the same steps as described in the previous example in order to find the most basic structural motifs of this system. Figure 4 shows the dendrogram, the sketchmap and a few color coded properties of this system to demonstrate their correlation with the classification. The most prominent feature for this molecule, which is evident in both the dendrogram and the sketch maps, is the presence of a group of outliers, that are clearly separated from the bulk of the conformers. Inspection of the cluster centroid (g) clarifies the structural basis of this separation. Whereas in most of the structures the excess charge lies on the lysine side chain as a NH_3^+ group, conformers in this cluster experienced a proton transfer event, with the excess proton attached to one of the carbonyl oxygen O_1, stabilized by H-bonding to N_2. This is a result of the database generation where ab initio replica-exchange molecular dynamics including high T trajectories where used for structure sampling during which protons can eventually transfer.

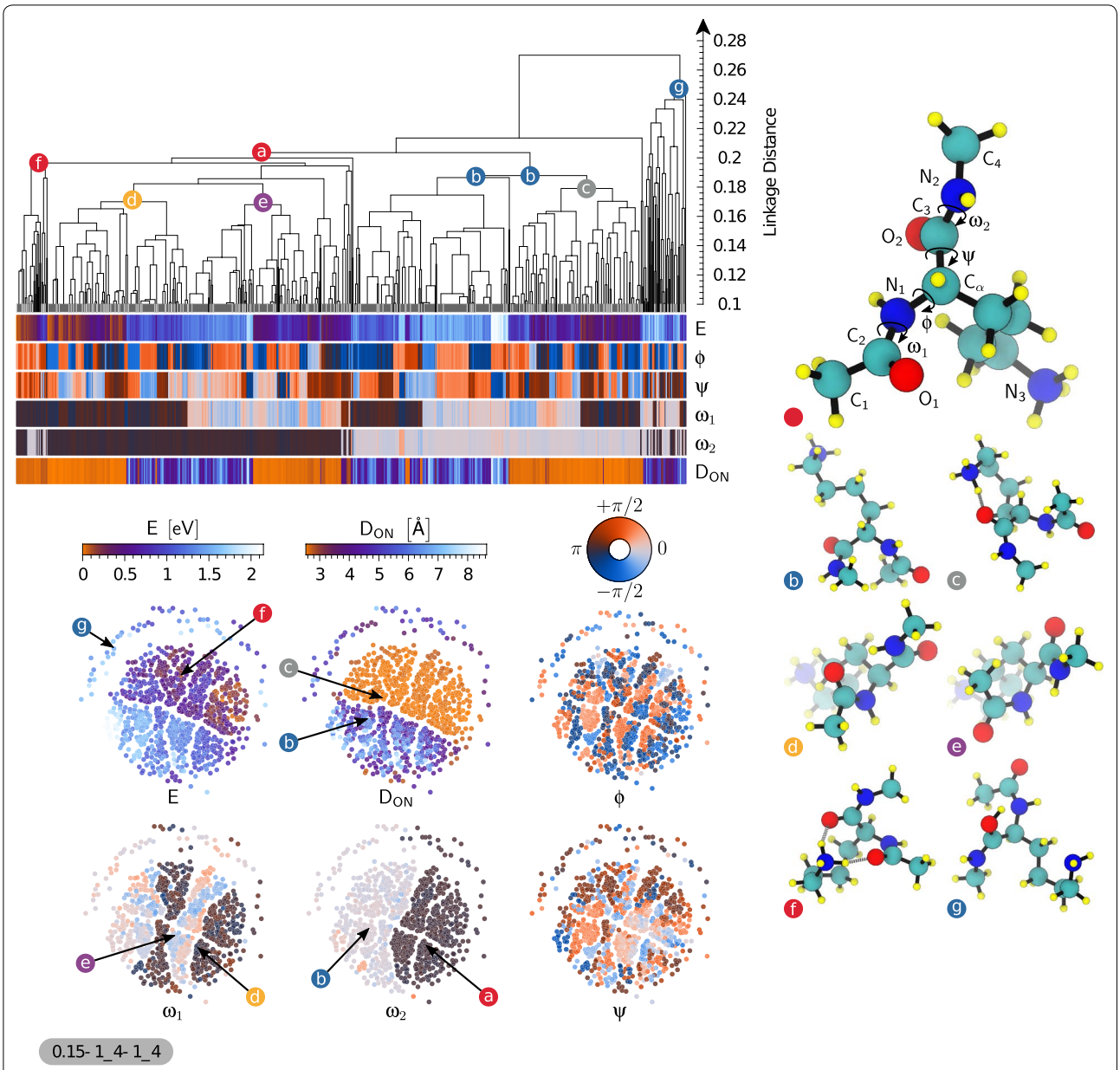

Fig. 4 Representation of the similarity matrix corresponding to the protonated lysine dipeptide dataset using the agglomerative clustering algorithm (*top*) and the sketchmap algorithm (*bottom*, projection parameters shown following the scheme *σ–A_B–a_b*). A few representative structures (see Eq. 7) of interesting clusters are shown (*right*) and their corresponding position on the sketchmaps and dendrogram representation is *highlighted*. The six sketchmaps are colored according to the conformational energy, the minimal distance between O_1 or O_2 with N_3 called D_{ON}, and the backbone dihedral angles ϕ, ψ, ω_1 and ω_2. The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using *color bars* below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as *thick gray bars* separated by *light-gray lines*. Clusters composed of only one structure are drawn as a *black line* reaching the bottom of the dendrogram. The main structural motifs of this set of structures are governed by the dihedral angles ω_1 and ω_2 and the distance D_{ON}. The two main clusters **a**, **b** are showing a global correlation with the angle ω_2 while the angle ω_1 splits them into well correlated sub-clusters (e.g. sub-clusters **d**, **e**). The other important sub-clustering parameter is the distance D_{ON}, e.g. sub-clusters (**c**) and (**b**), which also correlates well with the separation between low and high conformational energy shown on the sketchmaps. Two sub-clusters are particular: **g** is a clear 'outlier' due to a chemical change and **f** features a H-bonding pattern with the side chain NH_3^+ pointing to both carboxy groups that sets this cluster apart from all others

Moving on to the main cluster of structures, we can see that similar to our previous example of the neutral dipeptide and again due to the unbiased sampling protocol and the high energy range the peptide bond angles are again more important than Ramachandran's dihedrals. Conformers (a) and (b) are the representative structure for groups having *cis* and *trans* ω_2 peptide bonds respectively. Group (a) is further split based on the *cis/trans* state of ω_1 into the clusters represented by structures (d) and (e).

The presence of a charged side-chain leads to stronger H-bonds. As a consequence, peptide-bond isomerism plays a less crucial role in determining structural clustering than for the neutral dipeptide. An example of the importance of H-bonds is given for instance by the subcluster represented by conformer (f), in which the bent side chain leads to the formation of two H-bonds between NH_3^+ group and the carbonyl oxygens. H-bonds also dominate the partitioning of cluster (b), that is split into two groups—one of which is still best represented by the same conformer, and one that is epitomised by (c). Once again, inspection of these structural representatives reveals the organising principle behind the classification: (c)-like structures have an extended side chain, and are dominated by interactions among the peptide bond moieties, whereas (b)-like structures have a well-formed H-bond between the side chain and one of the two backbone O atoms. This structural pattern can be emphasized by color-coding conformers based on the parameter $D_{ON} = \min[d(O_1, N_3); d(O_2, N_3)]$: A small O-N distance indicates bending of the side chain and the formation of a H-bond between O and N. As it is clear from the sketchmap representation, there is a very strong correlation between the bending of the charged side chain and the energy of a conformer. All of the structures within 0.5 eV of the ground state feature this sidechain to backbone H-bonds.

It is worth noting that the importance of intramolecular H-bonds is a consequence of the gas-phase environment in which the structure search was performed. In a polar solvent like water, where intramolecular H-bonds that introduce strain compete with H-bonds with the surrounding water molecules, that do not require a bending of the side-chain, the energy balance might be different or less clear-cut. The analysis techniques we introduce in this work would be ideally suited to rationalize the changes in the (free) energetics of biological molecules when moving from the gas phase to (micro)solvated environments or to organic/inorganic interfaces.

Uncapped lysine

Our third example is a dataset containing 733 conformers of the un-capped lysine molecule in the gas phase. We follow the same steps as described in the previous examples to construct the dendrogram shown in Fig. 5. The map has a simple structure, with few well-separated groups. Being a smaller molecule with fewer degrees of freedom, the Ramachandran angles are not defined. The dihedral angles in the vicinity of the C_α atom display local structural correlation but once again they are not the main organizing factor that can rationalize the clustering. By juxtaposing representative conformers from the main clusters we could identify a better order parameter, that correlates strongly with H-bond patterns within the molecule. Namely, the distance (D_H) between the H atom in the OH group of the carboxyl function and the N atom in the backbone (N_1) discriminates well between structures based on H-bonding patterns [70] of *type I* between $N_1H \rightarrow O_2$ (e.g. conformer (b)) and of *type II* with a H-bond $O_1H \rightarrow N_1$ (e.g. conformer (a)). It can be seen from both the dendrogram and the sketchmaps that one could identify several subgroups based on particular values of D_H, representing specific orientations. Conformers (c) and (d) represent small groups of conformers having specific relative orientation between the OH and NH_2 groups. Conformer (e) is representative of a small outlier group with a well-defined bend of the side chain, leading to the formation of a further H-bond between the N_1 atom in the amino acid moiety and N_2, in the side chain. The lysine side chain is very flexible, and the distance between N and C_α only plays a role in defining the fine-grained structure of the dataset, but is minimally correlated with the most prominent features.

While it appears that even in this case we could identify the basic structural motifs that characterize the conformational landscape of this system, the correlation with energy is very poor. There are several instances, in both the dendrogram and the sketchmap, where two conformers that are detected as structurally very similar display very different stability. Understanding whether this inconsistency signals a problem with our analysis brings us to the topic of outlier detection and consistency checks, that we will discuss in details in the "Identifying outliers and checking for consistency" section.

Understanding the impact of perturbations on conformational space

Having elucidated the essential structural motifs that underlie the organization of a set of molecular conformers, one could also wonder how changes in the thermodynamic conditions, or other external perturbations such as solvation, the addition or subtraction of an electron [75] or that of an atom [76–78], modify the conformations of the molecule and their stability. In addition to bare oligopeptides, the database [15, 27] that we are using as an example contains sets of locally-stable conformers in the

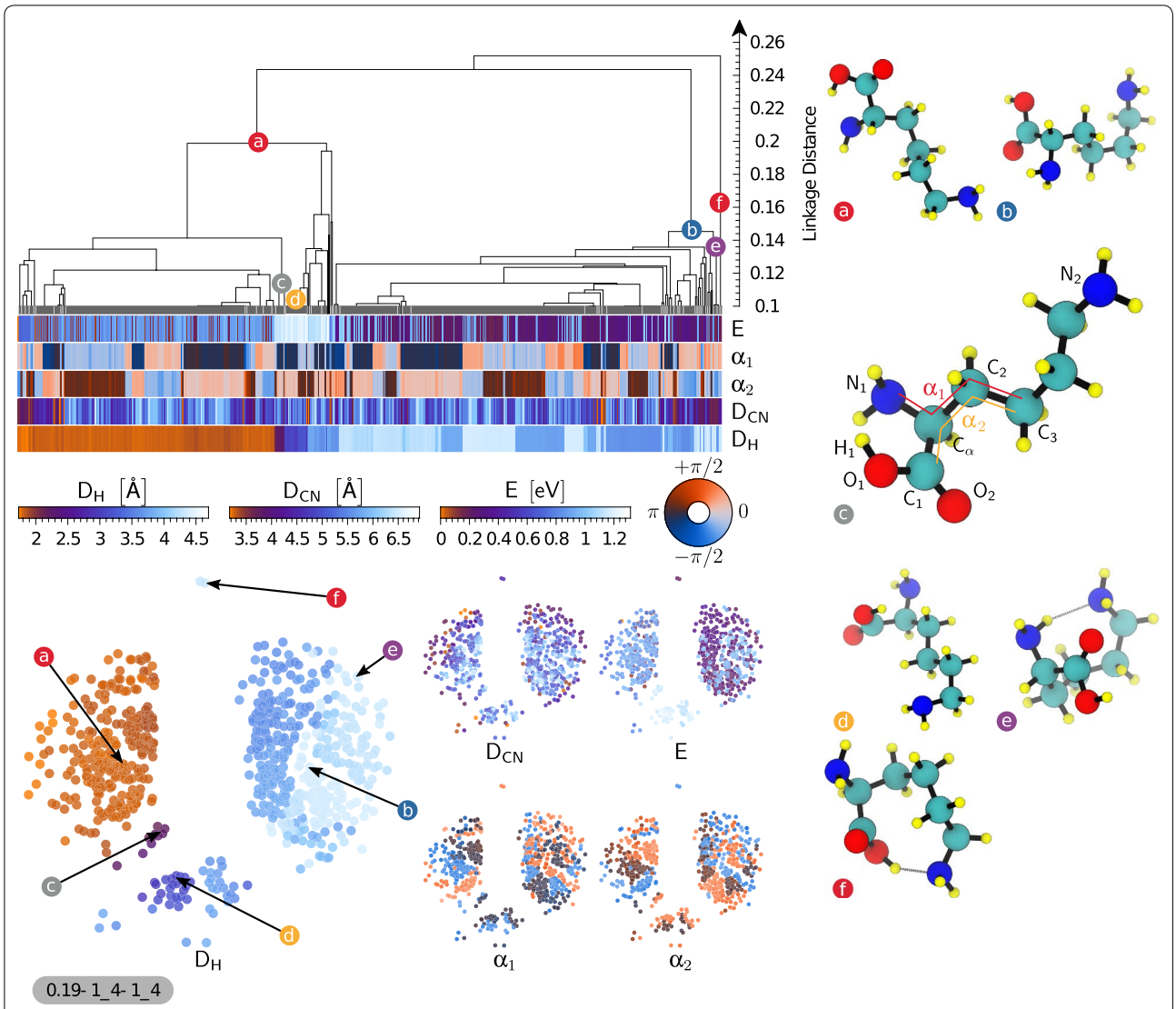

Fig. 5 Representation of the similarity matrix corresponding to the lysine uncapped dataset using the agglomerative clustering algorithm (*top*) and the sketchmap algorithm (*bottom*, projection parameters shown following the scheme σ–*A_B–a_b*). A few representative structures (see Eq. 7) of interesting clusters are shown (*right*) and their corresponding position on the sketchmaps and dendrogram representation is *highlighted*. The five sketchmaps are colored according to the conformational energy, the distance between N_1 and the hydrogen in the carboxilic group H_1 (labelled D_H), the distance between N_2 and C_α (labelled D_{CN}), and the dihedral angles α_1 and α_2 which are respectively computed with the following atoms (N_1, C_α, C_2, C_3) and (C_1, C_α, C_2, C_3). The dendrogram shows the clustering hierarchy of the structures of the dataset. Each structure is vertically aligned with its properties shown using *color bars* below the dendrogram. The dendrogram is cut at a linkage distance of 0.1 since structural properties are very similar below this threshold, and the clusters that are merged at this level are shown as *thick gray bars* separated by *light-gray lines*. Clusters composed of only one structure are drawn as a *black line* reaching the bottom of the dendrogram. The main structural motifs of the database are governed by the distance D_H. The two main clusters **a**, **b** are agglomerated according to the orientation of H_1 and the oxygen atom it is bonded to with respect to N_1 which is well described by the distance D_H. The sub-cluster **e** is composed of 'outlier' structures showing an H-bond between N_2 and an hydrogen of N_1 resulting in a folded side chain structural motif. Finally, the outlier cluster **f** contains a H-bond between the carboxy H and the side-chain NH_2, that can be seen as a precursor to the zwitterionic form

presence of cations of six different species, namely Ca^{2+}, Ba^{2+}, Sr^{2+}, Cd^{2+}, Pb^{2+} and Hg^{2+}. We consider the case of Ca^{2+} to describe how one can characterize its impact on the conformational space of the three molecular systems that we have discussed in our previous examples. We start by calculating the dissimilarity of all the conformers containing cations with their pure counterpart. In order to make the comparison on the same footings, we did not include the location of the cation in defining the SOAP kernels, so that the presence of Ca^{2+} only enters by

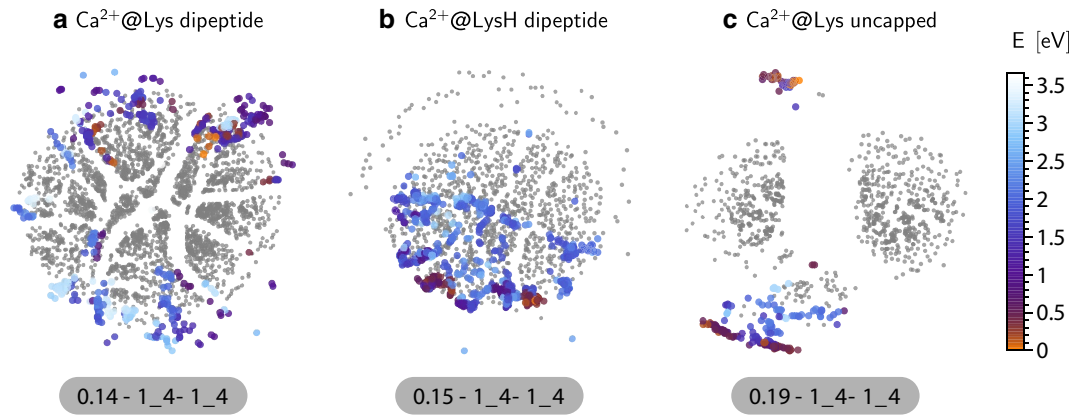

Fig. 6 The out-of-sample embedding of conformers with Ca^{2+} ion on the sketchmap of their pure counterpart, for the three systems we discussed in above: lysine dipeptide (**a**), protonated lysine dipeptide (**b**) and molecular lysine (**c**) systems. The projected conformers are colored with their energy where as the sketchmap on which they are projected are kept all in *grey color*. The location of the projected conformers allows us to understand how the conformational space of the pure conformers are affected due to presence of the Ca^{2+} ion

distorting molecular geometries and/or altering their relative stability. Using this information, we then projected the cation-containing dataset on the top of the sketchmap of structures for the bare molecule. This is done using sketchmap out-of-sample embedding, and we refer our reader to see the relevant literature [22–24] for more details about the method. In Fig. 6 we show the resulting projection, colored according to the stability of the conformers, on top of the sketchmap of the pure molecule shown in grey color as a reference. A close proximity of projected conformers with a pure conformer signifies their structural similarity. Segregation of the projected conformers with the cation in some area of the reference sketchmap represents the structural bias introduced by the strong electrostatic interaction with Ca^{2+}.

In the case of neutral lysine dipeptide (Fig. 6a), the presence of the Ca^{2+} ion induces relatively small distortions of the stable conformers, that get pushed towards the outer region of the map but are still clearly related to the locally stable structures for the bare molecule. Energies are dramatically changed, with the most stable cluster in the original map being completely absent in the presence of the cation. These observations highlight the importance of sampling high-energy conformers during high-throughput structure searches, since the relative stability can be modulated strongly by external perturbations. In particular, *cis* conformers become energetically more competitive and are topologically closer to the global minima. In the case of protonated lysine dipeptide (Fig. 6b), the same analysis shows an even clearer pattern. All the conformers with Ca^{2+} ions are projected in the lower part of the sketchmap, that correspond to conformers with an extended side chain (see Fig. 4). The Ca^{2+} ion preferably binds to the peptide O atoms, and the electrostatic repulsion with the protonated lysine residue strongly favors extended conformers, contrary to what we observed in the case of the bare molecule. Finally, one sees that for molecular lysine the addition of Ca^{2+} leads to conformers with very different structural motifs from those seen with the bare molecule, which is apparent in the sketchmap projection being concentrated far away from the unperturbed conformers (Fig. 6c). In fact, inspection of the structures shows that Ca^{2+} often triggers the transition to the zwitterionic form, with the cation coupled to the carboxylate group, and the protonated side chain NH_3^+ extending as far as possible away from it. In analogy with what was observed for Lennard-Jones clusters [24] and solvated polypeptide segments [79], sketchmap proved to be a powerful tool to analyze the response of the system to external perturbations and changes in the boundary conditions, and—in this specific example—to draw connections between different subsets of a high-throughput molecular database.

Identifying outliers and checking for consistency

The tools we introduced in this work are useful to address other important issues in data-driven science, such as outlier detection and consistency checks. We have already discussed the importance of detecting groups of outlier structures that are very different from the bulk of the dataset. These unusual items often signal the occurrence of unexpected effects that go beyond the original goal of the database construction effort. In the case of protonated lysine dipeptide, looking for outliers allowed us to reveal the presence of conformers with different chemical connectivity, or of strong H-bonds between

the backbone and the charged side chain. Similar observations can also be made in the case of the bare lysine molecule (Fig. 5). Moreover, one can observe a branch at the topmost level of the dendrogram, containing only two conformers. These are the only two cases where a H-bond is formed between the N of the side chain and the H atom of the OH group in the backbone. In the sketchmap, these two conformers are projected on the top, clearly isolated from rest of the groups, and bear the most resemblance to the zwitterionic conformers that are stabilized in the presence of a divalent cation. Obviously, the definition of a group of "outliers" can be more nuanced, and refer to small groups of structures appearing at deeper levels in the hierarchy. Overall, the possibility of clustering together the structures from a large dataset and inspecting a few representative conformers, rather than hundreds or thousands, greatly facilitates the task of identifying trends and spotting interesting or unexpected structures.

Outliers can signal interesting or important trends, but can also be a red flag for the presence of errors. The importance of database integrity has long been recognized by computer scientists [80–83], and several tools are available to monitor and correct inconsistencies from the technical point of view, in terms of reliability of storing and retrieving data [55–60]. The issue is also crucial when it comes to the maintenance of automatically-generated databases, and to repositories that store data of heterogeneous provenance [1–7]. In these cases, problems have generally little to do with the integrity of the storage, but rather with the consistency of the simulation details of different sets of calculations. Rather, inconsistencies should manifest themselves in the presence of structures that are geometrically very similar, but are associated to very different values of particular properties.

For example the lysine molecule dataset shows signs of this kind of issues, with energies that vary wildly within clusters that are very homogeneous in structure. This problem can be seen from the maps, i.e. when comparing the energy-colored sketchmap in Fig. 5 to the respective maps for the other systems. However, a more robust and easy-to-automate approach to identify structure/property inconsistencies starts from the hierarchical clusters, and compares the structural variability within each cluster σ_D (Eq. 8) with the variance of a given property, in this case energy, σ_E. Looking, for example, at a glassy energy landscape [84], one can observe configurations that are very different from a structural point of view, but have similar energy, giving rise to clusters with large σ_D and small σ_E. The data points in Fig. 7 each represent individual clusters of lysine dipeptide and uncapped lysine, respectively, and illustrate their variation in energy and structure. In the case of lysine dipeptide (Fig. 7a) one sees a clear correlation between the structural and energetical variation of the clusters. The two quantities σ_D and σ_E are not necessarily strongly correlated, but in general clusters that contain very similar structures also have a low spread in energy. For uncapped lysine (Fig. 7b), however, one can identify a group of points (which we manually highlighted in orange for clarity) that has a distinctively different behavior, with σ_E converging to a constant value other than zero as σ_D decreases. This kind of feature indicates that the metric based on which structures were classified cannot detect one specific effect that has a dramatic impact on energetics, signaling either a failure of the metric or, as in this case, an inconsistency in the generated data. Further investigation of the lysine molecule dataset revealed that a subset of structures that had been generated at a lower level of theory in the initial stages of the structure-search procedure made their way by mistake into the final dataset. Using this measure of cluster homogeneity on all systems of the amino acid database (see Supporting Information) revealed similar problems also for other molecules, for example Cys, Glu, and Arg. Thanks to this analysis we will be able to identify and rectify mistakes in all the affected datasets and subsequently update the on-line repository [27].

Conclusion

The increasing use of high-throughput computational screening of materials and molecules, and the compilations of curated databases of the resulting structures and properties, is making more and more urgent to adapt "big data" techniques to the problems that are specific to this field. In this work we have demonstrated how a toolbox of algorithms ranging from hierarchical clustering to non-linear dimensionality reduction can be used to navigate molecular databases, taking as a paradigmatic example some subsets of a database of oligopeptide structures in the gas phase. The software that was used to compute similarity data between molecules, as well as to generate dendrograms and sketch-maps, are open-source and available for download [85, 86].

We find that the use of REMatch-SOAP, a general and unbiased metric to compare different structures based on the combination of pair-wise similarity between molecular environments, makes these techniques particularly insightful. Rather than simply reflecting preconceived notions of what would be the key structural parameters to differentiate molecular conformers, this metric reveals for instance the importance of peptide bond isomerization in describing the high-energy portion of conformational space of oligopeptides, the possibility of changes in chemical connectivity in the course of the ab initio structural search, and the interplay between hydrogen-bonding,

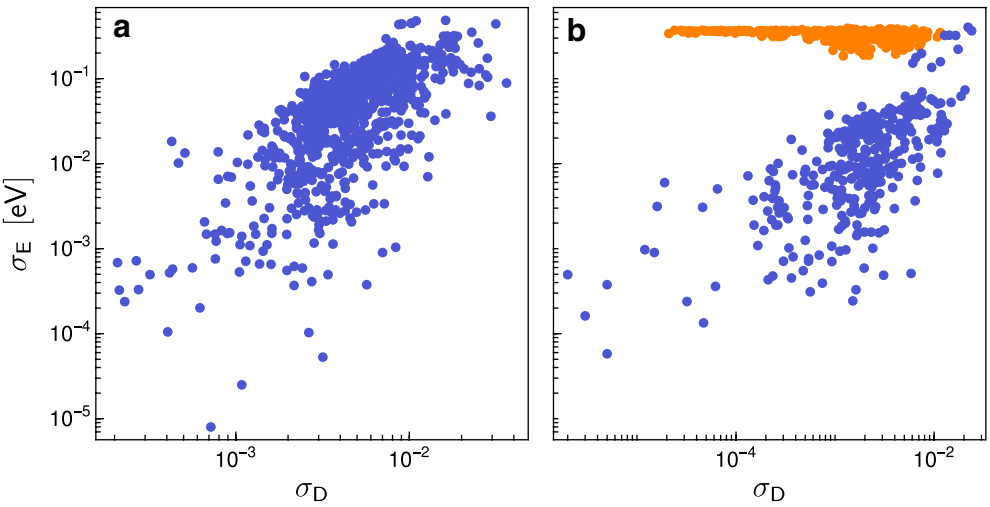

Fig. 7 This figure compares the homogeneity of clusters from the protonated lysine dipeptide (**a**) and the bare lysine uncapped (**b**) with respect to properties of their elements. The homogeneity of a cluster is probed using the standard deviation with respect to the distance between each cluster elements, σ_D, and the conformational energy, σ_E. The outliers of uncapped lysine (**b**) were manually highlighted in *orange*

backbone dihedrals an electrostatic interactions. Sketch-maps and hierarchical clustering proved to be complementary tools, with representative structures from the main clusters providing an easy way to compare visually different groups of conformers, and the low-dimensional map providing a quick, intuitive tool to verify hypotheses and visualize structure–property correlations.

Assumption-free first-principles molecular-structure search for data generation in combination with dimensionality reduction and clustering for data analysis provide a powerful tool box to study structure formation trends. We could highlight the presence of large portions of configurational space that consist of *cis* isomers of the peptide bond. Albeit energetically unfavorable, these conformers may play an important role in the dynamics of polypeptides. By comparing isolated molecules and their complexes with Ca^{2+}, we can also reveal how a strong electrostatic perturbation modifies the energetic landscape of a small molecule—be it by shifting the stability of different conformers, or triggering the formation of new structures that are not observed in the absence of a cation. Furthermore, we also demonstrate the importance of automated analysis techniques in assessing the integrity and the internal consistence of a database, by successfully identifying a subset of structures associated with ill-converged energetics.

All of the techniques we discussed should be readily extendable to heterogeneous databases of molecules and solids, where we expect that the possibility of defining an alchemical kernel within the REMatch-SOAP metric will make it possible to tune the relative weight of composition and structure in determining the notion of similarity. By simplifying the analysis and the interpretation of computational datasets containing thousands or millions of hypothetical compounds, these methods will be crucial to unleash the full potential of computational materials design.

Additional file

Additional file 1. Supplementary file contains dissimilarity matrix data for all the oligopeptides available in the database including the one discussed in this paper.

Authors' contributions

MC and SD designed the calculations and developed the methods. FM and SD performed calculations and analysis, and prepared the materials for the manuscript. CB and TI provided insights on the implications of the analysis of the oligopeptide database. All the authors contributed to the writing of the manuscript. All authors read and approved the final manuscript.

Author details

[1] National Center for Computational Design and Discovery of Novel Materials (MARVEL), Lausanne, Switzerland. [2] Laboratory of Computational Science and Modelling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [3] Theory Department of the Fritz Haber Institute, Faradayweg 4-6, 14195 Berlin-Dahlem, Germany.

Acknowledgements

S.D. would like to thank Czuee Morey (University of Geneva) for insightful discussion. C.B. thanks Matti Ropo (Tampere University of Technology) Volker Blum (Duke University) and Matthias Scheffler (Fritz Haber Institute) for support and discussion.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The oligopeptide database used as an example in this paper is already available online [27]. All the data required to generate the figures in this paper and same analysis data for other oligopeptides in the database are provided in the

Additional file 1. Software to perform the structural analysis, mapping of datasets and hierarchical clustering is freely available from GIT repositories [85].

Funding
S.D. and M.C. would like to acknowledge support from the NCCR MARVEL. M.C., T.I. and C.B. would like to acknowledge funding from the MPG-EPFL center for molecular nanoscience.

Received: 29 September 2016 Accepted: 17 January 2017
Published online: 02 February 2017

References

1. Pizzi G, Cepellotti A, Sabatini R, Marzari N, Kozinsky B (2016) AiiDA: automated interactive infrastructure and database for computational science. Comput Mater Sci 111(1):218–230
2. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, Gold-Parker A et al (2011) The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett 2(17):2241–2251
3. Ortiz C, Eriksson O, Klintenberg M (2009) Data mining and accelerated electronic structure theory as a tool in the search for new functional materials. Comput Mater Sci 44(4):1042–1049
4. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C (2013) Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). JOM 65(11):1501–1509
5. Villars P, Berndt M, Brandenburg K, Cenzual K, Daams J, Hulliger F et al (2004) The pauling file, binaries edition. J Alloys Compd 367(1–2):293–297
6. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S et al (2013) Commentary: The materials project: a materials genome approach to accelerating materials innovation. APL Mater 1(1):011002
7. White A (2012) The materials genome initiative: one year on. MRS Bull 37(08):715–716
8. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA (2012) Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett 108(5):058301
9. Ghiringhelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M (2015) Big data of materials science: critical role of the descriptor. Phys Rev Lett 114(10):105503
10. Huan TD, Mannodi-Kanakkithodi A, Ramprasad R (2015) Accelerated materials property predictions and design using motif-based fingerprints. Phys Rev B Condens Matter Mater Phys 92(1):14106
11. Botu V, Ramprasad R (2015) Learning scheme to predict atomic forces and accelerate materials simulations. Phys Rev B Condens Matter Mater Phys 92(9):94306
12. Kusne A, Gao T, Mehta A, Ke L, Cuong Nguyen M, Ho KM et al (2014) On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. Sci Rep 4:6367
13. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA (2014) Quantum chemistry structures and properties of 134 kilo molecules. Sci Data 1:140022
14. Arsenault LF, Lopez-Bezanilla A, Von Lilienfeld OA, Millis AJ (2014) Machine learning for many-body physics: the case of the Anderson impurity model. Phys Rev B Condens Matter Mater Phys 90(15):155136
15. Ropo M, Schneider M, Baldauf C, Blum V (2016) First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. Sci Data 3:160009
16. Rodriguez A, Laio A (2014) Clustering by fast search and find of density peaks. Science 344(6191):1492–1496
17. Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678
18. Yu G, Chen J, Zhu L (2009) Data mining techniques for materials informatics: datasets preparing and applications. In: 2009 2nd international symposium on knowledge acquisition and modeling, KAM 2009, vol 2, pp 189–192
19. Isayev O, Fourches D, Muratov EN, Oses C, Rasch K, Tropsha A et al (2015) Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem Mater 27(3):735–743
20. Balachandran PV, Theiler J, Rondinelli JM, Lookman T (2015) Materials prediction via classification learning. Sci Rep 5:13285
21. Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG (2010) Systematic determination of order parameters for chain dynamics using diffusion maps. Proc Natl Acad Sci USA 107(31):13597–13602
22. Ceriotti M, Tribello GA, Parrinello M (2011) From the cover: Simplifying the representation of complex free-energy landscapes using sketch-map. Proc Natl Acad Sci 108(32):13023–13028
23. Ga Tribello, Ceriotti M, Parrinello M (2012) Using sketch-map coordinates to analyze and bias molecular dynamics simulations. Proc Natl Acad Sci 109(14):5196–5201
24. Ceriotti M, Tribello GA, Parrinello M (2013) Demonstrating the transferability and the descriptive power of sketch-map. J Chem Theory Comput 9(3):1521–1532
25. Rohrdanz MA, Zheng W, Clementi C (2013) Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. Annu Rev Phys Chem 64(1):295–316
26. De S, Bartók AP, Csányi G, Ceriotti M (2016) Comparing molecules and solids across structural and alchemical space. Phys Chem Chem Phys 18(20):13754
27. Ropo M, Baldauf C, Blum V (2016) Berlin ab initio amino acid DB. http://aminoaciddb.rz-berlin.mpg.de/. Accessed 31 Jan 2017
28. Pietrucci F, Andreoni W (2011) Graph theory meets ab initio molecular dynamics: atomic structures and transformations at the nanoscale. Phys Rev Lett 107(8):85504
29. Szlachta WJ, Bartók AP, Csányi G (2014) Accuracy and transferability of Gaussian approximation potential models for tungsten. Phys Rev B Condens Matter Mater Phys 90(10):104108
30. Lopez-Bezanilla A, Von Lilienfeld OA (2014) Modeling electronic quantum transport with machine learning. Phys Rev B Condens Matter Mater Phys 89(23):235411
31. Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R (2013) Accelerating materials property predictions using machine learning. Sci Rep 3:2810
32. Bartók AP, Gillan MJ, Manby FR, Csányi G (2013) Machine-learning approach for one- and two-body corrections to density functional theory: applications to molecular and condensed water. Phys Rev B Condens Matter Mater Phys 88(5):054104
33. Rupp M, Proschak E, Schneider G (2007) Kernel approach to molecular similarity based on iterative graph similarity. J Chem Inf Model 47(6):2280–2286
34. Hirn M, Poilvert N, Mallat S (2015) Quantum energy regression using scattering transforms. arXiv preprint arXiv:150202077
35. Montavon G, Rupp M, Gobre V, Vazquez-Mayagoitia A, Hansen K, Tkatchenko A et al (2013) Machine learning of molecular electronic properties in chemical compound space. New J Phys 15(9):95003
36. Snyder JC, Rupp M, Hansen K, Müller KR, Burke K (2012) Finding density functionals with machine learning. Phys Rev Lett 108(25):253002
37. Ghasemi SA, Hofstetter A, Saha S, Goedecker S (2015) Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. Phys Rev B 92(4):045131
38. Von Lilienfeld OA (2013) First principles view on chemical compound space: gaining rigorous atomistic control of molecular properties. Int J Quantum Chem 113(12):1676–1689
39. Hansen K, Biegler F, Ramakrishnan R, Pronobis W, Von Lilienfeld OA, Müller KR et al (2015) Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. J Phys Chem Lett 6(12):2326–2331
40. Zhu L, Amsler M, Fuhrer T, Schaefer B, Faraji S, Rostami S et al (2016) A fingerprint based metric for measuring similarities of crystalline structures. J Chem Phys 144(3):034203
41. Cuturi M (2013) Sinkhorn distances: lightspeed computation of optimal transport. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, vol 26. Curran Associates Inc, Red Hook, pp 2292–2300
42. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1):37–52
43. Kruskal JB (1964) Nonmetric multidimensional scaling: a numerical method. Psychometrika 29(2):115–129

44. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science (New York, NY) 290(5500):2319–2323
45. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F et al (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci USA 102(21):7426–7431
46. Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319
47. Jain AK, Murty MP, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
48. Aggarwal CC, Reddy CK (2013) Data clustering: algorithms and applications. CRC Press, Boca Raton
49. Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: an overview. Wiley Interdiscip Rev Data Min Knowl Discov 2(1):86–97
50. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov 2(3):283–304
51. Jing L, Ng MK, Huang JZ (2007) An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans Knowl Data Eng 19(8):1026–1041
52. Su MC, Chou CH (2001) A modified version of the K-means algorithm with a distance based on cluster symmetry. IEEE Trans Pattern Anal Mach Intell 23(6):674–680
53. Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd international conference on knowledge discovery and data mining. AAAI Press, San Jose, pp 226–231
54. Ankerst M, Breunig MM, Kriegel HP, Sander J (1999) Optics: ordering points to identify the clustering structure. In: ACM Sigmod record. ACM Press, New York, pp 49–60
55. Zhao X, Liang J, Cao F (2014) A simple and effective outlier detection algorithm for categorical data. Int J Mach Learn Cybern 5(3):469–477
56. Yamanishi K, Takeuchi JI, Williams G, Milne P (2004) On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Min Knowl Discov 8(3):275–300
57. Petrovskiy MI (2003) Outlier detection algorithms in data mining systems. Program Comput Softw 29(4):228–237
58. Angiulli F, Pizzuti C (2002) In: Elomaa T, Mannila H, Toivonen H (eds) Fast outlier detection in high dimensional spaces, vol 2431. Springer, Berlin, pp 15–27
59. Breunig MM, Kriegel HP, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. ACM SIGMOD Rec 29(2):93–104
60. Aggarwal CC, Yu PS, Aggarwal CC, Yu PS (2001) Outlier detection for high dimensional data. In: Proceedings of the 2001 ACM SIGMOD international conference on Management of data—SIGMOD '01, vol 30, no 2, pp 37–46
61. Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X et al (2009) Ab initio molecular simulations with numeric atom-centered orbitals. Comput Phys Commun 180(11):2175–2196
62. Perdew JPJ, Burke K, Ernzerhof M, of Physics D, Quantum Theory Group Tulane University NOLJ (1996) Generalized gradient approximation made simple. Phys Rev Lett 77(18):3865–3868
63. Tkatchenko A, Scheffler M (2009) Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. Phys Rev Lett 102(7):073005
64. Tkatchenko A, Rossi M, Blum V, Ireta J, Scheffler M (2011) Unraveling the stability of polypeptide helices: critical role of van der Waals interactions. Phys Rev Lett 106(11):118102
65. Baldauf C, Pagel K, Warnke S, Von Helden G, Koksch B, Blum V et al (2013) How cations change peptide structure. Chem Eur J 19(34):11224–11234
66. Schubert F, Rossi M, Baldauf C, Pagel K, Warnke S, von Helden G et al (2015) 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 17(11):7373–7385
67. Schubert F, Pagel K, Rossi M, Warnke S, Salwiczek M, Koksch B et al (2015) Native like helices in a specially designed β peptide in the gas phase. Phys Chem Chem Phys 17(7):5376–5385
68. Rossi M, Chutia S, Scheffler M, Blum V (2014) Validation challenge of density-functional theory for peptides-example of Ac-Phe-Ala5-LysH(+). J Phys Chem A 118(35):7349–7359
69. Baldauf C, Rossi M (2015) Going clean: structure and dynamics of peptides in the gas phase and paths to solvation. J Phys Condens Matter Inst Phys J 27(49):493002
70. Ropo M, Blum V, Baldauf C (2016) Trends for isolated amino acids and dipeptides: conformation, divalent ion binding, and remarkable similarity of binding to calcium and lead. arXiv:160602151
71. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7(1):95–99
72. Fischer G (2000) Chemical aspects of peptide bond isomerisation. Chem Soc Rev 29(2):119–127
73. Dugave C, Demange L (2003) *Cis–trans* isomerization of organic molecules and biomolecules: implications and applications. Chem Rev 103(7):2475–2532
74. Weiss MS, Jabs A, Hilgenfeld R (1998) Peptide bonds revisited. Nat Struct Biol 5(8):676
75. De S, Ghasemi SA, Willand A, Genovese L, Kanhere D, Goedecker S (2011) The effect of ionization on the global minima of small and medium sized silicon and magnesium clusters. J Chem Phys 134(12):124302
76. Heidari I, De S, Ghazi SM, Goedecker S, Kanhere DG (2011) Growth and structural properties of MgN (N = 10–56) clusters: density functional theory study. J Phys Chem A 115(44):12307–12314
77. Ghazi SM, De S, Kanhere DG, Goedecker S (2011) Density functional investigations on structural and electronic properties of anionic and neutral sodium clusters Na N (N = 40–147): comparison with the experimental photoelectron spectra. J Phys Condens Matter 23(40):405303
78. Pochet P, Genovese L, De S, Goedecker S, Caliste D, Ghasemi SA et al (2011) Low-energy boron fullerenes: role of disorder and potential synthesis pathways. Phys Rev B Condens Matter Mater Phys 83(8):81403
79. Ardevol A, Tribello GA, Ceriotti M, Parrinello M (2015) Probing the unfolded configurations of a β-hairpin using sketch-map. J Chem Theory Comput 11(3):1086–1093
80. Baškarada S, Koronios A (2014) A critical success factor framework for information quality management. Inf Syst Manag 31(4):276–295
81. Van Den Broeck J, Cunningham SA, Eeckels R, Herbst K (2005) Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2(10):0966–0970
82. Gevorgyan A, Poolman MG, Fell DA (2008) Detection of stoichiometric inconsistencies in biomolecular models. Bioinformatics 24(19):2245–2251
83. Ferretti L, Colajanni M, Marchetti M (2014) Distributed, concurrent, and independent access to encrypted cloud databases. IEEE Trans Parallel Distrib Syst 25(2):437–446
84. De S, Willand A, Amsler M, Pochet P, Genovese L, Oedecker S (2011) Energy landscape of fullerene materials: a comparison of boron to boron nitride and carbon. Phys Rev Lett 106(22):225502
85. Code repositories from the Laboratory of Computational Science and Modelling at EPFL (2014). http://epfl-cosmo.github.io/
86. Libatoms (2014) http://www.libatoms.org/