Identifying conformational isomers of organic molecules in solution via unsupervised clustering

Veselina Marinova,[†,‡] Laurence Dodd,[†] Song-Jun Lee,[†] Geoffrey P. F. Wood,[¶] Ivan Marziano,[§] and Matteo Salvalaglio[*,†]

[†]*Thomas Young Centre and Department of Chemical Engineering, University College London, London WC1E 7JE, UK.*

[‡]*Department of Materials Science and Engineering, The University of Sheffield, Sheffield S1 3JD, UK*

[¶]*Pfizer Worldwide Research and Development, Groton Laboratories, Groton, Connecticut 06340, USA*

[§]*Pfizer Worldwide Research and Development, Sandwich, Kent CT13 9NJ, UK*

E-mail: m.salvalaglio@ucl.ac.uk

Abstract

We present a systematic approach for the identification of statistically relevant conformational macrostates of organic molecules from molecular dynamics trajectories. The approach applies to molecules characterised by an arbitrary number of torsional degrees of freedom and enables the transferability of the macrostates definition across different environments. We formulate a dissimilarity measure between molecular configurations that incorporates information on the characteristic energetic cost associated with transitions along all relevant torsional degrees of freedom. Such metric is employed to perform unsupervised clustering of molecular configurations based on the *fast search and find of density peaks* algorithm. We apply this method to investigate the equilibrium conformational ensemble of Sildenafil, a conformationally complex pharmaceutical compound, in different environments including the crystal bulk, the gas phase and three different solvents (acetonitrile, 1-butanol, and toluene). We demonstrate that, while Sildenafil can adopt more than one hundred metastable conformational configurations, only 12 are significantly populated across all the environments investigated. Despite the complexity of the conformational space, we find that the most abundant conformers in solution are the closest to the conformers found in the most common Sildenafil crystal phase.

Introduction

Conformational isomerism in organic molecules is an important characteristic which bears significance in a variety of problems. For example, binding properties of proteins in protein-ligand complexes are controlled by their conformational configuration by affecting association/dissociation rates, and by entropic contributions to the process.[1,2] Understanding the details and mechanisms of conformational changes that proteins undergo is an important part of modern drug discovery methodologies.[3] For small organic molecules the ability to adopt different conformational configurations can open the possibility for the formation of multiple crystal forms known as conformational polymorphs[4,5] - crystal structures of components with the same chemical formula but different molecular shape. This phenomenon is particularly important in the pharmaceutical industry where the uncontrolled occurrence of an undesired polymorphic form can affect the stability, shelf-life or efficacy of the drug. In the field of crystallisation, conformational rearrangements are not only relevant to polymorphism. In our previous work[6] on the study of ibuprofen conformational isomerism at the crystal/solution interface, we demonstrate how, even for relatively small systems, conformational rearrangements, crystal growth and dissolution are inherently coupled. Additionally state-to-state transitions of a molecule along its path of incorporation into the crystal from solution may be limited by conformational rearrangements.

Computational studies of conformational rearrangement in small organic molecules often use internal torsional angles to describe the adopted molecular configuration.[6–9] Torsional angles are a convenient way of describing rearrangements as they provide a fine-grained comprehensive picture of the internal molecular configuration space. To describe the conformation of larger molecules such as peptides or aliphatic chains, however, resorting to descriptors such as end-to-end distance or Root Mean Square Deviation (RMSD)[1,10] is a common choice, made necessary by the fact that the torsional angle space for these systems is high-dimensional and impractical to read and interpret. A critical drawback of this approach is that, by reducing the dimensionality of the descriptors' space used to represent configurations, degeneracy is introduced, and consequently information lost. More generally, reliable conformational descriptors are particularly important when implementing enhanced sampling techniques. Enhanced sampling techniques are heavily reliant on the use of appropriate system descriptors. Particularly for studying self-assembly processes, conformationally flexible systems currently present a major challenge,[11] thereby driving the search for a systematic approach to their classification.

Dividing the conformational space of large organic molecules is often done via partitional clustering methods, meaning that the data is assigned into groups without any hierarchical structure, based on a chosen criterion.[12] One of the best known partitioning algorithms is k-means.[13] The idea behind this algorithm is to define a k-centroid for each cluster and measure the distance between a data point and each of the cluster centres. Many computational works have achieved partitioning of molecular configurational space through k-means clustering based methodologies.[14–18] Despite its ease of use, it has a few important limitations. Cluster centres can be difficult to define arpiori, as well as, k-means can be very sensitive to outliers and noise.[19,20] A partition-based algorithm which has tackled the drawbacks of k-means clustering is Affinity Propagation (AP),[21] where all data points are regarded as potential cluster centres. The negative distance between data points is their affinity, so the bigger the sum of the affinity of one data point to other data points, the higher the probability of this data point being a cluster centre is. AP has been implemented in the study of protein conformations,[22] however it has also been regarded and complex and costly approach.[19,23]

As an alternative to distance-based algorithms, density-based algorithms have been developed.[20,24–28] They work on the principle of assigning densities to local points and are able to separate clusters based on high- and low-density regions. Such algorithms not only do

not require defining the number of cluster centres *apiori*, but also allow the identification of non-spherical clusters. Particularly suited to the analysis of molecular dynamics (MD) simulations is the Fast Search and Find of Density Peaks (FSFDP) clustering algorithm, developed by Rodriguez and Laio.[26] By computing the distance between all pairs of data points, the algorithm identifies the points with highest density in their neighbourhood as the cluster centroids.

Such tools, which allow the systematic classification of molecular configurations regardless of the dimensionality of the space of descriptors necessary to completely capture every conformational change, have the potential to improve existing methodologies for studying the effects of conformational rearrangements during crystal nucleation and growth.[6]

Here, we propose a methodology which enables the study of conformational isomerism in a general way for systems with a large number of torsional degrees of freedom. Our approach, based on the application of the Fast Search and Find of Density Peaks clustering algorithm,[26] allows to define a set of conformational states that is common to multiple environments (i.e. solvents), and enables a systematic assessment of their impact on the conformational landscape. We demonstrate this approach by studying the conformational rearrangement of sildenafil, a commercially available active pharmaceutical ingredient (API).[7]

Sildenafil is the main component of Viagra,[7] which is known to have two polymorphic forms.[29,30] Sildenafil is a relatively large molecule consisting of 63 atoms and a number of ring structures. The two forms of sildenafil (denoted form 1 and form 2) are morphotropically related to one another as a noncrystallographic rearrangement can transform one to the other,[30] with form I being the thermodynamically stable form. Both forms have two molecules in the asymmetric unit adopting different conformations.[29,30]

With the use of a data clustering approach we demonstrate how characteristic conformational configurations can be identified *a priori* for a molecule in the gas phase in order to then extract quantitative information on conformational states from enhanced sampling molecular dynamics simulations performed in solution under experimentally relevant conditions. Our methodology also enables the breakdown of the free energy of a conformational state into enthalpic and entropic contributions, providing a valuable insight into the effect of the solvent on conformational isomerism. With this work we aim to propose a method for conformational analysis which provides a route towards achieving rapid and automated conformational classification, enabling the comprehensive study of conformational isomerism in solution for systems for which it is currently impractical.

Methods

In this work, molecular dynamics (MD) simulations are used to study the conformational isomerism of sildenafil in the crystal bulk, in the gas phase and in three solvents - acetonitrile, 1-butanol and toluene. MD is combined with well-tempered metadynamics (WTmetaD) to enable the study of the conformational rearrangements of sildenafil in solution. The Fast Search and Find of Density Peaks (FSFDP) clustering method, developed by Rodriguez and Laio,[26] is used to identify sildenafil conformers in the gas phase and generate a characteristic fingerprint in torsional angle space for each of them. The metric used to define the similarity between configurations includes information on the free energy cost associated to transitions in every degree of freedom explicitly considered. The fingerprints are used to post-process the biased trajectories in solution and assign a conformational macrostate for each trajectory step. Through the implementation of a reweighting procedure, the equilibrium probability of conformers, as well as enthalpic and entropic contributions to the free energy of each conformer for each of the solvents considered is obtained.

In this study, the molecular rearrangement of the drug is described in torsional angle space by considering all internal dihedral angles as shown in Figure 1. By including all torsional degrees of freedom of the molecule, this study

adopts a systematic and transferable strategy for tackling conformational rearrangement.

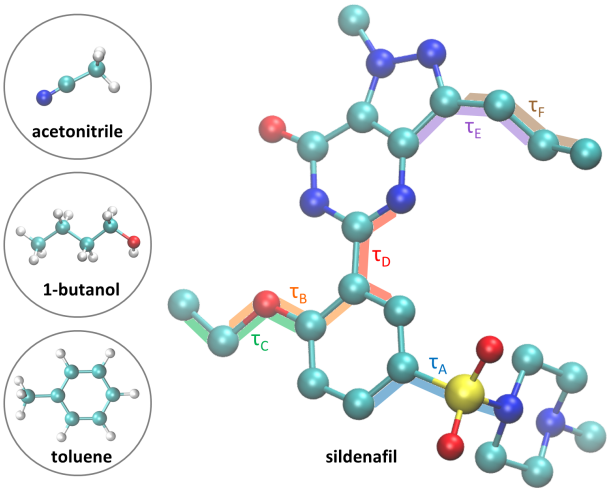


Figure 1: Sildenafil structure, where hydrogen atoms have been excluded for simplicity. The six internal torsional angles, labelled τ_A, τ_B, τ_C, τ_D, τ_E and τ_F, are marked on the structure. All images of molecular structures shown have been generated with VMD.[31]

Molecular Dynamics Setup

Molecular dynamics simulations of form I and form II crystal polymorphs of sildenafil, a sildenafil molecule in the gas phase and a sildenafil molecule in three different solvents were performed using the Generalised Amber Force Field (GAFF).[32] For all systems considered in this work GAFF is able to reproduce properties consistent with experimental data (Section A of the Supplementary Material). MD simulations were performed with Gromacs 5.1.4[33] with an explicit representation of the solvent. Force field parameters for solvent molecules were obtained from the Virtual Chemistry solvent database.[34,35] A standard cut-off distance of 1.0 nm for the non-bonded interactions was chosen, along with including long-range intermolecular interactions using the particle-mesh Ewald (PME) approach.[36] A time step of 2 fs was used. Temperature and pressure control have been implemented through the use of the Bussi-Donadio-Parrinello thermostat,[37] Berendsen barostat[38] and Parrinello-Rahman barostat.[39] More detail on the applied pressure control and the recovered system density in each case can be found in Section A of the Supplementary Material.

Simulations of the Crystal Bulk Supercells of size 3.5 × 3.5 × 5.0 nm and 7.0 × 3.5 × 2.3 nm, representing crystal forms I and II respectively, were set up containing 96 molecules each. Crystal structure *.cif* files[29] at ambient temperature and pressure were obtained from the CSD under deposition codes QEGTUT and QEGTUT02. In both cases unbiased MD simulations were performed for 10 ns in the isothermal-isobaric ensemble, implemented by applying an anisotropic pressure control.

Simulation of Sildenafil in the Gas Phase A 450 ns unbiased simulation of a sildenafil molecule in a box of 2.0 × 2.0 × 2.0 nm in vacuum was carried out in the canonical ensemble. A free energy profile of each torsional angle as denoted in Figure 1 was obtained. All one-dimensional free energy profiles are reported in Section B of the Supporting Information.

Simulations of Sildenafil in Solution Simulations in solution were set up by solvating a single sildenafil molecule with each of the three solvents used in this study - acetonitrile, 1-butanol and toluene - by using the `insert-molecules` utility in Gromacs in a box of approximate size of 4 × 4 × 4 nm. MD simulations were performed in combination with well-tempered metadynamics. All simulations were performed in the isothermal-isobaric ensemble (NPT) at pressure of 1 bar and temperature of 300 K.

Well-Tempered Metadynamics Setup

Metadynamics was implemented in order to enhance fluctuations in the internal rearrangement of sildenafil in solution for computational efficiency. The bias was applied as a function of the τ_A torsional angle as shown in Figure 1 in blue. The choice of the CV was made based on

10 ns exploratory MD simulations, which revealed that overcoming the barrier associated with the rotation of τ_A in an efficient way requires enhanced sampling in all three solvents. The biasing protocol was applied in the form of Gaussian functions with a width of 0.3 rad and height of 2.5 k_BT at a rate of every 500 simulation steps with a bias factor of 15 K. The use of WTmetaD was implemented through plumed 2.4.[40] All the data and input files, required to reproduce the results reported in this paper, are available on plumed-NEST,[41] the public repository of the PLUMED consortium as plumID:20.032.

Clustering Procedure

Studying the conformational isomerism of systems with several internal degrees of freedom in a systematic and transferable way requires the need of grouping molecular configurations and identifying relevant conformational states. Achieving this provides the opportunity of reducing the dimensionality of the problem and enables the analysis of biased molecular dynamics trajectories in order to obtain useful kinetic and thermodynamic information.

To achieve this partitioning of configurational space in an unsupervised data-driven manner, a recently developed data clustering method is applied. In this work, the **Fast Search and Find of Density Peaks (FSFDP)** algorithm, developed by Rodriguez and Laio,[26] is implemented. The algorithm can be used to group molecular configurations into clusters of structures based on their similarity by calculating a distance matrix between configurations. The distance matrix refers to a matrix containing the distance between any two molecular configurations i and j as a function of the chosen system descriptors e.g. internal torsional angles, with no limit on the dimensionality, as shown in Eq. 1.

$$d_{ij} = \sqrt{\sum_{n=1}^{N_{CV}} d_n^2} \tag{1}$$

Once the distance matrix has been calculated, the algorithm operates by calculating the density of each point i, evaluated by considering the number of neighbours within a distance cut-off d_c according to:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{2}$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise.

$$\delta_i = \min_{j:p_j > p_i} d_{ij} \tag{3}$$

Cluster centres are characterised by having the highest density within a cluster of points and a large distance from points with a higher density as shown in Eq. 3. Each cluster has an associated core set of data points and a halo, evaluated based on the given cut-off. The core set are the points which belong exclusively to a cluster and can be found within the selected distance cut-off from the cluster centre, while the halo is considered as the noise around the cluster core, while still affiliated with the given cluster. This grouping algorithm is particularly powerful as the clustering procedure is such that the number of clusters arises intuitively, which makes it particularly suited to identifying characteristic conformational configurations, described by a highly multidimensional set of CVs.

Here, we consider each frame of an MD trajectory of a sildenafil molecule in the gas phase as a single point in the 6-dimensional torsional angle space and so the distance d_{ij} between every two frames i and j, is calculated according to Eq.4:

$$d_{ij} = \sqrt{\sum_{n=1}^{N_{CV}} (d_n w_n)^2} \tag{4}$$

where N_{CV} refers to the total number of system descriptors, which in this case is 6 torsional angles, while w stands for a *weight* applied to each dimension as described below.

In the above equation d_n is refers to the absolute difference in the values of any of the given torsional angles for points i and j. For example, d_1 is calculated as:

$$d_1 = |\tau_{Ai} - \tau_{Aj}| \tag{5}$$

Should the value of d_n be greater than π, periodicity is accounted for by subtracting the value from $2pi$ as follows:

$$d_n > \pi : d_n = 2\pi - d_n \tag{6}$$

Clustering structures in a meaningful way requires a distance matrix calculation which is able to distinguish between conformational *transitions* and conformational *adjustments*. These terms refer to the nature of the conformational rearrangement within the molecule. A conformational *transition* describes the conversion of one stable conformational state into another, usually associated with overcoming a free energy barrier higher than k_BT. An *adjustment* is, on the other hand, a term used to describe a minor rearrangement which is not associated with a new conformer, but rather a relaxation of the structure from one configuration to another, both of which occupy the same free energy minimum in collective variable space.[5] In order to resolve these two cases when calculating the distance matrix, a slight modification to the algorithm is applied by including a weight associated with the free energy barrier of rotation, w, of each torsional angle in order to scale the distance d. In such a way, rotations which lead to new conformational configurations through rare, activated transition, will have a higher impact on the distance compared to those associated with an adjustment or a fast conversion.

Table 1: Free energy barrier of rotation to each torsional angle obtained from simulations in vacuum, along with the corresponding rescaling used as a weight in calculating the distance matrix.

Angle	**Barrier Height** $[k_BT]$	**Weight**
τ_A	6	5
τ_B	1.5	1.3
τ_C	2	1.7
τ_D	1.2	1
τ_E	2.5	2.1
τ_F	4	3.3

The weight factor, w, is obtained by recording the lowest rotational barrier for each torsional angle according to the calculated one-dimensional free energy profiles (see Section B of the SI) along each dimension of the CV space. The barriers are then normalised with respect to the lowest one as shown in Table 1. An example for the case of τ_A is shown in Figure 2.

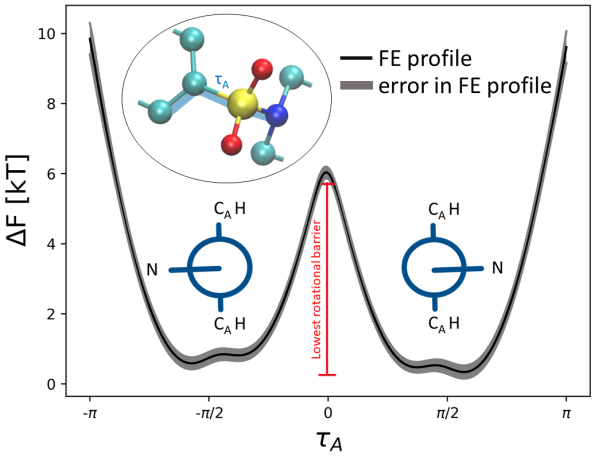


Figure 2: Free energy profile of torsional angle τ_A obtained from a vacuum simulation, along with Newman projections of the angle in the free energy minima. The lowest free energy barrier of rotation was extracted for the clustering weights.

We note that in the case of larger, more complex solutes with a higher number of conformational degrees of freedom one may need to analyse large datasets, and in such cases the calculation of the distance matrix may become an efficiency bottleneck for the process. Recent work[27] addresses this problem pushing the limit of applicability of FSFDP to the million-configurations range.

Conformational Classification

Clustering of the molecular configurations was performed using the FSFDP algorithm, with a distance cut-off d_c tuned to obtain an average number of neighbours for every datapoint equal to 2% of the entire dataset. Based on the distance matrix calculation and the chosen cut-off, the sildenafil configurations sampled in the gas phase are grouped into clusters, each consisting of a cluster centre structure, a core set of configurations and a halo. All configurations assigned

to a cluster were used to generate a structural *fingerprint* for the identified conformer. The term fingerprint refers to the probability distribution of each torsional angle as illustrated in Figure 1. All fingerprints are provided in Section D of the Supporting Information.

Conformational classification of each frame of the trajectories of sildenafil in solution was carried out in order to represent conformational change of the molecule in a one-dimensional space and hence enable further analysis of the enthalpic and entropic contributions to conformational isomerism in solution. To achieve that, an algorithm which compares the instantaneous value of the torsional angles, defined to describe the molecular configuration, to each of the given fingerprints for every trajectory frame was set up. For every instantaneous torsional angle value where the corresponding probability density in the fingerprint is nonzero a value of 1 is assigned. The total number of variables used for the classification is 6 and therefore a score of 6 means that the molecular configuration in the given frame matches a fingerprint and therefore is assigned the corresponding conformer number. A score lower than 6 indicates at least one mismatch between the given configuration and the fingerprint and is therefore assigned a value of 0 signifying that it remains unclassified.

Conformational Equilibrium Probability Distribution

A characteristic fingerprint in torsional angle space for each dominating sildenafil conformer in the gas phase was generated. Choosing the gas phase as a reference is inspired by the work of Cruz-Cabeza and Bernstein,[5] who use the same conditions to define reference conformational states. To calculate the conformational population of sildenafil in solution, each frame of the biased trajectory was assigned a characteristic conformer following the procedure discussed in the previous section. A discrete probability distribution in one-dimensional space can then be straightforwardly calculated, with the caveat that the bias potential deposited throughout the duration of the simulation needs to be accounted for in a procedure referred to as *reweighting*.

In this work, the total metadynamics bias potential applied as a function of τ_A and recovered at the end of the simulation of a sildenafil molecule in solution, $V^{total}(\tau_A)$, is used in the reweighing scheme.[42] The trajectory is post-processed so that each time frame, with a corresponding value for each torsional angle, as well a conformational cluster number, will also have a value associated with the total bias deposited in that particular point in the CV space of τ_A. For simplicity, let us refer to this value as $V_i^{total}(\tau_A)$ where i stands for the trajectory frame number. The weight W_i applied to each frame when reconstructing the *unbiased* probability distribution of conformational isomers in solution is a Boltzmann weight associated with a rescaled value of $V_i^{total}(\tau_A)$ according to:

$$W_i = e^{\beta(V_i^{total} - max(V_i^{total}))} \tag{7}$$

In such a way, the weight associated with points in CV space where the maximum total external bias was deposited will have a value of 1, as it corresponds to the lowest point in the free energy profile in τ_A and all other frames will have a correspondingly lower weight. This reweighting scheme was implemented to reconstruct the population of sildenafil conformers in different solvents, as well as obtaining a two-dimensional probability density function of the conformer number and its associated potential energy, used in the free energy decomposition discussed in the next chapter.

Enthalpy and Entropy Contributions to Free Energy Differences between clusters

Free energy differences between conformational macrostates i and j of sildenafil in solution were computed from their equilibrium probability P_i and P_j as $\Delta G_{i,j} = -k_B T \ln \frac{P_j}{P_i}$ and decomposed into their enthalpic and entropic contributions following the procedure outlined by Gimondi et. al.[43,44] For instance, the difference in free energy between clusters i and j, $\Delta G_{i,j}$ can be expressed as $\Delta H_{i,j} - T\Delta S_{i,j}$, where

$\Delta H_{i,j}$ and $\Delta S_{i,j}$ are the enthalpy and entropy differences between states i and j. Hence the entropic contribution can obtained by difference as $T\Delta S_{i,j} = \Delta G_{i,j} - \Delta H_{i,j}$, once the term $\Delta H_{i,j}$ is known. Since conformational transitions of sildenafil are not associated to a change in the ensemble average of the system's volume, $\Delta H_{i,j} = \Delta U_{i,j} + P\Delta V_{i,j} \simeq \Delta U_{i,j}$, where $\Delta U_{i,j}$ is the difference in internal energy between conformational macrostates i and j. Moreover, since states i and j are conformational isomers sampled at the same temperature, the internal energy difference reduces to the difference in the ensemble average of the potential energy between states i and j as:

$$\Delta U_{i,j} = \langle E_P \rangle_j - \langle E_P \rangle_i \tag{8}$$

where $\langle E_P \rangle_i$, the ensemble average of the potential energy of configurations classified in cluster i, is computed as:

$$\langle E_P \rangle_i = \int E_P(\mathbf{r}) p(E_P(\mathbf{r}) | \mathbf{r} \in i) dE'_P \tag{9}$$

where $p(E_P(\mathbf{r})|\mathbf{r} \in i)$ is the potential energy probability density, conditional to the system being classified in cluster i. The probability density $p(E_P(\mathbf{r})|\mathbf{r} \in i)$ is computed as discussed in Gimondi et al.,[43] using the reweighting strategy described in the previous paragraph. In models in which explicit solvents are employed, the $\langle E_P \rangle_i$ is typically dominated by the contribution of the solvent molecules, and it is associated with large fluctuations that affect the convergence and accuracy of the $\Delta U_{i,j}$ estimate. In order to improve the statistical accuracy in the $\Delta U_{i,j}$, we follow the procedure outlined by Kollias et al.[44] and decompose the ensemble average of the potential energy in three components, namely $\langle E_P \rangle_i^{solute}$, $\langle E_P \rangle_i^{solvent}$ and $\langle E_P \rangle_i^{solute-solvent}$. The $\langle E_P \rangle_i^{solute}$, and $\langle E_P \rangle_i^{solvent}$ contributions account respectively for potential energy terms associated with interactions between atoms that belong exclusively to the solute and to the solvent species. The solute-solvent term $\langle E_P \rangle_i^{solute-solvent}$ accounts instead for non-bonded interactions between solute and solvent atoms. Since conformational changes in the solute do not affect the solvent-solvent contribution to the potential energy the term $\langle E_P \rangle_i^{solvent} = \langle E_P \rangle_j^{solvent} = const$, and thus $\Delta U_{i,j}$ reduces to:

$$\begin{aligned} \Delta U_{i,j} = \left(\langle {E_P}^{solute} \rangle_j - \langle {E_P}^{solute} \rangle_i \right) + \\ + \left(\langle {E_P}^{solute-solvent} \rangle_j - \langle {E_P}^{solute-solvent} \rangle_i \right) \end{aligned} \tag{10}$$

which is not affected by the large fluctuations of $\langle E_P \rangle_i^{solvent}$ that would mask the contribution of conformational transitions to $\Delta U_{i,j}$, and hamper the convergence of the enthalpy and entropy contribution to free energy differences. Despite implementing this strategy the convergence of the enthalpic and entropic contributions require a substantial sampling of the configuration space of the explicitly solvated system. Here we achieve sufficient sampling by running WTmetaD simulations for 0.5 to 0.6 μs. In the case of larger, more complex solutes with a higher number of conformational degrees of freedom we anticipate the need for replica exchange methods to exhaustively sample the configuration space and successfully apply this decomposition approach.

Results

The following sections summarise the analysis carried out on sildenafil conformers in the gas phase, as well as on the conformational rearrangements of sildenafil occurring in the crystal bulk and in solution. The results are organised as follows. First, the conformational freedom of sildenafil in the crystal bulk is reviewed by considering the torsional angle distribution obtained from MD simulations of forms I and II. Next, the results obtained from the clustering algorithm are reported, along with drawing a comparison between structures in the gas phase and those in the solid. The last section reports on the equilibrium distribution of sildenafil conformers in different solvents, obtained with the aid of WTmetaD, along with a breakdown of their corresponding free energy into enthalpy and entropy contributions.

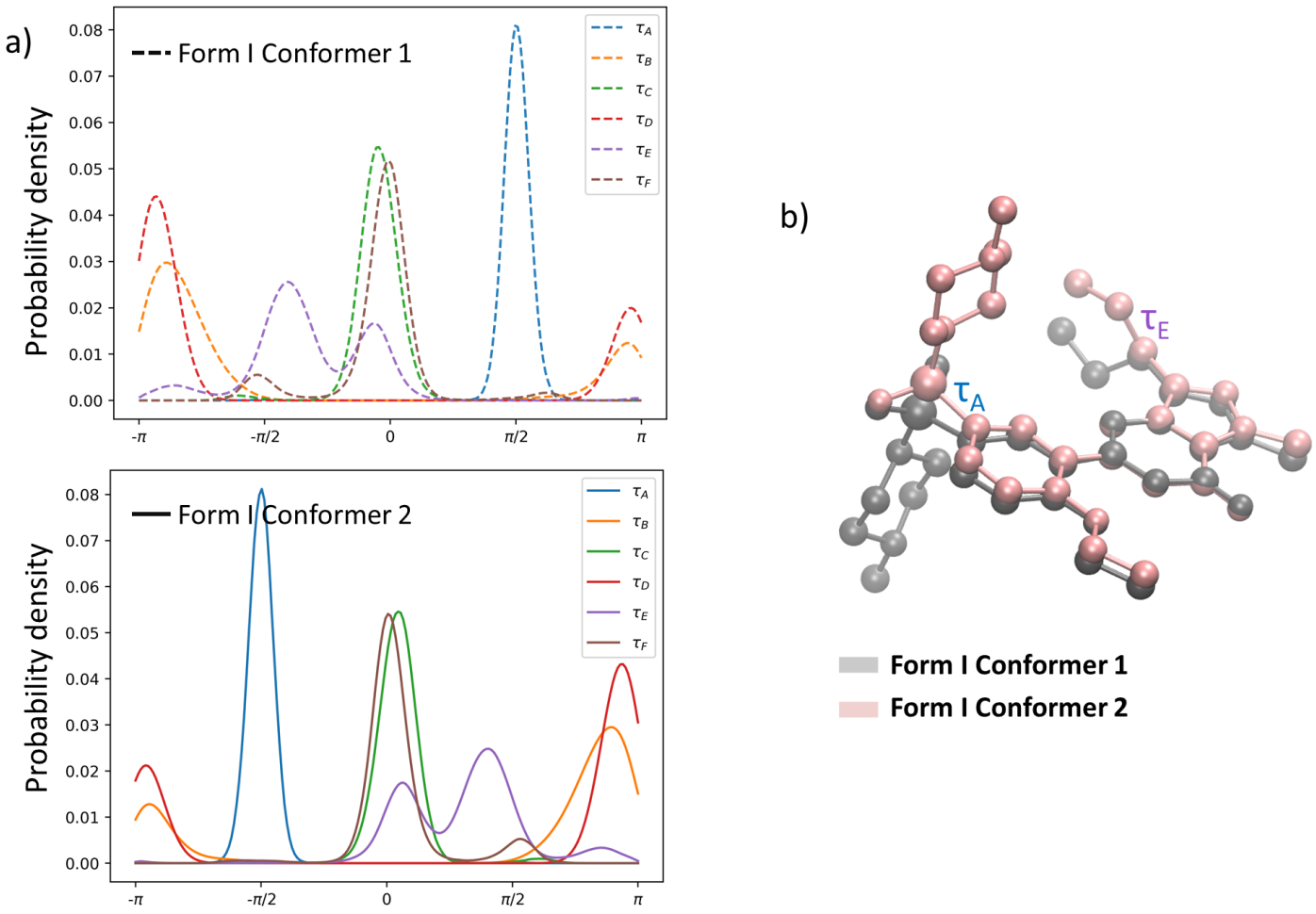


Figure 3: a) Probability density plots of each torsional angle for crystal conformers 1 and 2 obtained from a 10 ns MD simulation of form I sildenafil. b) Image of crystal form I conformers 1 and 2, obtained from the *.cif*, generated with VMD, where hydrogen atoms are removed for simplicity.

Conformational Rearrangements in the Crystal Bulk

The conformational rearrangement of sildenafil was first investigated for the case of a molecule in the crystal bulk in each of the two polymorphic forms. As mentioned, each polymorph contains two conformational isomers of the molecule. A probability distribution of each torsional angle representative of the conformational state adopted by sildenafil in the solid was obtained, which enables to gain insight into the degree of conformational freedom available in each crystal form.

The results reveal that while several of the dihedral angles of sildenafil are completely restrained by the crystal packing, a surprising amount of flexibility is accessible to the rest. A probability distribution of each torsional angle for crystal conformers 1 and 2 in form I is shown in Figure 3a. In the figure, a narrow and mono-modal distribution corresponds to each of the torsional angles τ_A, τ_C and τ_F, shown respectively in blue, green and brown. A mild degree of conformational adjustment is associated with torsional angles τ_B (in orange) and τ_D (in red) which are both distributed around $\pm\,\pi$. These torsional angles are associated with adjacent substituents on the phenyl ring, suggesting that the rearrangement is possibly related to releaving steric hindrance. The highest degree of rotational freedom is observed in the case of torsional angle τ_E, shown in purple, represent-

ing the rotation of the propanyl substituent of the pyrazole ring (see Figure 1), displaying a multi-modal distribution. The flexibility of τ_E reveals a moderate degree of conformational rearrangement available to the conformer, despite the restrictive environment traditionally associated with a crystal. This observation has been addressed by Barbas et al.[29] who have recorded the presence of a dynamic disorder in the propyl groups at room temperature in form I, which disappears at temperatures lower than 100K. The authors justify this observation with the fact that these functional groups do not establish strong intermolecular interactions with the surrounding atoms within the crystal, resulting in a moderate degree of flexibility in the chain.

Analysing the torsional angle distributions of each of the two conformers found within the crystal structure of form I reveals that they are conformational isomers, where a rotation along τ_A and τ_E can convert conformer 1 into conformer 2, as shown in Figure 3b. According to the results obtained from the MD simulation of form I, the internal rearrangement of crystal conformer 2 displays an identical behaviour as to that of conformer 1.

Similarly, the conformational freedom of sildenafil in form II was analysed through an unbiased MD simulation. A comparison of the probability distribution of each torsional angle of sildenafil between form I and form II reveals a similar conformational configuration in the two structures, with the only difference being a marginally lower flexibility of τ_E in form II compared to that observed in form I. A detailed comparison between the conformers in each form is provided in Section C of the Supporting Information. An experimental comparison between the two crystal forms is provided by Barbas et al.[30] who make a similar observation to the one reported here, and stress that any differences between conformers in the two crystals can be classified as conformational *readjustments* of the same gas phase conformer, validating the conclusions made on the basis of our MD simulations. As mentioned, the probability distribution of torsional angle τ_E, associated with the rearrangement of the pronanyl group, shows that the degree of rearrangement is counterintuitively *lower* in form II compared to form I, despite the presence of larger structural voids in the former. The experimental publication does not report measurements of the degree of disorder in form II, however the authors speculate that the conformational rearrangement of the propanyl groups will be dominated by a drive to keep the cavities in the structure empty, which could explain the conformational behaviour observed in the MD simulation of form II.

Structure Clustering in the Gas Phase

This section reports on the results obtained from implementing a data clustering algorithm in order to group structures from an MD trajectory of sildenafil in the gas phase and identify characteristic conformational configurations. A trajectory of a molecule in vacuum was chosen for this purpose as in the absence of solvent effects the internal rearrangement of the molecule is unhindered and thorough sampling of all possible molecular configurations can be achieved efficiently. This allows to consider the full conformational space of sildenafil in the clustering procedure. Such approach has the potential to provide a more meaningful and robust method of identifying representative conformational isomers over methods which rely on generating conformers thorough random search and local minima strategies. By considering the free energy profile of each torsional angle in the gas phase, as discussed in Section B of the SI, all possible combinations of structural local minima of the molecule in torsional angle space is estimated to be 144. However, in reality, each local minimum comprises of an ensemble of configurations, meaning that 144 structures is a rather conservative estimate, and in practice, there is a swarm of possible stable molecular configurations. For that reason, failing to explore the collective variable space thoroughly encounters the risk of missing out important structures due to the sheer number of available configurations, even for systems of moderate flexibility.

A typical output of the clustering algorithm

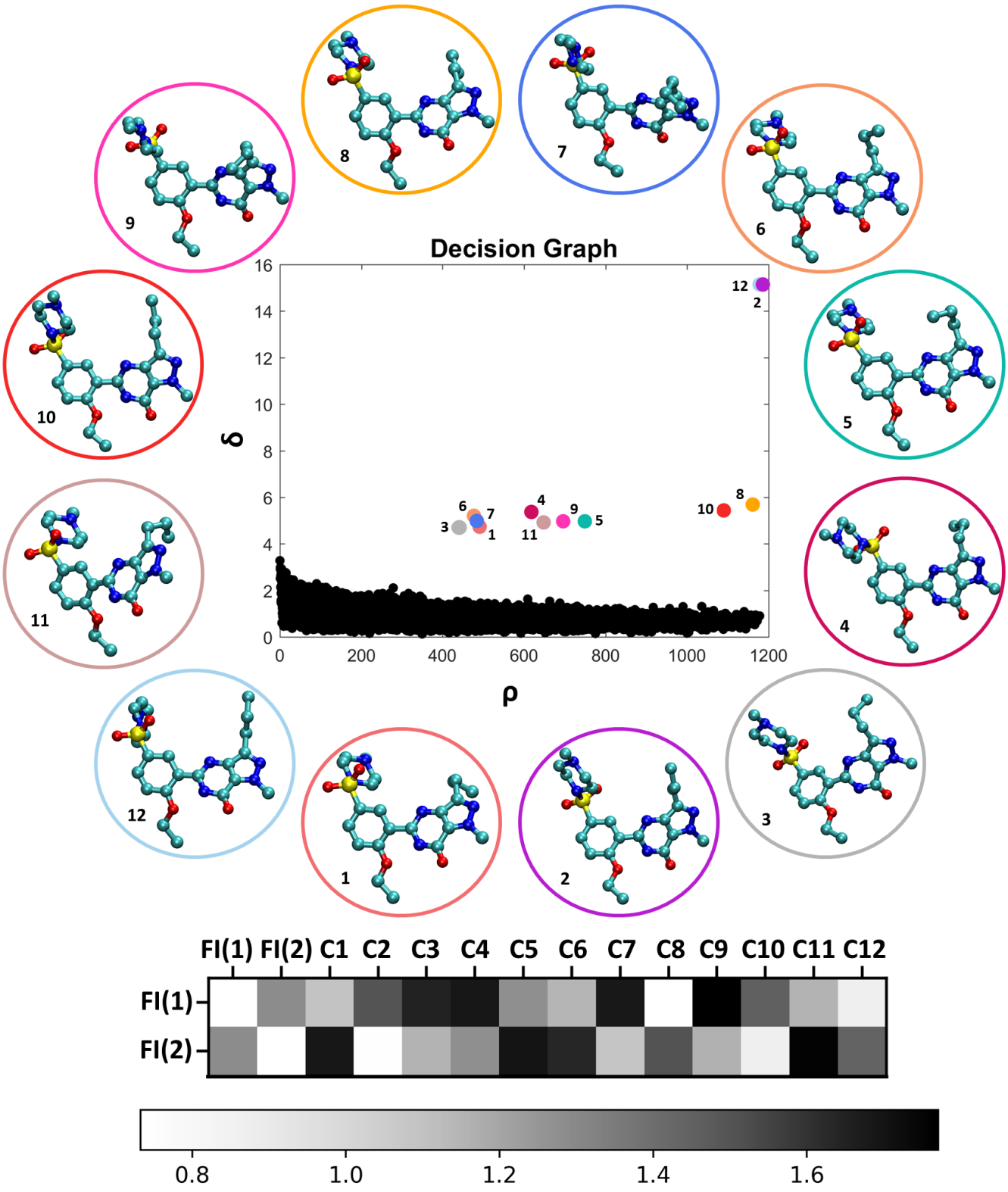
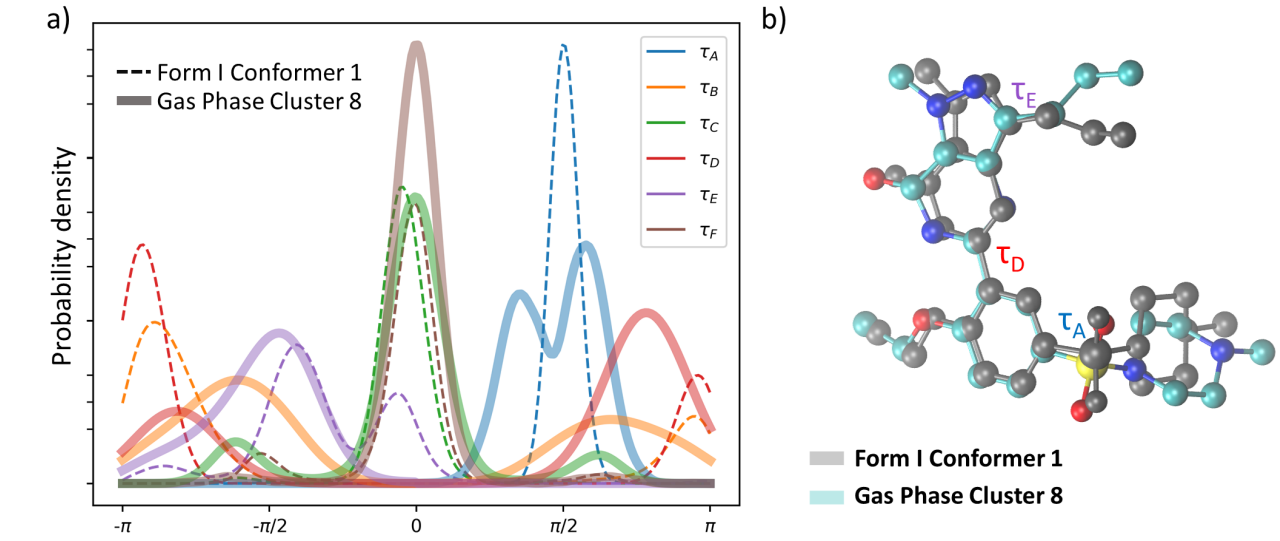


Figure 4: Decision graph of the structure clustering performed on the trajectory in the gas phase. The circled structures represent the cluster centres. The figure below represents the distance matrix between distributions of conformers 1 and 2 in the crystal and the identified from the algorithm cluster centres.

is a decision graph, displaying all data points as a function of their density ρ (number of neighbours) and distance from the nearest point of higher density δ (See Eq. 2 and 3). When ap-



	τ_A	τ_B	τ_C	τ_D	τ_E	τ_F
Distance [rad]	0.25	0.88	0.13	0.92	0.19	0.03

Figure 5: a) Probability density of each torsional angle for crystal conformer 1 (dashed lines) and cluster centre 8 (thick line). The numerical distance between distribution peaks is provided in the table below. b) Comparison of crystal conformer 1 structure and the representative structure for C8 cluster group.

plied, the clustering algorithm determined 12 cluster centres as shown in the decision graph in Figure 4. Each cluster centre corresponds to the most representative conformer structure for those belonging to a cluster (core and halo). A visual representation of the molecular shape corresponding to each cluster centre is displayed around the decision graph. The conformers identified at this stage are labelled C1 to C12, each of which has an associated characteristic fingerprint in torsional angle space as discussed in Section D of the Methods. All fingerprints can be found in Section D of the SI.

Before proceeding further into the analysis of the conformational population of sildenafil in solution, it is useful to compare the structures of the cluster centres identified in vacuum and those of the crystal conformers. To this aim, a distance matrix comparing crystal conformer 1 and 2 found in form I and each of the 12 cluster centre structures is generated as shown in Figure 4. In the plot, the colour scale corresponds to the distance in torsional angle space, where white signifies the lowest distance, i.e the most similar structures. The distance matrix was obtained by taking the most probable value for each angle and calculating the distance between points in the 6D space of all dihedrals.

The distance matrix reveals that conformer C8 is the most similar structure to crystal conformer 1, while C2 is the most similar to conformer 2. The fingerprints corresponding to C8 and conformer 1 structures in torsional angle space are visually compared in Figure 5a. The two plots show that the conformational rearrangement of C8 into crystal conformer 1 would involve readjustment in torsional angles τ_A, τ_B and τ_D, which display a comparatively broader distribution in vacuum. The overlap in the distributions demonstrates that the exact crystal conformer is found in the gas phase, and it accounts for only 3% of all configurations. Figure 5b shows a comparison between the cluster centre structure of C8 and crystal conformer 1 as taken from the CCDC database, prior to performing MD. The overlap demonstrates the readjustment in τ_A and τ_D necessary to convert C8 into conformer 1. Torsional angle τ_E differs significantly according to the figure, however, as discussed, it has a moderate flexibility in the

solid and it relaxes to a configuration closer to that of C8 during MD as shown in the probability distribution in Figure 5a in purple.

The second most similar to the crystal group of structures are conformers C10 and C12 which relate to respectively conformers C2 and C8 via a rotation of torsional angle τ_E.

Conformational Isomers of Sildenafil in Solution

Equilibrium Probability

MD simulations in combination with WT-metaD were used to investigate the conformational rearrangement of sildenafil in three different solvents - acetonitrile, 1-butanol and toluene. The biased MD trajectory of each solvent case was analysed using the characteristic fingerprints associated with structure clusters C1 to C12, following the procedure outlined in the Methods. The probability of each conformer cluster was generated by accounting for the deposited bias potential through the reweighting procedure discussed earlier.

The obtained probability for all solvent cases is shown in Figure 6. The results show that, 95% of configurations in solution are accounted for, indicating that the proposed procedure of identifying conformational structures via unsupervised clustering is a fast and reliable way of determining conformational configurations in solution for systems of moderate to high degree of conformational complexity. Examining the distribution, small but significant variations in the probability in different solvents are observed. These findings correlate with what was observed in in our previous work for the case of ibuprofen.[6] Structure types C2 and C8 are found to account for 15 to 20% of the structures in solution each. As discussed, these two groups represent the conformers with a structure closest to the crystal-like configuration of sildenafil. Furthermore, C10 and C12, which relate to C2 and C8 via a rotation along the fast-converting torsional angle τ_E, each account for further a 15% of conformational isomers in solution. Therefore, given the high flexibility of τ_E even within the crystal structures, overall 60 to 70% of the structures found in solution resemble configurations close to those observed in the crystal, inferring that at the crystal-solution interface 30% of structures will have to encounter a more significant conformational rearrangement to adopt a crystal-like configuration and promote crystal nucleation or growth.

Enthalpy and Entropy Contributions to Conformational Stability

In order to gain further insight into the conformational isomerism of sildenafil in solution, the free energy ΔG of each state is calculated, along with a breakdown into potential energy and entropy. In Figure 7 the relative free energy of each state with respect to structure C2 for each solvent can be found in blue. As expected, the free energy difference between the four structures dominating the probability density plot - C2, C8, C10 and C12 is less than 1 kJ/mol for all solvents. The difference in free energy between the latter group and the rest of the structures varies between 2.5 and 5.5 kJ/mol, with the exception of C6 for the case of acetonitrile. Despite the minor variations in ΔG between different solvent cases however, the potential energy and entropy reveal more significant differences induced by the solvent. The relative potential energy difference (with respect to C2) rarely exceeds 5 kJ/mol in the case of acetonitrile, which also translates into minor entropic contributions in most states. The exception is once again state C6 for which the free energy is dominated by configurational entropy. A significantly different observations can be made for the cases of 1-butanol and toluene, where the relative potential energy of states is much larger, varying between 5 and 13 kJ/mol. Equally, entropy contributions in these two solvents are much more substantial than for the case of acetonitrile, indicating that solute - solvent interactions are much more dynamical. This is particularly prominent for the case of 1-butanol and it is possibly related to the inherent flexibility of its aliphatic chain.

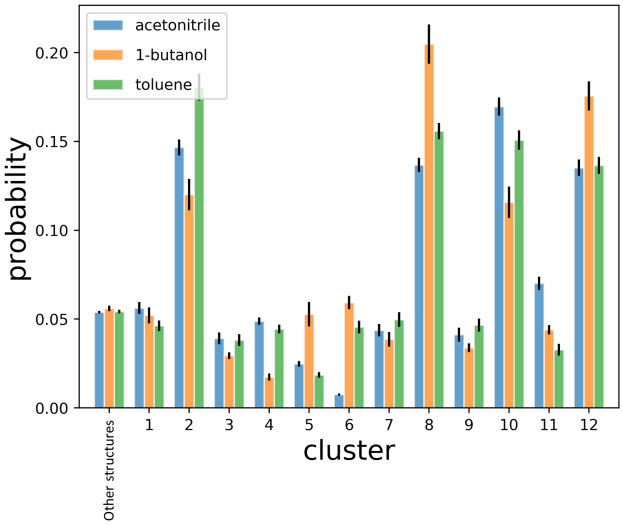


Figure 6: Probability of each conformer configuration C1-C12 along with crystal conformer 1 and crystal conformer 2 for all three solvents.

Conclusions

In this paper we develop a systematic approach to partition the configuration space of flexible molecules with an arbitrary number of rotatable bonds into conformational macrostates. The approach is based on the development of a distance metric between configurations that incorporates qualitative information on the energetic cost associated to transitions along each degree of freedom, and the subsequent application of unsupervised clustering. We apply this approach to investigate the conformational landscape of sildenafil in the crystal bulk, in the gas phase and in solution. A key aspect of the methodology introduced in this work is that the cluster centres are identified only once, for a reference state in the gas phase. These cluster centres configurations then used to classify configurations sampled in solution. This approach provides a self-consistent identification scheme for clusters in the condensed phase. Using this methodology, 95% of structures in three different solvents are unambiguously assigned to a cluster, demonstrating the effectiveness of the proposed classification procedure. We demonstrate that this classification strategy can be coupled with reweighting strategies to compute the free energy of conformational states and to further decompose it into its enthalpy and entropy contributions. This analysis leads to new insights into the role of solvent in the definition of the conformational landscape of an organic molecule. It is found that, while the relative free energy variation between states in different solvents is limited, solvents cases 1-butanol and toluene cause an increase in the entropic contribution to the conformational free energy. Combining this approach with existing strategies for studying effects of conformational rearrangements on processes can prove invaluable in understanding the effect of conformational isomerism in the process of crystal nucleation, growth and dissolution for systems of any size and level of conformational complexity.

Conflict of interest

There are no conflicts to declare.

Acknowledgements

The authors acknowledge financial support by Pfizer. We are grateful to the UK Materials and Molecular Modelling Hub which is partially funded by EPSRC (EP/P020194/1) and the UCL High Performance Computing Facilities and associated support services for computational resources.

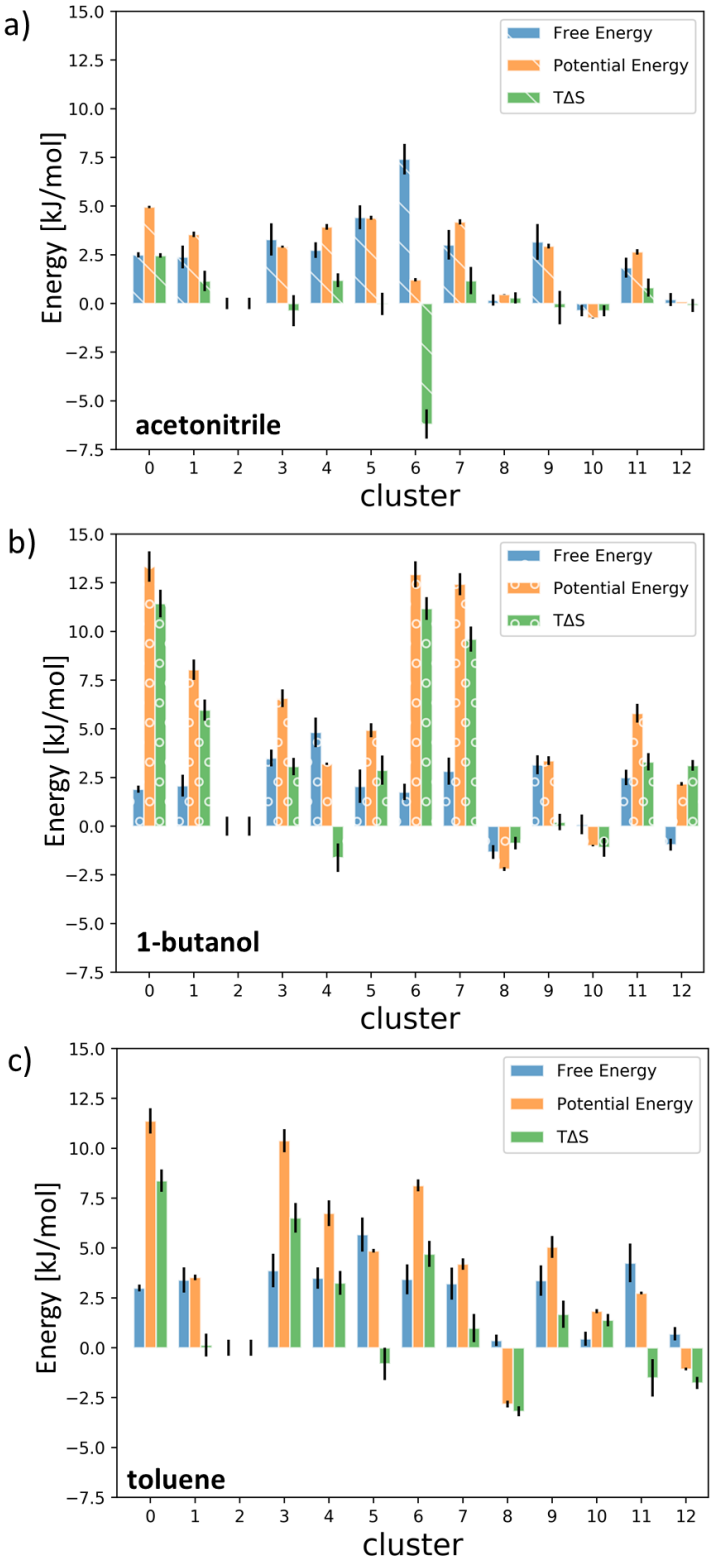


Figure 7: Relative free energy, potential energy and entropy of each cluster configuration with respect to C2 for the case of a) acetonitrile, b) 1-butanol and c) toluene.

References

(1) Amaral, M.; Kokh, D. B.; Bomke, J.; Wegener, A.; Buchstaller, H. P.; Eggenweiler, H. M.; Matias, P.; Sirrenberg, C.; Wade, R. C.; Frech, M. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nat. Comm.* **2017**, *8*, 1–14.

(2) Ertekin, A.; Massi, F. *eMagRes*; John Wiley Sons, Ltd: Chichester, UK, 2014; Vol. 3; pp 255–266.

(3) Saladino, G.; Gervasio, F. L. New Insights in Protein Kinase Conformational Dynamics. *Curr. Top. Med. Chem.* **2012**, *12*, 1889–1895.

(4) Bernstein, J.; Hagler, A. T. Conformational Polymorphism. The Influence of Crystal Structure on Molecular Conformation. *J Am Chem Soc* **1978**, *100*, 673–681.

(5) Cruz-Cabeza, A. J.; Bernstein, J. Conformational polymorphism. 2014.

(6) Marinova, V.; Wood, G. P.; Marziano, I.; Salvalaglio, M. Dynamics and Thermodynamics of Ibuprofen Conformational Isomerism at the Crystal/Solution Interface. *J. Chem. Theory Comput* **2018**, *14*, 6484–6494.

(7) Goldstein, I.; Burnett, A. L.; Rosen, R. C.; Park, P. W.; Stecher, V. J. The Serendipitous Story of Sildenafil: An Unexpected Oral Therapy for Erectile Dysfunction. 2019.

(8) Weise, C. F.; Weisshaar, J. C. Conformational analysis of alanine dipeptide from dipolar couplings in a water-based liquid crystal. *J. Phys. Chem. B* **2003**, *107*, 3265–3277.

(9) Lucaioli, P.; Nauha, E.; Gimondi, I.; Price, L. S.; Guo, R.; Iuzzolino, L.; Singh, I.; Salvalaglio, M.; Price, S. L.; Blagden, N. Serendipitous isolation of a disappearing conformational polymorph of succinic acid challenges computational polymorph prediction. *CrystEngComm* **2018**, *20*, 3971–3977.

(10) Granata, D.; Camilloni, C.; Vendruscolo, M.; Laio, A. Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics. *Proc. Nat. Acad. Sci.* **2013**, *110*, 6817–6822.

(11) Elts, E.; Greiner, M.; Briesen, H. In Silico Prediction of Growth and Dissolution Rates for Organic Molecular Crystals: A Multiscale Approach. *Crystals* **2017**, *7*, 288.

(12) Saxena, A.; Prasad, M.; Gupta, A.; Bharill, N.; Patel, O. P.; Tiwari, A.; Er, M. J.; Ding, W.; Lin, C. T. A review of clustering techniques and developments. *Neurocomputing* **2017**, *267*, 664–681.

(13) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28*, 100.

(14) Malmstrom, R. D.; Lee, C. T.; Van Wart, A. T.; Amaro, R. E. Application of molecular-dynamics based markov state models to functional proteins. *J. Chem. Theory Comput* **2014**, *10*, 2648–2657.

(15) Thayer, K. M.; Lakhani, B.; Beveridge, D. L. Molecular Dynamics-Markov State Model of Protein Ligand Binding and Allostery in CRIB-PDZ: Conformational Selection and Induced Fit. *J. Phys. Chem. B* **2017**, *121*, 5509–5514.

(16) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.* **2010**, *132*, 1526–1528.

(17) Wang, Y.; Makowski, L. Fine structure of conformational ensembles in adenylate kinase. *Proteins* **2018**, *86*, 332–343.

(18) Sun, L. W.; Li, H.; Zhang, X. Q.; Gao, H. B.; Luo, M. B. Identifying Conformation States of Polymer through Unsupervised Machine Learning. *Chinese J Polym Sci* **2020**, *38*, 1403–1408.

(19) Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci* **2015**, *2*, 165–193.

(20) Sittel, F.; Stock, G. Robust Density-Based Clustering To Identify Metastable Conformational States of Proteins. *J Chem Theory Comput* **2016**, *12*, 2426–2435.

(21) Frey, B. J.; Dueck, D. *Clustering by Passing Messages Between Data Points*.

(22) North, B.; Lehmann, A.; Dunbrack, R. L. A new clustering of antibody CDR loop conformations. *J. Mol. Biol.* **2011**, *406*, 228–256.

(23) Vlasblom, J.; Wodak, S. J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics* **2009**, *10*.

(24) Cheng, Y. Mean Shift, Mode Seeking, and Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 790–799.

(25) Fukunaga, K.; Hostetler, L. D. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Trans Inf Theory* **1975**, *21*, 32–40.

(26) Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496.

(27) Liu, S.; Zhu, L.; Sheong, F. K.; Wang, W.; Huang, X. Adaptive partitioning by local density-peaks: An efficient density-based clustering algorithm for analyzing molecular dynamics trajectories. *J. Comput. Chem.* **2017**, *38*, 152–160.

(28) Wang, G.; Bu, C.; Luo, Y. Modified FDP cluster algorithm and its application in protein conformation clustering analysis. *Digit Signal Process* **2019**, *92*, 97–108.

(29) Barbas, R.; Font-Bardia, M.; Prohens, R. Polymorphism of Sildenafil: A New Metastable Desolvate. *Cryst. Growth Des.* **2018**, *18*, 3740–3746.

(30) Barbas, R.; Font-Bardia, M.; Paradkar, A.; Hunter, C. A.; Prohens, R. Combined Virtual/Experimental Multicomponent Solid Forms Screening of Sildenafil: New Salts, Cocrystals, and Hybrid Salt-Cocrystals. *Cryst. Growth Des.* **2018**, *18*, 7618–7627.

(31) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

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

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

(34) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput* **2012**, *8*, 61–74.

(35) van der Spoel, D.; van Maaren, P. J.; Caleman, C. GROMACS molecule and liquid database. *Bioinformatics* **2012**, *28*, 752–753.

(36) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(37) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(38) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(39) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. App. Phys.* **1981**, *52*, 7182–7190.

(40) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(41) Bonomi, M. et al. Promoting transparency and reproducibility in enhanced molecular simulations. 2019; `https://nomad-coe.eu`.

(42) Branduardi, D.; Bussi, G.; Parrinello, M. Metadynamics with adaptive gaussians. *J. Chem. Theory Comput* **2012**, *8*, 2247–2254.

(43) Gimondi, I.; Tribello, G. A.; Salvalaglio, M. Building maps in collective variable space. *J. Chem. Phys.* **2018**, *149*.

(44) Kollias, L.; Cantu, D. C.; Glezakou, V.-A.; Rousseau, R.; Salvalaglio, M. On the Role of Enthalpic and Entropic Contributions to the Conformational Free Energy Landscape of MIL-101 (Cr) Secondary Building Units. *Adv. Theory Simul.* **2020**, *3*, 2000092.