Sampling and Classifying High-Dimensional Conformational Free Energy Landscapes of Active Pharmaceutical Ingredients

Alexandre Van Bronkhorst Ferreira

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Chemical Engineering

University College London

July 23, 2025

I, Alexandre Van Bronkhorst Ferreira, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

We present a new method for calculating the high-dimensional conformational free energy landscapes of flexible drug-like molecules. Using Density Peaks Advanced's density estimator, the free energy associated with individual configurations sampled from an enhanced sampling simulation can be calculated in a gridless manner, thus enabling the mapping of conformational ensembles in dimensionalities computationally inaccessible to grid-based methods. Due to the physics-based configurational sampling, conformers can be characterized by the configurations corresponding to the density peaks. The gridless nature of this method enables this characterization in the full dimensionality of a flexible molecule's conformation space. This method can produce per-point free energy maps, which enable the study of conformational interchanges in a level of detail previously inaccessible. This method is initially demonstrated on molecules with 2, 4, and 11-dimensional conformational spaces and is presented alongside a set of consistency checks which enable the quality of the high-dimensional results to be assessed.

Finally, to further demonstrate the utility of the method, a study of the conformational landscapes of 4 different molecules is presented. Each molecule is subjected to two distinct solvent environments, which have been linked experimentally to conformational changes in these systems. The subsequent impact on the high-dimensional free energy landscapes is explored through the use of Sketch-map projections and visual inspections of the conformers generated by the method.

Impact Statement

The work carried out in this project enables the study of conformational mechanisms at a level of detail previously only possible at high computational cost. One of

the key motivators for this work is the challenge of understanding the link between molecular flexibility and the polymorphism observed in pharmaceutical molecules. The tools developed in this work will enable the understanding of how specific environments affect the conformers assumed by pharmaceutical molecules and will enable these environments to be tuned, allowing for selective production of the desired product.

The time it takes to bring a new drug to market is lengthy, and the need for novel medicines and therapies is often urgent. Thus, tools which accelerate the drug design process will be of great benefit to society. Computational modeling tools, such as the one presented in this work, have the potential to enhance our understanding of molecular behavior and eliminate the time-consuming, trial-and-error-based methods which are so prevalent in drug design.

Acknowledgements

I am grateful to my supervisor, Prof. Matteo Salvalaglio, for his expertise, advice, and guidance throughout this project and for the education in computational science I have received under his supervision. Our many meetings were essential to the work presented here, whether we spent them confused by broken free energy surfaces or arguing over reweighing schemes.

I am also grateful to Dr. Rui Guo and Dr. Ivan Marziano for their support of this project, and to Pfizer Inc. for its funding and scientific contributions to this project.

I would like to thank Dr. Aaron Finney for helping me get to grips with life as a researcher during my first year in the group, and for his advice on life as a scientist in general.

I would like to thank my colleagues at the Molecular Modeling and Engineering group at UCL for all the ideas, help, and discussions they have given me over the past four years, whether crowded around a computer monitor or at a lunch table.

I would like to thank my family for their total support of my scientific ambitions, which I have been able to count on throughout my life, and for letting me move into their basement when my rent was raised.

And of course, I am deeply grateful to my partner, Jodi, who, for some reason, did not run away when I told her I wanted to pursue a PhD, and is still here to watch me submit this thesis.

I am in no way grateful to Foxtons Estate Agents or my former landlord in any way whatsoever.

Contents

List of Figures

List of Tables

Chapter 1

Context and Motivation

The role of flexibility in biology has long been known to be important. As early as 1958, it was proposed that molecular substrates interact with enzyme active sites not through rigid 'lock-and-key' interactions, but through an induced fit, in which the geometry of both the substrate and enzyme dynamically react to one another, resulting in the binding interaction and the specific biological function of the enzyme [2]. It is therefore not surprising that modern pharmaceuticals are often flexible small molecules with variable geometries [3]. This flexibility poses challenges to the pharmaceutical industry, as highly flexible molecules do not limit their variations in geometry solely to biological contexts, but rather exhibit diverse conformational landscapes in many environments, including the gas phase, solution phase, and solid state. That an active pharmaceutical ingredient (API) can assume different conformations when organized into a crystal lattice, a phenomenon called conformational polymorphism [4, 5], is of chief interest to the pharmaceutical industry. The pharmacological properties of APIs can depend very strongly on the crystal's polymorphism, and understanding an API's polymorphism has significant biological, regulatory, and financial implications on the drug design and manufacturing process [6].

Lucaioli et al. explored the conformational polymorphs of succinic acid in a 2018 study [7] and found that a novel conformational polymorph of succinic acid crystallized out of a methanol solution that was laced with a series of impurities. The novel polymorph was folded, in contrast with the two previously observed poly-

morphs, which exhibited succinic acid in a flat conformation. The study further establishes through molecular dynamics (MD) simulations [8] that the conformation of the novel polymorph is the dominant conformer in water, and crystal structure prediction (CSP)[9] determined that the novel polymorph was more thermodynamically stable than either of the previously observed polymorphs. It is remarkable that even for a molecule as simple as succinic acid, there is no clear relationship between the conformational distribution in solution, the thermodynamic stability of the conformational polymorphs in theory, and the prevalence of experimentally observed conformational polymorphs.

Marinova et al. have undertaken two studies on the conformational landscape of ibuprofen, a pharmaceutical molecule of great importance [10]. They defined the conformation space of ibuprofen in terms of two freely rotating dihedral angles, here termed torsions, in the molecular structure. In the first study [11], the conformational landscape of ibuprofen in a variety of solvents was explored through MD simulations. The simulations are carried out in bulk solvent, and then with an ibuprofen molecule simulated in proximity to a slab of crystalline ibuprofen. The study found that the extent to which the solvents hinder and promote conformational pathways was exacerbated by the presence of the crystal face. The second study [12] extended this approach to an analysis of ibuprofen's conformational distribution in the proximity of all of ibuprofen's crystal faces, finding that conformational dynamics is impacted not only by solvent or crystal face choice, but rather the combination of the two.

The cases of succinic acid and ibuprofen demonstrate that the nature of a molecule's conformational dependence on its environment can be extremely complex. However these studies focused on changes in geometry brought about by changes in a small number of torsions. In the case of very flexible molecules, wherein geometries are the function of a large number of torsions, the conformational space itself is complex and high-dimensional. Marinova et al. combined enhanced sampling MD with unsupervised clustering methods to study the conformational space of sildenafil [13], another highly relevant pharmaceutical product

[14]. Sildenafil's conformation is described in terms of six torsions, making its conformation space more complex than that of ibuprofen or succinic acid. As they did to ibuprofen, Marinova et al. ran enhanced sampling MD simulations of sildenafil in a variety of different solvents. Unlike ibuprofen, sildenafil's 6-dimensional conformation space is both impossible to visualize and computationally challenging to analyze. Density-based clustering, an analysis tool suited to high-dimensional datasets, was therefore used to partition the conformation space into distinct conformers, and inform estimates of the free energy.

It is clear that understanding the conformational landscapes of APIs and how they relate to their environments is an important and relevant challenge. Inspired by the examples highlighted above, the aim of this project is to create general workflow for the study of these conformational landscapes. As in the approach of Marinova et al. above, a molecule's conformation is defined using the values of the molecule's torsions [4] [13] . The conformation space is therefore always bounded and periodic in all dimensions. Veber's rules [15], a set of heuristics originally designed to predict whether a molecular structure would possess pharmacological properties, can be used to identify the relevant torsions in any given molecular structure. Characterizing conformers through the values of a set of torsions is not without precedent, and a number of approaches are based on this definition. For example, Torsiflex[16] is a software package that aims to explore a single molecule's potential energy surface utilizing a semi-random exploration of conformational spaces defined by torsions. Other methods for studying the conformational distribution of flexible molecules depend on quantum mechanical (QM) calculations [17, 18], machine learning algorithms trained on structural data (GOAT) [19], or complex combinations of MD and QM calculations (CREST) [20].

Rather than relying on stochastic algorithms for conformation space exploration, MD simulations are used throughout this project as a means of sampling the conformational spaces of the subject molecules. MD sampling is directly physics-inspired, so the datasets generated through these simulations are rich in conformations that are physically relevant to the system within the simulated environment.

The relatively low cost of MD enables long simulation times in any range of explicitly simulated environments. MD can also be augmented using enhanced sampling methods, which can be tailored specifically to promote the exploration molecular conformation spaces.

This project also continued the work of Marinova et al. by implementing density-based clustering tools in the analysis of high-dimensional conformational datasets, which reduce the computational cost of the analysis drastically as well as enables a more human-intuitive understanding of these high-dimensional spaces. The key relationship in this project is the natural partnership between MD simulations and density-based clustering. Evaluating the density of a configuration in conformation space where the configurations are sampled by MD is equivalent to evaluating the relative probability associated with the region occupied by the conformation. This is a direct result of the physics-inspired sampling of MD and makes possible the calculation of free energies assigned directly to conformations. Enhanced sampling methods may distort the underlying physical distributions, but the same reweighing methods [21] which exist to correct biased distribution on grids can be applied to the densities instead. As will be demonstrated here, by treating configurations sampled by MD as data points in conformation space in lieu of grid points, higher-dimensional free energy landscapes become accessible which preserve a high 'resolution' in the more relevant, low free energy regions of conformation space.

The treatment of molecular conformation spaces as high-dimensional periodic spaces defined by torsions is very general, and this project aims to present an approach which is applicable to the study of a wide range of molecular systems in any environment accessible by the wide range of MD simulation techniques. In this spirit, all the techniques presented in this work are implemented as open source software, available at https://github.com/ucecvan/Twister. Included in this repository is an accessible tutorial, enabling the application of this work to novel systems. A static version of this tutorial is shown in Appendix C.

Chapter 2

Theoretical Background

2.1 Molecular Dynamics

This section introduces the key concepts behind Molecular Dynamics (MD), the sampling method used in this project. The term Molecular Dynamics describes a family of methods for running computer simulations of physical systems. In an MD simulation, a computational model representing a physical system is evolved in time according to the laws of classical mechanics [8]. The simulation is carried out by an MD engine, which keeps track of the position and velocity vectors of all the particles (often atoms) in the simulation at every frame, and uses the forces experienced by the particles at the latest frame to predict how the particles will move in a given amount of time, generating the next frame. The time between frames is referred to as the timestep (dt), and the record of the position and velocity vectors of all particles in the system at every frame is referred to as a trajectory. The forces experienced by particles are determined by their interactions with one another, which are a function of their identity and relative position in space. These functions are referred to as the force field used by an MD simulation [22]. The following sections will explore each of these elements of an MD simulation, with a focus on the specific methods used during this project.

2.1.1 The First Frame

As described briefly above, MD evolves a system in time by considering the forces brought about by the position vectors of all the system's particles. Choosing an

initial set of positions and velocities for the particles in a system is therefore an important part of the process. For some simple systems, selecting the positions of all the particles could be as simple as manually specifying each atomic position, but for more complicated systems, such as simulations of solvated systems, energy minimization techniques are used to steer the system away from unphysical starting states (such as those with overlapping atom positions or impossible bond angles) by finding a local minimum in the potential energy of the system. The forces acting on the system of N particles in terms of the potential energy is given by[22]:

$$F(r) = -\frac{dV(r)}{dr}, \tag{2.1}$$

where r is the vector of all individual particle positions r_i, $F(r)$ is the vector of all the forces f_i experienced by the particles in the system. $V(r)$ is provided by the force field, which is covered in detail in section 2.1.3.

$$F(r) = \begin{bmatrix} f_{1x} & f_{1y} & f_{1z} \\ f_{2x} & f_{2y} & f_{2z} \\ \vdots & \vdots & \vdots \\ f_{Nx} & f_{Ny} & f_{Nz} \end{bmatrix} \tag{2.2}$$

In order to find a local minimum in the potential energy, steepest descent [23] (also called gradient descent) can be employed; from some starting configuration, the local gradient of the potential energy is calculated and the system is moved a small step down that gradient[24]. This process is repeated until a minimum is reached. In practice, this is done by considering the forces acting on each particle f_i, given by

$$f_i(r_i) = -\frac{\partial V(r)}{\partial r_i}, \tag{2.3}$$

Each particle i's position is moved in the direction of f_i according to

$$r_i^* = r_i + \frac{f_i(r_i)}{\max(F(r))}h, \tag{2.4}$$

where $\max(F(r))$ is the highest force experienced by any individual particle and h is the step size parameter. Following this, the potential energy of the new configuration, $V(r^*)$ is evaluated. If $V(r^*) < V(r)$, this configuration is accepted, and the process repeats from the new position, with h increased by a factor of 1.2. If $V(r^*) > V(r)$, this new configuration is rejected, and the process repeats from the previous configuration with h reduced by a factor of 5. The algorithm continues until interrupted or after F stops changing significantly with each configurational change. The resulting system should be in a configuration which corresponds to a local minimum in the potential energy surface and therefore serves a reasonable, physically viable starting point for a molecular dynamics simulation.

Steepest descent is one of several viable energy minimization algorithms which can be used to prepare a model system for MD simulation. It is described here because it is the method used throughout the project.

Once the initial positions of all the particles have been determined, their initial velocities must also be determined. This is done by selecting the system's temperature T, generating the Maxwell-Boltzmann distribution of kinetic energies [25] for each particle for that temperature and the particle's mass,

$$f(v_x, v_y, v_z) = \left(\frac{m}{2\pi kT}\right)^{3/2} \exp\left[-\frac{m(v_x^2 + v_y^2 + v_z^2)}{2kT}\right], \tag{2.5}$$

then randomly assigning the velocity components to each particle in accordance with that distribution, leaving each particle with velocity v_i. In equation 2.5, v_x, v_y, and v_z all correspond to the velocities in the x, y, and z planes, m is the mass of the particle, and k is the Boltzmann constant. Once positions and velocities have been assigned to all particles in the system, MD propagation can begin, evolving the system according to Newton's equations of motion.

2.1.2 Propagation Algorithms

From an initial configuration, an MD simulation evolves under a propagation algorithm, which uses the forces calculated by the force field, based on the positions of the particles in the initial frame, to compute what the positions and velocities of

the particles would be a short time dt later [26]. There are a range of propagation algorithms available, but the Leapfrog algorithm will be outlined here, as it is the algorithm used in the simulations carried out in this project. Firstly, for all particles, their net acceleration is calculated from their net force experienced f_i and their mass m_i,

$$a_i = \frac{f_i}{m_i}, \tag{2.6}$$

of each particle. Next, we describe the initial positions of all particles $r(t)$ as occurring at time t and the initial velocities $v(t - \frac{1}{2}dt)$ as being the velocities half of a timestep before t. A new set of velocities is described at $t + \frac{1}{2}dt$ by

$$v\left(t + \frac{1}{2}dt\right) = v\left(t - \frac{1}{2}dt\right) + dt \cdot a(t). \tag{2.7}$$

These new velocities are used to calculate new positions $r(t + dt)$ as

$$r(t + dt) = r(t) + dt \cdot v\left(t + \frac{1}{2}dt\right). \tag{2.8}$$

From the new positions, the force field reevaluates the forces on each particle and then equations 2.6 - 2.8 are repeated in order to evolve the system a further timestep. The process is repeated by a user-specified number of timesteps, with the result being a record of particle positions and velocities at each timestep, called frames. This forms a dataset commonly described as a trajectory.

It should be noted that under the Leapfrog algorithm, velocities and positions are known at intervals offset by $\frac{1}{2}dt$, hence the name of the algorithm. When the velocity of particles at positions $r(t)$ are needed, the velocities of the succeeding and preceeding half-step are averaged,

$$v(t) = \frac{1}{2}\left(v\left(t - \frac{1}{2}dt\right) + v\left(t + \frac{1}{2}dt\right)\right). \tag{2.9}$$

2.1.3 Force Fields

In the above sections, the forces acting on each atom and their use in evolving the positions and velocities of the system in time are frequently mentioned. These forces are calculated as a function of the positions of the simulated particles, as well as any specific interactions between them (bonds, etc.). Since a typical system of interest will consist of atoms of many different elements, arranged into a variety of environments, the number of parameters involved in replicating the number of interactions between all the atoms will be large. For this reason, numerous force fields have been developed, which contain all the potential functions and parameters necessary to evaluate how every particle in a simulation interacts with every other particle in given configuration, and the resulting potential of the system at said configuration [27] [8]. The force acting on a particle then corresponds to the first derivative of the potential at the position of said particle.

The force field used in this project is the General Amber Force Field (GAFF) [28], a force field developed to be applicable to small, organic molecules. GAFF is a Class I force field, which means it evaluates the potential of the system through a function of form

$$V(r) = \sum_{\text{bonds}} \frac{1}{2} k_r (r_{ij} - r_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta_{ijk} - \theta_0)^2 + \sum_{\text{torsions}} \sum_{n} k_{\phi,n} [cos(n\phi_{ijkl} + \delta_n) + 1] + \sum_{\text{non-bonded pairs}} \left[\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right] \tag{2.10}$$

The first term considers interactions between all atoms i, j which are joined by a bond and separated by a distance of r_{ij}. Here we we treat this interaction as a harmonic potential [29], with an equilibrium distance of r_0 and a force constant k_r.

The second term considers the potential as it varies with bond angles θ_{ijk} for all sequences of three sequentially bonded atoms i, j, k. This interaction is also treated as a harmonic potential in terms of the angle and has a force constant of k_θ and an equilibrium angle of θ_0.

The third term considers the potential as it varies with the torsions ϕ_{ijkl} between all sequences of four sequentially bonded atoms i, j, k, l. This is modelled with a cosinusoidal function which has an amplitude of $k_{\phi,n}$ and splits the torsion angle space into n minima.

The fourth term considers the potential between all atom pairs i, j except those which are part of the same molecule and separated by only one or two bonds. Otherwise, this term applies to all atom pairs, whether or not they are part of the same molecule. These are modelled by a Lennard-Jones potential [30], and a classical electrostatic point charge potential [25], with r_{ij} representing the distance between the atoms, $q_i\ q_j$ the charges on each atom, A_{ij} and B_{ij} are parameters related to the equilibrium distance between neutral atoms. ε_0 is the vacuum permittivity.

In equation 2.10, r_{ij}, θ_{ijk}, and ϕ_{ijkl} are functions of the positions of the atoms in the simulation. The remaining parameters are selected by the force field depending on the properties of the atoms themselves. GAFF, for instance, selects the parameters in equation 2.10 based on 35 basic atom types, which include 5 carbon types, 8 nitrogen types, 3 oxygen types, 5 sulfur types, 4 phosphorous, 6 hydrogen, and 1 each for fluorine, chlorine, bromine, and iodine, as well as a further set of advanced atom types. For elements with more than one atom type, the atom type, and therefore the parameter set, is chosen based on the environment of the atom. For instance, carbon atoms of different hybridizations are assigned different atom types [28].

2.1.4 Common Tools and Approximations

The above section outline what can be described as the essential aspects of a simulation: the use of a force field to describe the potential energy of a configuration, and the use of the resulting forces and a propagation algorithm to evolve the configuration in time by a small timestep. In practice, however, many additional tools and approximations are applied to a typical MD simulation in order to reduce the computational cost or better control aspects of the simulation. The most relevant methods to this project will be described in this section.

2.1.4.1 Periodic Boundary Conditions and Cutoffs

One challenge in MD is the simulation of bulk materials and solutions using a simulation box of limited size. Typically, this is achieved using periodic boundary conditions [31], wherein the simulation box is treated as a unit cell in an infinite lattice of replicas [8]. In practice this means that when an atom passes through a box boundary, it reappears on the other side of the box. This also means that atoms interact with each other across particle boundaries, and with their own replicas. It should be noted that periodic boundary conditions cannot perfectly replicate bulk conditions. For instance, if a system experiences fluctuations at a certain wavelength, these will be impossible to reproduce with a box smaller than that wavelength.

In equation 2.10, presented in **2.1.3**, the fourth term for bonded pairs considers non-bonded interactions between all possible atom pairs (excluding those separated by one or two bonds), regardless of the distance between them. This term in equation 2.10 is the sum of van der Waal's (Lennard-Jones) potential and a classical electrostatic potential, but here only the handling of the van der Waal's potential through a cutoff will be discussed (long-range electrostatics are handled using Particle Mesh Ewald[32], a method outlined in the next subsection). With periodic boundary conditions in place, considering all interactions regardless of distance becomes unfeasible, since the sum of every particle interacting with every other particle an infinite number of times cannot be practically computed. As interatomic interactions weaken as their separations increase, a cutoff distance r_c can be defined [33] beyond which we do not consider particles to interact with one another. In order to minimize errors, this cutoff distance needs to be large enough that the interactions are truly negligible at that distance. In the case of a Lennard-Jones potential, the long-range term is the attractive term which decays with r^{-6}. This rapid decay means a relatively short cutoff distance (in the order of 1 nm) can be used.

In order to ensure that the potential function is continuous, the value of the potential at the cutoff distance is subtracted from the entire potential

$$V_{tr}(r) = V_{nb}(r) - V_{nb}(r_c) \tag{2.11}$$

This leaves $V_{tr}(r)$ as the truncated potential, which ensures that the potential smoothly vanishes at the cutoff distance. Since the entire function is shifted by a constant amount, its derivative, and therefore the forces acting on the particle, are unaffected by the shift. In applying the cutoff, the van der Waal's force acting on each particle is limited to contributions from the nearest particles only. This is a reasonable approximation for evolving the position and velocities of each particle, but the use of a cutoff and the resulting shift will have an impact on the calculation of the overall properties of the entire system. This can be rationalized by considering that even though interactions between particles grow weaker with increasing distance, the number of neighbors on a spherical shell radius r increases with r^2. Thus while the interactions between distant particle pairs are small, they are large in number, and contribute meaningfully to the system's potential energy and pressure. Calculations of the system's potential energy and pressure must therefore be corrected to account for the missing long-range van der Waal's interactions[34]. This long-range correction for the potential energy takes the form

$$V_{LR} = \frac{1}{2}N\left(\left(\frac{4}{3}\pi\rho r^3 - 1\right)C_6 S - \frac{4}{3}\pi N\rho C_6 r_c^{-3}\right), \tag{2.12}$$

where the potential being treated is the attractive term of the Lennard-Jones potential, represented by $C_6 r^{-6}$, ρ is the average number density of the system, S is the amount the potential function has been shifted by, and N is the number of particles in the system. The first term accounts for the potential lost to the shift, while the second term corrects missing interactions beyond the cutoff.

The pressure of an ideal gas is given by

$$P = \frac{2}{3V}E_{kin},$$

and to correct for non-ideality brought about by van der Waal's interactions, the share of the system's kinetic energy which derives from the interaction of particles with each other through van der Waal's interactions must be accounted for,

$$P = \frac{2}{3L^3}(E_{kin} - \Xi), \tag{2.13}$$

Where Ξ is this share of the kinetic energy, known as the system's virial. The virial of a pair of particles is described by

$$\Xi_{ij} = -\frac{1}{2} r_{ij} F_{ij} = 3C_6 r_{ij}^{-6} \tag{2.14}$$

The virial's dependence on the intermolecular forces means it will also be affected by the cutoff, and so a correction term for the long-range virial is necessary to compute the pressure of the system. This correction takes the form

$$\Xi_{LR} = \frac{N\rho}{2}(4\pi C_6 r_c^{-3}) \tag{2.15}$$

2.1.4.2 Handling of long-range electrostatics with particle-mesh Ewald

As detailed above, the computation of long-range interactions can be controlled by introducing a cutoff distance beyond which the interactions are not considered to impact the evolution of the system, and the calculation of the overall potential can be corrected by adding a correction term to the total potential expression. Since this expression requires the evaluation of the integral of the potential up to an infinite distance, it is crucial that this integral converges with distance. As mentioned above, this is indeed the case for van der Waals interactions modelled with a Lennard Jones potential (see the fourth term of equation 2.10), but for electrostatic and dipolar interactions (also in the fourth term of equation 2.10), this is not the case [8]. Therefore, another method is needed to account for the electrostatic contribution to the total potential in an infinite system. This can be accomplished using a computational method called Particle-Mesh Ewald (PME) [32].

To begin, an Ewald sum will be introduced. The total Coulombic potential of a system of N point charges distributed inside a simulation box with periodic boundary conditions in place is given by

$$V_{coul} = \frac{1}{2}\sum_{i=1}^{N} q_i \phi(r_i). \tag{2.16}$$

The net charge of the system is 0. This is the sum of the potentials $\phi(r_i)$ at position of particle i, r_i, multiplied by i's charge q_i. These individual potentials are given by

$$\phi(r_i) = \sum_{j,n} \frac{q_j}{|r_{ij} + nL|}, \tag{2.17}$$

where the sum is over all neighboring particles j in all periodic images n. L is the length of the sides of the periodic box, assumed to be cubic in this case. This sum is not guaranteed to converge, and may converge slowly. To correct for this, a diffuse cloud of the exact opposite charge is centered on each point charge. These clouds diffuse according to a Gaussian distribution, and they serve to screen the point charge. Thus, any potential between point charges is a function of only the charge q which is not screened, a fraction which decays to 0 as the Guassian does. Due to the rapidly decaying nature of the potential between charges, the total sum can be computed as it must now converge. However, the effect of the screening charges must be accounted for, so another set of diffuse charges, also centered on the point charges is added with the same charge as the point charge, such as to compensate for the charge of the screening charge distribution. These three terms- the point charges, the screening cloud charges, and corrective cloud charges- sum together to the same charge distribution is the just the point charges. Because the rapidly decaying nature of the point charge is combined with the screening charge, it can be handled in real-space. The corrective charges therefore account for the long-range effect on the potential. For this reason, they are evaluated in reciprocal space, which is computationally advantageous. The total Coulombic potential energy function, combining these real and reciprocal terms, takes the form

$$V_{coul} = \frac{1}{2}\sum_{i\neq j}^{N} \frac{q_i q_j \mathrm{erfc}(\sqrt{\alpha} r_{ij})}{r_{ij}}$$

$$+\frac{1}{2L^3}\sum_{k\neq 0}\frac{4\pi}{k^2}|\rho(k)|^2\exp(-k^2/4\alpha)$$

$$-(\alpha/\pi)^{1/2}\sum_{i=1}^{N}q_i^2, \tag{2.18}$$

where α is a parameter that determines the width of the Gaussian shape of the screening and corrective charge clouds, ρ is the charge distribution in the reciprocal space, and k is the frequency vector in reciprocal space. The erfc function has the form $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-u^2}du$.

The first term of Equation 2.18 describes the potential of the point charges and their screening clouds. The interactions of each particle i with all other particles j calculated and summed. The factor of $1/2$ accounts for the fact that adding up interactions for every particle doubles the number of interactions between particle pairs.

The second term of Equation 2.18 describes the potential of the corrective clouds. The charge distribution surrounding each particle repeats infinitely in every direction, due to the periodic boundary conditions. Therefore, this charge distribution is a periodic function and its Fourier transform can be used to evaluate its potential as a function of the frequency vectors of this periodic charge distribution.

The third term is a correction for the fact that by summing the first term and second term together, each point charge in the first term is interacting with its own corrective Gaussian in the second term. To correct for this, the interaction between an overlaid point charge and the Gaussian charge is subtracted for every charge in the system.

It should be noted that while both the first and second terms of equation 2.18 converge with increasing r and k, respectively, the rates of these convergences have opposite dependencies on the widths of the Gaussian-shaped charge clouds, α. Decreasing α leads to more rapid convergence of the first, real space term, but slower convergence of the second, reciprocal space term. Therefore, in all practical methods based on the theory of Ewald sums, a compromise value of α must be selected.

In practice, Ewald summation is applied to the long-range interactions in the

MD simulations carried out in this project through Particle-Mesh Ewald (PME). In PME, the first term of equation 2.18 is calculated with a distance cutoff, similarly to how van der Waal's interactions are treated. This is possible because the combination of a screening charge and a point charge has a potential that decays rapidly with distance. The second term of equation 2.18 is treated similarly, but with a cutoff in reciprocal space instead. This is possible because high values of k contribute little to the overall long-range potential of the charge distribution. In PME, the charge distribution itself is assigned to a fine but discrete mesh, instead of being distributed freely through space. This is because computing the Fourier transform $\rho(k)$ of the charge distribution can be performed relatively cheaply using the fast Fourier transform (FFT) algorithm.

2.1.4.3 Bond Constraints

Simulations carried out throughout this project utilize the LINCS algorithm[35] for constraining all bonds terminating in a hydrogen atom. The low mass of a hydrogen atom results in high vibrational frequency when modeled with a classical potential. Replacing this with a static bond can be more accurate in replicating the properties [35] of a true bond, and allows a longer timestep to be used when modeling the system. Since the vibrational modes of the H-bonds in the systems studied here are not of interest, LINCS can be implemented without significant compromise on the dynamics of the system. LINCS provides a framework for handling the evolution of the system when certain bond lengths must be preserved after every evolution of the system, regardless of the forces experienced by the atoms. In a system with K constrained bonds, constraint functions are defined as

$$g_i(r) = |r_{i_1} - r_{i_2}| - d_i = 0, i = 1, 2, ..., K, \tag{2.19}$$

where r_i represents the positions of the two atoms in bond i and d_i is the constrained bond length. These constraints are worked into the equation of motion for the system,

$$\frac{d^2r}{dt^2} = -M^{-1}\frac{\partial}{\partial r}(V - \lambda g), \tag{2.20}$$

where g is the vector of all of the constraint functions, V is the unconstrained potential, and M is the matrix of atomic masses. It can be seen that λg is acting in the manner of a constraint potential, where the parameter vector λ carries the magnitude needed for each constraint to cancel out the potential V and preserve the bond distance. The quantity B is now introduced:

$$B = \frac{\partial g}{\partial r},$$

so that the dot product $B^T\lambda = \lambda \cdot B$ can be seen as analogous to a constraint force, with B containing the directions of all the constrained bonds. An expression for the constraint force can be deduced to be

$$\begin{aligned} B^T\lambda &= -B^T(BM^{-1}B^T)^{-1}\frac{dB}{dt}\frac{dr}{dt} \\ &= MT\frac{dB}{dt}\frac{dr}{dt}. \end{aligned} \tag{2.21}$$

Where T is shorthand for $M^{-1}B^T(BM^{-1}B^T)^{-1}$. This leads to a new, constrained equation of motion:

$$\frac{d^2r}{dt^2} = (I - TB)M^{-1}f - T\frac{dB}{dt}\frac{dr}{dt}, \tag{2.22}$$

where I is the identity matrix. This constrained equation of motion can be used to determine the propagation equations for the leap-frog algorithm (for details on the unconstrained leap-frog algorithm, see section 2.1.2, in particular equations 2.7 and 2.8). The positions are propagated as follows:

$$r(t+dt) = (I - T(t)B(t)\left(r(t) + v\left(t - \frac{1}{2}dt\right)dt + M^{-1}f(t)dt^2\right) + T(t)d \tag{2.23}$$

where the atom positions are first evolved according to the unconstrained po-

tential, before setting the constrained bond lengths in direction $B(t)$ to the distance d. Velocities are then propagated as follows:

$$v\left(t+\frac{1}{2}dt\right) = \frac{r(t+dt)-r(t)}{dt}. \tag{2.24}$$

2.1.4.4 Thermostats

To better replicate experimental conditions, simulations carried out in this project are often run in the NVT ensemble (also called the canonical ensemble). This involves the use of a thermostat, which replicates the effect of coupling the system to a heat bath, allowing the exchange of energy in order to keep the system at a constant temperature. Commonly used thermostats include the Berendsen thermostat [36], the Langevin thermostat [27] and the Nose-Hoover thermostat [37, 38]. The thermostat used for simulations in this project is the Bussi-Donadio-Parrinello, or Bussi thermostat [39]. The temperature of a system is related to its kinetic energy by

$$K = \frac{N_f}{2\beta}$$
$$\beta = (k_B T)^{-1},$$

where N_f is the system's number of degrees of freedom, T is the absolute temperature, and k_B is the Boltzmann constant. The Bussi thermostat modifies the velocities of particles in the system such that the average kinetic energy of the system will reflect the target temperature. This is done by multiplying the velocities of all particles by a scaling factor

$$\alpha = \sqrt{\frac{K_t}{K}}, \tag{2.25}$$

where K is the current kinetic energy of the system, and K_t is a kinetic energy randomly sampled for the canonical distribution of kinetic energies:

$$P(K_t) \propto K_t^{(N_f/2)-1} e^{\frac{-K_t}{k_B T^*}}, \tag{2.26}$$

where T^* is the target kinetic energy. Applying the scaling factor α over the course of the simulation will result in an average kinetic energy which corresponds to the target temperature. Because each time the scaling factor is applied the canonical distribution for kinetic energies is being sampled, the canonical distribution in kinetic energies is preserved. Additionally, since the same scaling factor is applied to all velocities, this thermostat does not affect the bond lengths of constrained bonds. This preservation of the system's dynamics is crucial for this project, since the principal aim is to study how the molecule's geometric configurations are physically sampled.

2.1.5 Calculating a Free Energy Surface from a Molecular Dynamics Trajectory

The overall free energy of a system[27] in the NVT ensemble is given by

$$F = -k_B T \ln(Q), \tag{2.27}$$

where k_B is Boltzmann's constant, T is the absolute temperature, and Q is the partition function of the system:

$$Q = \sum_i \exp(-E_i/k_B T), \tag{2.28}$$

Where E_i is the energy of state i, and the Boltzmann factor is the exponent being summed over all microstates. It is possible to evaluate the free energy of a subset of the states, by only considering those microstates in equation 2.28. Each microstate of the system corresponds to a unique set of atomic positions r_i, but these microstates can be classified by defining a function of these atomic positions $S(r)$, known as a collective variable (CV), and grouping microstates which lead to an equivalent value of $S(r)$.

The probability of finding the system in state j is related to the partition function by:

$$P_j = \frac{\exp(-E_j/k_BT)}{\sum_i \exp(-E_i/k_BT)} = \frac{\exp(-E_j/k_BT)}{Q}. \tag{2.29}$$

This link between the probability of encountering a microstate and its free energy is what allows for the use of MD for the estimation of free energy surfaces[21]. If two discrete regions m and l in $S(r)$ are considered, the free energy difference between regions m and l are given by:

$$F_m - F_l = -k_BT\ln(P_mQ) + k_BT\ln(P_lQ) = k_BT\ln\left(\frac{P_l}{P_m}\right). \tag{2.30}$$

Equation 2.30 demonstrates that the difference in free energy between two regions of CV space does not depend on knowledge of Q. Thus, applying equation 2.27 to the probability distribution in terms of S, $P(S)$ yields a free energy surface where the relative free energies between different regions are correct:

$$F(S) = -k_BT\ln(P(S)). \tag{2.31}$$

If the CV space is evenly split into discrete regions in $S(r)$, an MD trajectory can be used to build a normalized histogram $H(S)$ which track how many frames correspond to each region of $S(r)$. Each bin then corresponds to the probability of encountering the system within that region of $S(r)$.

$$F(S) = -k_BT\ln(H(S)). \tag{2.32}$$

Using equation 2.32 on a histogram constructed from an MD trajectory assumes that all regions of the CV space were visited in the course of the simulation, and that the simulation has run for enough time that the time spent in each region is proportional to the probability of that region being sampled. This condition, ergodicity, is not trivial: simulations may run for a long time and at great computational cost while stuck in local minima, leaving regions relatively unexplored or even completely unexplored at the end of the simulation. An entire branch of MD techniques, known as enhanced sampling techniques[40], have arisen to achieve ergodic simulations in reasonable timeframes. In this project, one such technique, Well-Tempered

Metadynamics (WTMetaD), is used to ensure ergodicity in simulations.

2.2 Well-Tempered Metadynamics

As described in Section 2.1.5, using a trajectory to calculate a FES requires the trajectory to have ergodically sampled the relevant CV space. For systems where this requires prohibitively long simulation times, a family of enhanced sampling techniques offer solutions which enhance the sampling of the CV space without necessitating longer simulation times. These include parallel tempering [41], bias exchange metadynamics [42], umbrella sampling [43], and steered metadyamics [44].

Here, Well-Tempered Metadynamics is used to promote sampling [45].The principle of the process is to add a biased potential term to the overall potential of the system, in such a way that frequently visited regions of CV space experience artificial forces which push the system into previously unexplored regions. In order to apply this bias, an appropriate set of CVs, $S(r)$ must be described which suitably split the system into regions of interest. The bias is defined in terms of these CVs, and ensures the system explores these CVs. The form of the bias is

$$V_B(S,t) = k_B \Delta T \ln\left(1 + \frac{\omega H(S,t)}{k_B \Delta T}\right), \tag{2.33}$$

where ΔT is a parameter with the units of temperature, ω is the energy rate, a parameter which determines how much potential is deposited over the course of the simulation, and $H(S,t)$ is the histogram of the system's configuration in CV space. The time-dependency arises since the bias potential is added as the simulation evolves, so the histogram is built as the simulation progresses. This bias potential is a monotonic function of $H(S,t)$, which means more bias will be present in the more frequently visited region, ensuring that its force displaces the system into unexplored regions. Because the potential is applied over time as the system evolves, the time derivative of the potential, the rate of deposition of the bias, is more relevant to the implementation of metadynamics,

$$\dot{V}_B(S,t) = \omega \exp\left(\frac{-V_B(S,t)}{k_B \Delta T}\right) \delta(S,t), \tag{2.34}$$

where $\delta(S,t)$ are delta functions in S and t, the time derivative of the histogram $H(S,t)$. In practice, delta functions are not used, and bias is instead deposited along CV space in a Gaussian shape,

$$\dot{V}_B(S,t) = \omega \exp\left(\frac{-V_B(S,t)}{k_B \Delta T}\right) W \exp\left(\frac{S(r) - S(r(t))}{2\sigma^2}\right), \tag{2.35}$$

where W is the height of the Gaussian and σ is the width of Gaussian. Equation 2.35 decays with $1/t$, and the deposited Gaussians are distributed in CV space $S(r)$ centered on the point in CV space then occupied by the simulation, $S(r(t))$. The time decay of Equation 2.35 means that the total deposited potential converges in the limit $t \to \infty$,

$$V_B(S, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(S) + C, \tag{2.36}$$

where T is the temperature of the system and C is a constant that can largely be ignored, since the relative free energy of different points on the surface is the main interest. The bias affects the probability distribution of S as follows

$$P(S) \propto \exp\left(-\frac{F(S)}{k_B(T + \Delta T)}\right). \tag{2.37}$$

Equation 2.37 demonstrates the utility of the ΔT parameter. At $\Delta T = 0$, Well-tempered metadynamics deposits no bias and the probability distribution follows on from section 2.1.5. As $\Delta T \to \infty$, the bias deposited by the system does not converge with time, corresponding to un-tempered, or standard metadynamics. Tuning ΔT allows control over the rate of decay of bias deposition, and thus the convergence of the bias.

Equation 2.37 allows for the recovery of $F(S)$, but the bias deposited in terms of S also distorts the probability distribution of variables $T(r)$ in the system which were not explicitly biased. Recovering the probability distribution of non-biased

variables requires reweighing the distribution of those variables to account for the bias[46]. The unbiased histogram of any variable can be recovered from the biased histogram by applying the factor

$$H_0(T(r)) = \exp\left(\frac{V_B(S(r), t \to \infty)}{k_B T}\right) H(T(r)), \tag{2.38}$$

which assumes that the final bias $V_B(S(r), t \to \infty)$ has been applied for the entirety of the simulation. This assumption becomes valid as simulations run for long after the deposited V_B has converged, the rate of which can be managed by modulating ΔT.

2.3 Unsupervised Clustering

Unsupervised clustering [47] refers to a family of computational methods which partitions a dataset into discrete regions, termed 'clusters'. The adjective 'unsupervised' refers to the fact that unlike other, supervised machine learning methods, the algorithm does not require training on a labelled dataset. Instead, unsupervised clustering detects some underlying structural pattern within the data itself.

This project uses unsupervised clustering on a dataset of a molecule's conformers obtained from an MD trajectory, where each conformer is described by the values of torsions in the molecule. Within this conformation space, more frequently observed conformations from the MD trajectory will form denser clusters of data points, while uncommon conformations will mean regions of conformation space remain sparsely populated. For this reason, a clustering method which separates data based on density would be capable of sorting the dataset into discrete conformers.

Density-based clustering techniques form a family of unsupervised clustering algorithms that group data points within spatial datasets into clusters based on the distance between data points within the data space. Algorithms in this family include DBSCAN [48] and Fast Search and Find of Density Peaks (FSFDP) [49]. There is precedent for the use of FSFDP in molecular conformation spaces, Marinova et al. used it to study the conformation space of Sildenafil[13]. The working

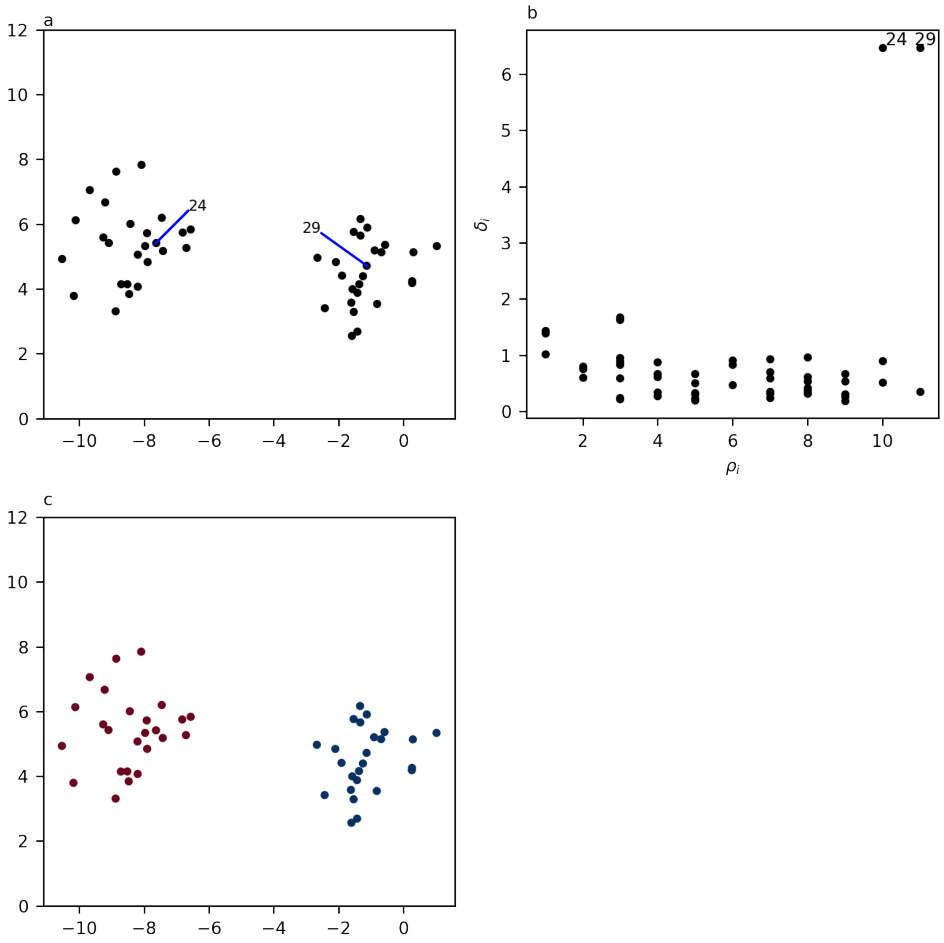


Figure 2.1: a: An example of a two-dimensional spatial dataset, where the data points are clearly separated into two clusters. The points representing the two density peaks, as identified using the decision graph, are indicated. b: the FSFDP decision graph ($d_c = 1$), where the indicated points 24 and 29 stand out clearly as the two density peaks. c: The data points are colored according their assigned cluster when points 24 and 29 are used as density peaks.

principles of FSFDP and its successor, Density Peaks Advanced (DPA)[50] will be outlined here by exploring how these algorithms operate on unbiased MD-generated conformational data.

2.3.1 Fast Search and Find of Density Peaks

FSFDP is a robust and computationally efficient density-based clustering method, suitable for the types of conformational datasets generated by MD. The algorithm determines, for each data point i in conformation space $\mathbf{S}$, a local density ρ_i, calculated by counting the number of other data points within a distance d_c of i. The

algorithm then calculates for each data point i the distance, δ_i, to the nearest neighbor of higher local density. The point with the highest local density in the entire dataset has no potential neighbor of higher density and is thus assigned a δ_i equal to the distance between itself and the furthest point from itself.

For most points, the expected value of δ_i is small, as there should be neighboring points of higher density. The exceptions to this will be points of extremely low local density, which will have long distances between themselves and any neighbors, and points which are the densest point in their neighborhood, and so represent density peaks. These density peaks are surrounded by points of lower density, and so must search into other neighborhoods to find points of higher local density. Therefore, these density peaks represent something of an anomaly: they will have a high ρ and a high δ simultaneously, and so can readily be identified by plotting ρ and δ for all points. This plot is called decision graph, and cluster centers are identified from the high-δ high-ρ regions of the decision graph. These cluster centers, once identified, become representative of each cluster. An example of this process carried out on a simple two-dimensional dataset is presented in Figure 2.1 with an example decision graph shown in Figure 2.1b

The remaining data points are assigned to the same cluster as that of their nearest neighbor of higher density. The clusters, now complete, are split into core and halo regions. For each cluster, this is done by identifying all points within d_c a point belonging to another cluster. These points are termed border points, and the density of the densest border point becomes the cluster's threshold density. All points denser than the threshold are considered part of the cluster core, while those with a lower density are part of the cluster halo.

This clustering algorithm is suitable for this project's applications for several reasons. Firstly, it clusters based on density peaks, which reflects the manner in which a molecular dynamics simulation samples conformation space. It also has the number of clusters arise naturally as part of the algorithm, instead of requiring it as a user input prior to clustering taking place, as required in non-density based algorithms, such as k-means clustering. In fact, the only user-specified parameter

at all is d_c, the distance used in local density calculations and in halo-core region assignment. However, FSFDP's use of a fixed radius in the calculation of a data point's density renders it less powerful in datasets where points populate regions with a range of densities. A successor algorithm, Density Peaks Advanced, uses a modified density estimator capable of adapting the size of the neighborhood considered when calculating each point's density.

2.3.2 Density Peaks Advanced

Density Peaks Advanced (DPA)[50], a successor to FSFDP, is the primary clustering algorithm used throughout this project. DPA estimates the probability associated with the ensemble of configurations projected in a given point i of the configuration space $\mathbf{S}$ using the PAk density esitmator[51]. Like the simple radius based density estimation in FSFDP, PAk is based on the Euclidean distances between point i and its nearest neighbors. Unlike FSFDP, the size of the hyperspherical neighborhood centered on i is determined by PAk for each individual data point. An underpinning assumption of this method is that the density is constant in the neighborhood of the point i. Hence, the number of nearest neighbors k is selected to be as large as possible to maximize the data used in calculating the local density while still representing a hypervolume of constant density. Each neighbor l can be said to occupy the volume v_l of the hyperspherical shell enclosed between hyperspheres of radii r_l and r_{l-1}. The sum of these volumes up to neighbor k is equal to volume V_k of a hypersphere with radius r_k. DPA leverages the fact that for a region of constant density, the volumes will be drawn from an exponential distribution [52] with a rate of this density ρ and that thus the log-likelihood function of ρ given a set of k neighbors is given by

$$L_{i,k}(\rho) = k\log(\rho) - \rho V_k, \tag{2.39}$$

The PAk estimator selects an appropriate k value using two models with distinct assumptions. Model one, $M1$, assumes that the densities of point i and its $j = k+1$ nearest neighbor are independent, while model two, $M2$, assumes these

densities are identical. Their log-likelihood functions are:

$$L_{M1} = k\log\left(\frac{k^2}{V_k V_j}\right) - 2k \tag{2.40}$$

$$L_{M2} = 2k\log\left(\frac{2k}{V_k + V_j}\right) - 2k. \tag{2.41}$$

The two models are compared with a likelihood ratio test [53],

$$D_k = -2(L_{M2} - L_{M1}) \tag{2.42}$$

which increases as the two models differ. If D_k grows over a threshold D_{thr} ($D_{thr} = 23.928$ according to Ref. [50]) then the densities of i and j cannot be considered constant. As such, PAk selects an appropriate k value by iteratively calculating D_k to increasing values of k until the threshold is passed.

Once k has been determined for all points, a density value ρ_i is known for each point through Equation 2.39.

In the classification step, DPA identifies peaks in the density as cluster centers. To avoid every fluctuation in density from being identified as a distinct peak, the DPA classifier is set to merge clusters which are separated by a saddle point between the density peaks with a density difference less than a user-specified threshold from the lower density peak. This provides an additional advantage over FSFDP, as a manual interpretation of the decision graph is no longer required, and classification is carried out fully automatically.

Once cluster centers have been determined, all remaining configurations are assigned membership to the same cluster as their nearest neighbor of higher density [50], as in FSFDP.

For both FSFDP and DPA, the most computationally expensive part of the process is the construction of a distance matrix which captures the distances between all pairs of data points. We therefore expect the cost of this approach to scale with the square of dataset size. As the dimensionality of the space increases, an additional term is added to the Euclidean distance calculation carried out. We therefore

expect the cost of the algorithm to scale linearly with dimensionality.

2.4 Two-Dimensional Projections with Sketch-Map

This project uses Sketch-map, a tool devised by Ceriotti et al [54], to project high-dimensional spatial datasets into two dimensions, which allows them to be visualized. The structure of a sample of configurations taken from an MD trajectory in a high-dimensional conformation space is of central interest to this project. Visualizing some representation of this structure would aid in enabling an intuitive analysis of these conformational spaces. Sketch-map has been designed with the specific purpose of creating a two dimensional projection of high-dimensional data generated from molecular simulations, aiming to preserve the local structure between isolated basins in the data space.

Sketch-map is based on an observation made when considering the histogram of pairwise distances for a set of molecular configurations distributed in a high-dimensional, periodic space. Figure 2.2 shows the distribution of pairwise distances for configurations of sulfadiazine in its 4 dimensional conformation space, as sampled by the simulations carried out in Chapter 4, but the shape of the histogram is largely representative of these types of datasets in general. Ceriotti et al. observe that the distribution assumes two shapes at either extreme of the distribution. Between 0 and 1 rad, the distribution takes on Gaussian form, while at distances longer than 3.5 rad, the shape of the distribution matches that of a uniform distribution of points in a periodic space[54]. Ceriotti et al. propose the Gaussian section of the distribution contains the distances between points within the same basins, and that the distances within the uniform distribution are those between points in two separate basins, distant from one another. Therefore, the intermediate stretch, between 1 rad and 3.5 rad, contains the distances between configurations in neighboring basins. Sketch-map seeks to create a two-dimensional projection of the high-dimensional dataset by preserving these intermediate distances. It does this by transforming the distance in both the original high-dimensional space and the low-dimensional projection using two sigmoid functions which reduce all short distances to 0 and extend

all long distances to 1,

$$F(R_{ij}) = 1 - (1 + (2^{a_D/b_D} - 1)(R/\alpha)^{a_D})^{-b_D/a_D}, \tag{2.43}$$

$$f(r_{ij}) = 1 - (1 + (2^{a_d/b_d} - 1)(r/\alpha)^{a_d})^{-b_d/a_d}, \tag{2.44}$$

where $F(R_{ij})$ maps the high dimensional distance R_{ij} to a sigmoid function, and $f(r_{ij})$ maps the low dimensional distance r_{ij} to a sigmoid function. a, b, and σ are the parameters of the sigmoid functions, determining the rate at which the function approaches 0, the rate at which the function approaches 1, and the distance at which the function is equal to 0.5, respectively. σ is the same for both $F(R_{ij})$ and $f(r_{ij})$, but the two functions have distinct a and b parameters, distinguished with a D or d subscript. These parameters are all user specified. The mapping from R_{ij} to r_{ij} is obtained by minimizing

$$\chi^2 = (\sum_{j \neq i} w_i w_j)^{-1} \sum_{j \neq i} w_i w_j [F(R_{ij}) - f(r_{ij})]^2 \tag{2.45}$$

where w_i and w_j are the weights on points i and j. In practice, minimizing the above equation scales in cost with the square of the number of data points within the dataset. Typically, a small sample of N landmark points are randomly selected from the datset and χ^2 for this landmark dataset is minimized to generate a set of low-dimensional landmarks. To ensure that the distances between points in regions of higher density are weighed more heavily, each landmark is assigned a weight w equal to the number of data points from the complete dataset within its Voronoi polyhedron within the landmark dataset [54].

Any non-landmark point X in the high-dimensional dataset can be projected into x in the low dimensional space by minimizing

$$\chi^2 = (\sum_{i=1}^{N} w_i)^{-1} \sum_{i=1}^{N} w_i [F(X - X_i) - f(x - x_i)]^2 \tag{2.46}$$

Where X_i is one of the high-dimensional landmark points and x_i is its low di-

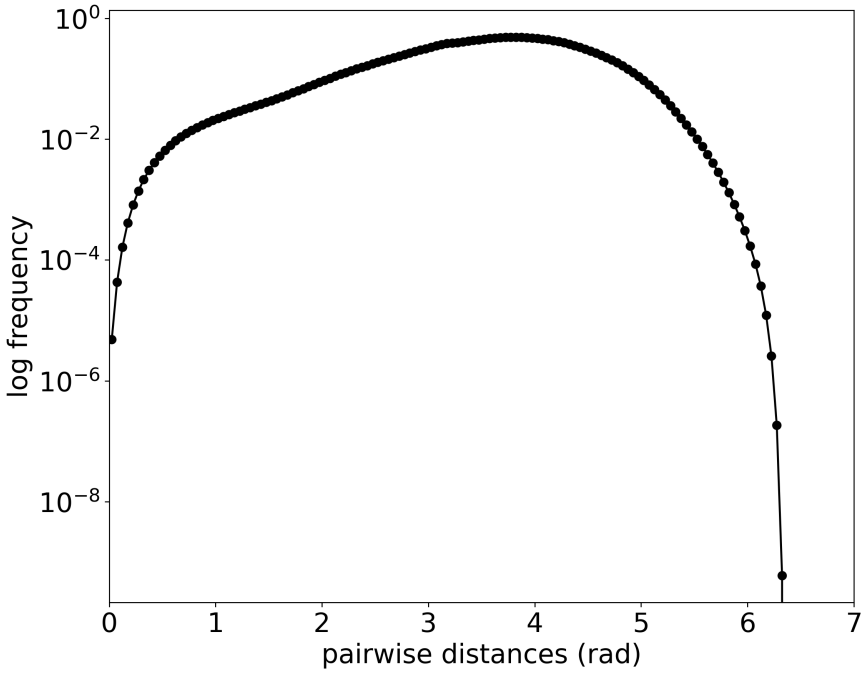


Figure 2.2: A histogram of the distribution of pairwise distances between configurations of sulfadiazine distributed in a periodice, 4-dimensional conformation space. The distribution resemble a Gaussian between 0 and 1 rad and a uniform distribution between 3.5 and 7 rad. The intermediate distribution is reflective of the structures of the basins within the conformation space, and it is these distances preserved by Sketch-map.

mensional projection. This is typically done for all remaining non-landmark points in the dataset.

Chapter 3

Methods

3.1 Summary of Approach

As an introduction to the workflow developed in this project, this section will provide a summary of the workflow, with subsequent sections covering the individual components in more detail [55]. To begin, the subject molecule's conformation is defined using the values of the molecule's torsions. [4] [13]. The space of possible conformations is termed the molecule's conformation space,

$$\mathbf{S} = [\gamma_1, \gamma_2, ..., \gamma_D],$$

where γ_i is one of the D torsions of a given molecule. The conformation space is therefore D-dimensional and always bounded and periodic in all dimensions.

Molecular dynamics (MD) simulations can be used to sample the conformational space and map the conformational FES of a given molecule. Sampling the probability distribution with MD offers two key advantages; MD sampling is inherently physics-inspired and allows for the analysis of the molecule in various environments and conditions. As discussed in Chapter 2, the physics-based nature of the sampling means the distribution sampled in conformational space by the MD simulation can be converted into a FES [21] through the relationship:

$$F(\mathbf{S}) = -k_B T \ln p(\mathbf{S}), \tag{3.1}$$

where $p(\mathbf{S})$ is the probability distribution in the conformational space, $F(\mathbf{S})$ is therefore the FES in conformational space $\mathbf{S}$, k_B is Boltzmann's constant, and T is the temperature.

To compute $F(\mathbf{S})$, it is thus necessary to obtain an estimate of $p(\mathbf{S})$, where all energetically relevant regions are ergodically sampled. For the ergodic sampling of a well-defined configuration space, metadynamics, which involves depositing penalty biases dynamically as the simulation proceeds to promote sampling[40], would be a typical approach. A description of well-tempered metadynamics is presented in Chapter 2. However, for the exhaustive sampling of a conformation space, this approach is limited by the computational feasibility of storing the bias values on a grid of the same dimensionality as the conformation space, which is the same issue faced with the conventional method of FES construction. For this reason, in practical applications, conventional metadynamics biases constructed in dimensionalities higher than three are very rare. Alternatives for high-dimensional CV spaces, such as bias-exchange metadynamics [42], have been developed. Still, these generally depend on the running of multiple replica simulations overseen by an exchange scheme. To widely sample conformational space with a single simulation, concurrent [56] well-tempered metadynamics (WTMetaD) [45] is used here.

DPA, developed by d'Errico et al., splits a set of data points distributed in space into clusters by grouping points within density peaks, a term referring to regions of high data density (N.B. In this work, the term 'density' refers to the density of data points in $\mathbf{S}$, unless otherwise noted). It does this partly by calculating the local density of every region centered on every single point in the dataset. This calculation is a function of the Euclidean distances between the point and its nearest neighbors. A full description of DPA is provided in Chapter 2. Here, the process is applied to a sample of N configurations in conformation space sampled by the MD simulation. These local density calculations are extremely powerful in this context for two key reasons: firstly, each additional dimension in $\mathbf{S}$ adds a single term to the Euclidean distance calculation, so the cost with increasing dimensionality scales linearly, and secondly these local densities map a distribution in much the same

way as the previously described histogram, so the same reweighing and inversion procedure may be applied to calculate the free energy. These free energies, unlike in the above histogram, are not associated with a defined region of conformation space. Rather, they are associated with a specific configuration sampled by the simulation. Thus, this *per point* FES has no fixed spatial resolution; data is rich in regions that have been heavily sampled and sparse in regions that have not been frequented. This is advantageous as it means that while data in the local minima remains highly dense due to the frequent sampling, little cost is incurred considering data from the rarely visited, largely irrelevant high-energy regions. This approach contrasts with the grid-based approach, where these high-energy regions are modeled in as high a resolution as the more relevant local minima. Figure 3.1 shows a summary of the workflow outlined here. Subsequent sections of this chapter will describe each stage in more detail.

3.2 Simulation Details and Enhanced Sampling Setup

3.2.1 Simulation Setup

As the high-dimensional free energy landscapes produced here depend solely on a configurational dataset defined by the sampling of a molecule's torsions, the analysis process subsequent to simulation is completely independent of the simulation environment. As such, vacuum simulations are sufficient to demonstrate the impact of conformational complexity on the resulting free energy landscapes, as is done in Chapter 4. The impact of solvation on the conformation free energy surface is explored in Chapter 5.

GAFF [28] force field parameters were used. The working principles of a Class I force field, such as GAFF, are presented in 2.1.3. For all simulations undertaken in this project, molecular structures were passed to Antechamber, part of the AmberTools [57] package, which assigned atomtypes to all atoms within the structures. Suitable GAFF force field parameters were then selected using LEaP, another component of AmberTools [57]. GAFF does not include an exhaustive selection

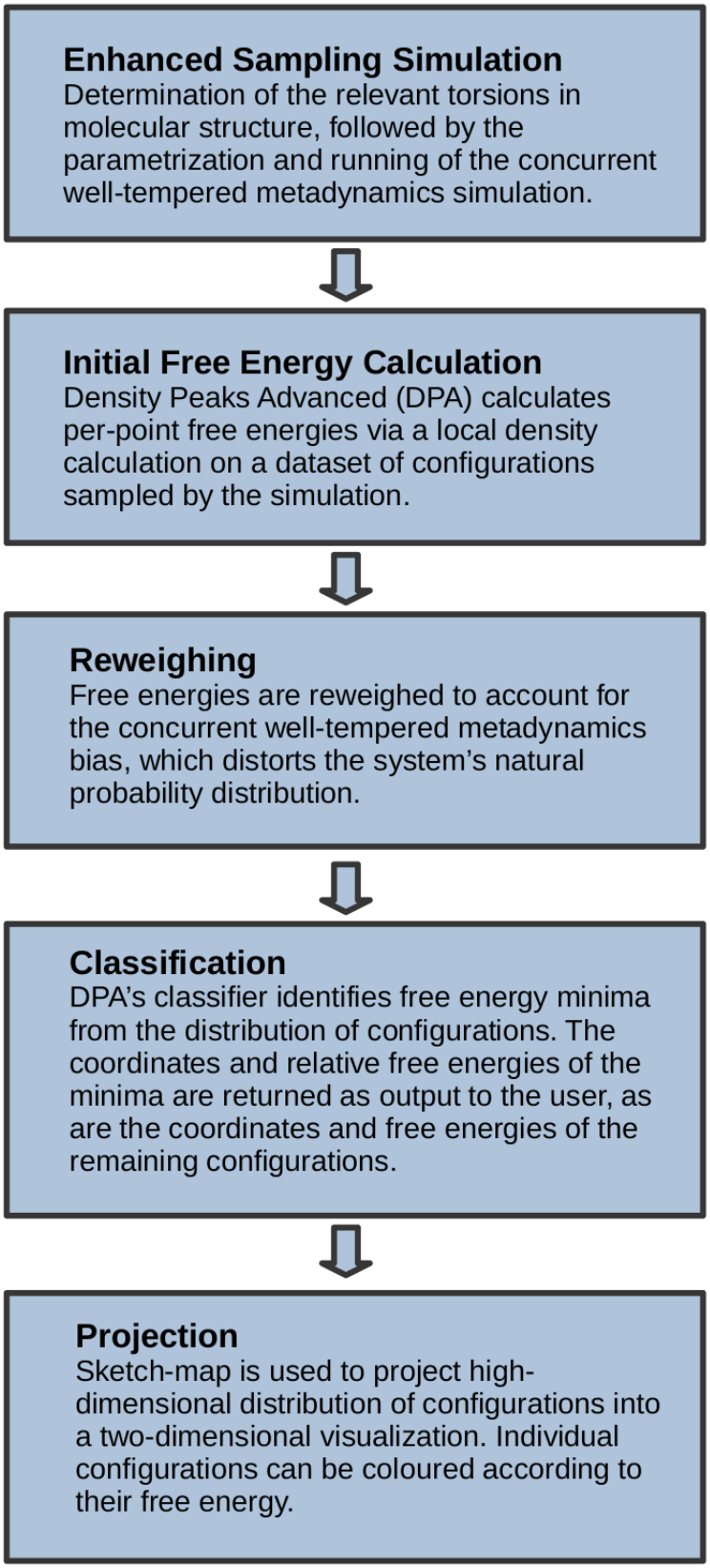


Figure 3.1: A summary of the workflow presented in this chapter. Note that each component operates independently, and the methodology used in each component can be modified or replaced if desired by the user.

of 4-atom torsion parameters. Where these were missing, GAFF's 2-atom (specific to the two central atoms) torsion parameters were substituted. As the simulations in this project serve chiefly to provide data to exhibit the density-based free energy estimation technique (see Chapter 4), and provide an example workflow for a set of computational experiments on single molecules both in vacuum and in solution (see Chapter 5), this degree of parameterization is sufficient. Note that the cost of the density-based analysis is completely independent of the level of theory used in the simulation.

Production simulations were run for 1 microsecond. GROMACS[23] was the MD engine used. WTMetaD was carried out using the Plumed [58] plugin for GROMACS. The simulations were carried out in the NVT ensemble, at a temperature of 300K, maintained using the velocity-rescaling thermostat developed by Bussi et al.[39]. The simulation box walls were set to be 5nm from the nearest atom of the simulated molecule upon initialization, and all three spatial dimensions were simulated using periodic boundary conditions. Short range van der Waal's and electrostatic interactions both had cut-offs of 1.5 nm. Long range electrostatics were handled with the Particle Mesh Ewald (PME) method [32], as implemented by GROMACS. All bonds terminating in a hydrogen atom were restrained with LINCS [35].

For simulations carried out in solvent, the simulation box was set up as above, and then populated with the solvent molecules. A steepest descent energy minimization was carried out, followed by two 1 ns equilibration runs. The first equilibration is conducted in the NVT ensemble, using the Bussi thermostat, and the second is carried out in the NPT ensemble, using the Berendsen [36] barostat. For all solvated simulations carried in water, the SPC/E water model was used [59].

For the determination of WTMetaD parameters, a short 10 ns unbiased simulation was run. The marginal FES in each torsion was computed on a 100-bin 1-dimensional histogram. A Gaussian Mixture Model (GMM) consisting of a sum of Gaussian terms was fitted to the resulting FES, and the smallest width parameter of one of the individual Gaussian hills, corresponding to the narrowest local mini-

mum in the marginal FES, was considered as the minimum reference width for the marginal under consideration. The width of the σ terms used to update the metadynamics bias was set to half of the minimum reference width, or 0.1 rad, whichever is larger.

3.2.2 Biasing in High Dimensional Spaces with Concurrent Metadynamics

The conventional method [60] for the construction of a 2D conformational FES is as follows. The WTMetaD biases are deposited in the 2D conformational space, defined by two torsions, ensuring that the entire space is fully sampled over the course of the simulation. The space is split into a 100 by 100 histogram. The distribution of MD configurations throughout the histogram follows the system's natural probability distribution as distorted by the metadynamic bias. The total bias deposited in each bin is known, allowing this distortion of the probability distribution to be reweighted. The resulting FES has a resolution equal to the fineness of the grid, in this case $(2\pi)/100$ rad. This methodology is robust but scales poorly to higher dimensional FESes. Both the bias deposition during the simulation and the estimate of the probability distribution require the construction of a grid with the same dimensionality as the conformation space. If the same resolution is desired, increasing the number of torsions incurs an exponential cost on computational resources, rapidly becoming unfeasible. We present these results in alanine dipeptide in Chapter 4 to facilitate comparison. Here, an alternative way of both biasing the simulations and modeling the high-dimensional probability distributions is proposed which does not depend on high-dimensional grids. Furthermore, methods for analyzing the resulting high dimensional FESes are explored, as their complete visualization is impossible beyond 2 dimensions.

To avoid the exponentially increasing costs of depositing WTMetaD biases in a high-dimensional conformational space, concurrent metadynamics [61] is used in its place to promote sampling. This entails simultaneously depositing a single one-dimensional bias for each torsion in the molecule, thus promoting exploration of the rotation of that torsion.[40]. The cost of this approach scales linearly with

dimensionality, as one additional monodimensional grid is needed for each additional torsion considered. The cost savings of this approach come with a trade-off; conventional metadynamics promotes the exploration of the entire conformational phase space and guarantees that previously visited configurations will be penalized accordingly. Concurrent metadynamics does not explicitly bias the combinations of any torsion values. It instead promotes the escaping from local free energy wells by driving the rotation of individual torsions.

3.3 Per-Point Free Energies with DPA and Biased Simulations

Following the end of the WTMetaD simulation, configurations from the simulation were paired with the total deposited bias in each dihedral at the corresponding position in conformation space, in line with the final bias approximation. This was achieved using Plumed, resulting in a dataset of configurations defined by their dihedral angles, accompanied by the final bias in each dihedral.

Using DPA's PA*k* density estimator on configurations sampled from a WTmetaD simulation produces densities that reflect a probability distribution perturbed by the applied biases. These densities can be reweighed using the Zwanzig approach [21] as:

$$\rho_i^* = \rho_i e^{\beta(\Sigma_{t=1}^{D} V_i^t)}, \tag{3.2}$$

where ρ^* is the reweighed density, β is equal to $1/k_BT$, V_i^t is the bias in torsion t, $\sum_{t=1}^{D} V_i^t$ represents the sum of the concurrent biases acting on the D torsions, for configuration i. Practically, we evaluate ρ, the biased density, from configurations generated in a quasi-static bias regime, as the bulk of the bias is deposited during the early stages of the simulation and the bulk of sampled configurations are visited when bias deposition is negligible. We therefore apply the *final bias* approximation to obtain a time-independent value of V_i^t acting on configuration i [46, 62, 63]. The free energy F_i, associated with configuration i (thus termed *per point*), is computed as

$$F_i = -k_B T \ln \bar{\rho}_i^*, \tag{3.3}$$

where $\bar{\rho}_i^*$ the smoothed density of point i. The procedure for determining the smoothed densities is presented in 3.6.1.

3.4 Conformer Classification

In the classification step, DPA identifies peaks in the density as cluster centers, i.e., distinct conformers. This operation is equivalent to identifying local minima in the D-dimensional FESes. For this step, we use the set of reweighted, smoothed densities $\bar{\rho}_i^*$.

Moreover, to avoid every fluctuation in density from being identified as a distinct peak, the DPA classifier is set to merge clusters separated by a saddle point between the free energy basins (a conformational transition state) with free energy less than 1 $k_B T$ higher than one of the cluster centers it connects. Once cluster centers have been determined, all remaining configurations are assigned membership to the same cluster as their nearest neighbor of higher density [50]. With all configurations classified, clusters with a population smaller than 1% of the total sample are discarded to avoid spurious clusters identified from anomalously isolated configurations.

The ultimate product of this process is a set of cluster center configurations representing the local minima of the D-dimensional conformational FES, called a cluster set. Each cluster's lowest free energy configuration, i.e., the cluster center, provides *the most* representative configuration of a given conformer.

3.5 Consistency Analysis

The potential high dimensionality of $\mathbf{S}$ makes the visualization of per-point free energies difficult. A series of checks on the ergodicity of the sampling and consistency of the DPA classification, therefore, provide confidence in the results.

Firstly, the convergence of the D monodimensional marginal free energy surfaces of each individual torsion is monitored to assess the ergodicity of the sam-

pling. For a simulation of length τ, convergence of the marginals is assessed by monitoring, on D histogams of n_{hist} points, the quantity:

$$\delta F_M(t) = n_{hist}^{-1} \sum_{i=1}^{n_{hist}} |F_i^\tau - F_i^t| \tag{3.4}$$

where simulation time t runs from $[0, \tau]$, $F(t)$ is a monodimensional FES obtained with data gathered up to time t, $F(\tau)$ is the same quantity computed with all the data available. This quantity represents the average free energy difference per histogram bin in any of the D monodimensional free energy surfaces.

Figure 4.2a displays an example of $\delta F_M(t)$ computed for ϕ and ψ torsional angles of alanine dipeptide during concurrent metadynamics. The flattening of these differences as the fraction of utilized trajectory increases indicates that the simulation has been run for long enough that these torsions have been ergodically sampled.

This check is computationally inexpensive and offers a first qualitative check on the quality of the configurational exploration obtained with concurrent WT-metaD. If the marginal FES associated with a torsion is still evolving rapidly at time τ, i.e., when the simulation ends, the sampling has not yet reached the ergodic limit with respect to the configurations discovered.

However, the convergence of 1D marginal FESes tells us little about exploring the conformational space in its full dimensionality. This is important as even substantial amounts of data may be distributed extremely sparsely in high dimensions.

As such, to build confidence in our results, we evaluate the statistical significance of the conformer classification as a function of the dataset size.

For this purpose, a consistency check has been devised, which offers a similarity score between two cluster-sets generated from different configurations. A cluster set generated from a dataset of size N, C^N can be compared with a reference cluster set C^{ref}, which is generated with the largest number of configurations feasible. Each cluster center C_i^N is matched with the nearest center in the reference set C_i^{ref}, according to the Euclidean distances in $\mathbf{S}$ between members of the two cluster-sets. This matching process is demonstrated in Fig. 3.2. Differences in free energy ΔF_i and position Δd_i are determined and averaged across all matched pairs as $\bar{\Delta F}^N$

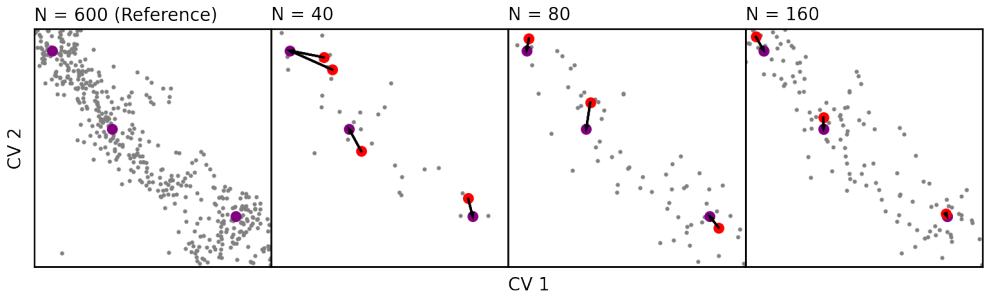


Figure 3.2: Sketch illustrating the pairing procedure used by the consistency metrics to compare cluster-sets generated from datasets of different sizes. The configurations shown are drawn from the simulation of alanine dipeptide but the principle illustrated is general. The cluster centers obtained from the largest amount of data form the reference set (shown in purple). Cluster centers generated from smaller amounts of data (shown in red) are paired to the nearest cluster center in the reference set, allowing comparison of distances and energy differences between cluster centers. It is expected that as dataset sizes grow, the positions and energies of the cluster centers will converge, as seen in Fig. 4.2c,d.

and $\Delta\bar{d}^N$. This comparison to C^{ref} can be repeated for cluster-sets generated from datasets of increasing N, and evolution of $\Delta\bar{F}^N$ and $\Delta\bar{d}^N$ with growing N can thus be assessed. Once datasets are large enough, the positions and relative free energies of minima would be expected to be independent of dataset size. The results of this analysis on the case of alanine dipeptide are shown in Figure 4.2b,c,d.

The evolution of $\delta F_M(t)$ for all torsions is presented for every molecule in this project, as are the evolutions of $\Delta\bar{F}^N$ and $\Delta\bar{d}^N$ as the number of configurations rise from 5000 to 45000 in increments of 5000. Also presented are the number of conformers detected by DPA at each dataset size.

3.6 Additional Heuristics

To augment the consistency and accuracy of the workflow described here, a set of additional heuristics were implemented.

3.6.1 Density Smoothing

To mitigate the noise introduced by exponential reweighting[21], the density of each point is averaged over a small hyperspherical domain. This step generates a new smoothed set of densities $\bar{\rho}_i^*$, at the cost of a small controllable loss in spatial

resolution. This is done by replacing each point i's density ρ_i^* with the average of all densities within a radius of 0.1 rad, $\bar{\rho}_i^*$. This mitigates the noise arising from the reweighing process, which can be quite substantial, since densities are multiplied by an exponential factor. Over the course of the project, a range of different radii were tested for smoothing, and 0.1 rad was settled on and used for all the results shown here. In the software package associated with this project, the radius is a user-set parameter and can be tailored to the desired application. The cost of addressing noise in this way is the loss of spatial resolution, which increases with the smoothing radius. Larger smoothing radii include a larger amount of data in the average, but risk factoring in data from regions far from the data point being smoothed. The default smoothing radius is 0.1 rad.

3.6.2 Cluster Merging

As previously mentioned, DPA uses a density peak merging procedure to prevent every small fluctuation in the local density from manifesting as a density peak. In a modification from the original DPA methodology [50], density peaks identified here were merged on energetic criteria: if two density peaks separated by a saddle point were distributed such that the difference in density between the lower-density peak and the saddle point was less than the equivalent of 1 k_BT in energy, the lower density peak would be merged into the higher density peak. This is functionally equivalent to merging free energy basins separated by a free energy barrier smaller than 1 k_BT in height. This is a natural way to sort significant density peaks from those arising from fluctuations, and is tunable. All the results presented in this work use a value of $1k_BT$ as the merging parameter, but different values can be used. Increasing the value of the parameter will result in a smaller number of conformers, which may result in better consistency when data is sparse. The trade-off for this gain in consistency is the potential loss of thermodynamically distinct conformers separated from more stable conformers by a relatively labile barrier.

3.6.3 Filtering and Sampling Heuristics

Two heuristics are used to filter less relevant data from the workflow, in order to optimize the usage of computational resources and eliminate less meaningful results. Firstly, when creating a sample of configurations from the MD simulation to analyze, the subset of configurations chosen are taken from the final two-thirds of the simulation. By not considering the first third of the simulation, we remove from consideration a stage of the simulation which is sampling the configuration space as significant amounts of metadynamic bias is being deposited, as this is incompatible with the final-bias approximation we utilize in our reweighing scheme. Because the number of configurations sampled by the MD simulation is much larger than the upper size limit of the configurational dataset to be analyzed, this can be done without compromising the size of the dataset.

Secondly, upon the first calculation of per-point free energies for the molecules studied in vacuum, all configurations calculated to have a free energy over 100 kJmol^{-1} are removed from consideration. Due to sampling promoted by concurrent WTmetaD, the sampling of such high energy regions are inevitable. Despite this, these configurations can be assumed to not be physically significant, and their presence incurs both increased computational cost and increases the level of noise encountered in the classification step. These high energy points are especially common for systems with high-dimensional conformation spaces, as isolated configurations may be very distant from any other points and thus have a very high free energy.

Chapter 4

Exploring Conformations in the Gas Phase

4.1 Method Validation: Alanine Dipeptide

The workflow[55] outlined in this work was first tested on the Ramachandran plot [64] of alanine dipeptide. This system was chosen for several reasons: the Ramachandran plot of alanine dipeptide is a commonly used model system in the field of MD and enhanced sampling techniques, making it one of the best-studied conformational FESes available. Additionally, its low dimensionality allows for both the visualization of the FES and access to more conventional methods of exploring this conformational space. The principal results of this are shown in Figure 4.1. The structure of alanine dipeptide, with ϕ and ψ indicated, is shown in Figure 4.1d. The conventional FES of alanine dipeptide, obtained through histogramming and reweighing of a trajectory generated through a WTMetaD MD simulation[60], is shown in Figure 4.1a. A per point FES, generated from the same simulation but with free energies calculated using the local densities of the sampled configurations, as discussed in the Methods section, is shown in Figure 4.1b. From a visual comparison, it clearly appears that the two FESes are in agreement, demonstrating that the reweighted-DPA density estimate leads to results virtually indistinguishable from standard histogramming-based approaches. Figure 4.1c shows a per point FES generated using a trajectory from a concurrent metadynamics simulation where ϕ

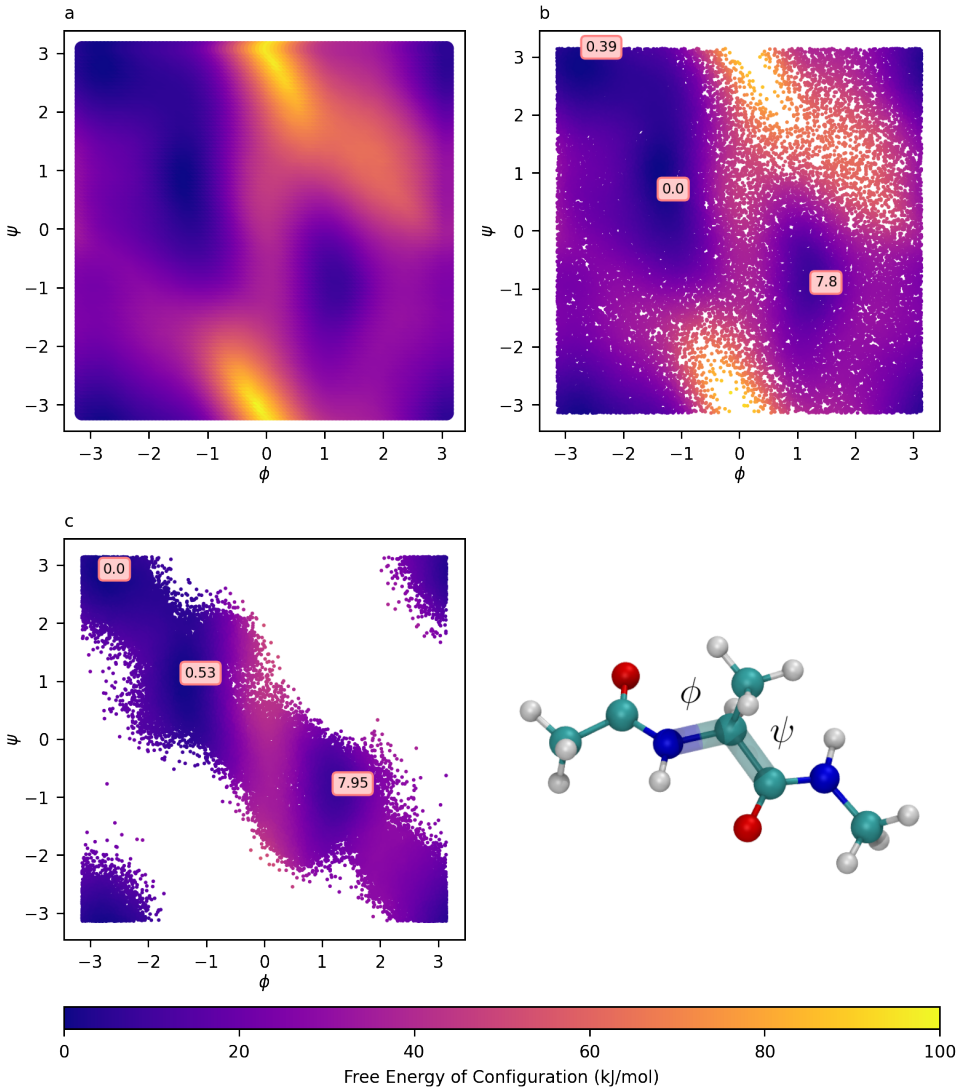


Figure 4.1: a: A conformational free energy surface for alanine dipeptide, obtained conventionally, through constructing a reweighted probability distribution on a histogram. b: The same free energy surface, constructed from a probability distribution derived from the local densities of individual configurations sampled from the simulation. The positions and free energies of local minima, as identified by DPA, are overlaid. c: The same free energy surface, also constructed from local densities, sampling a simulation using 2 $\times$1D WTmetaD biases in place of a conventional 2D bias. The positions and free energies of local minima, as identified by DPA, are overlaid. d: Alanine dipeptide, with the two relevant torsions ϕ and ψ highlighted. For a, b, c, the energies of the FESes are indicated by a colormap in units of kJ mol^{-1}.

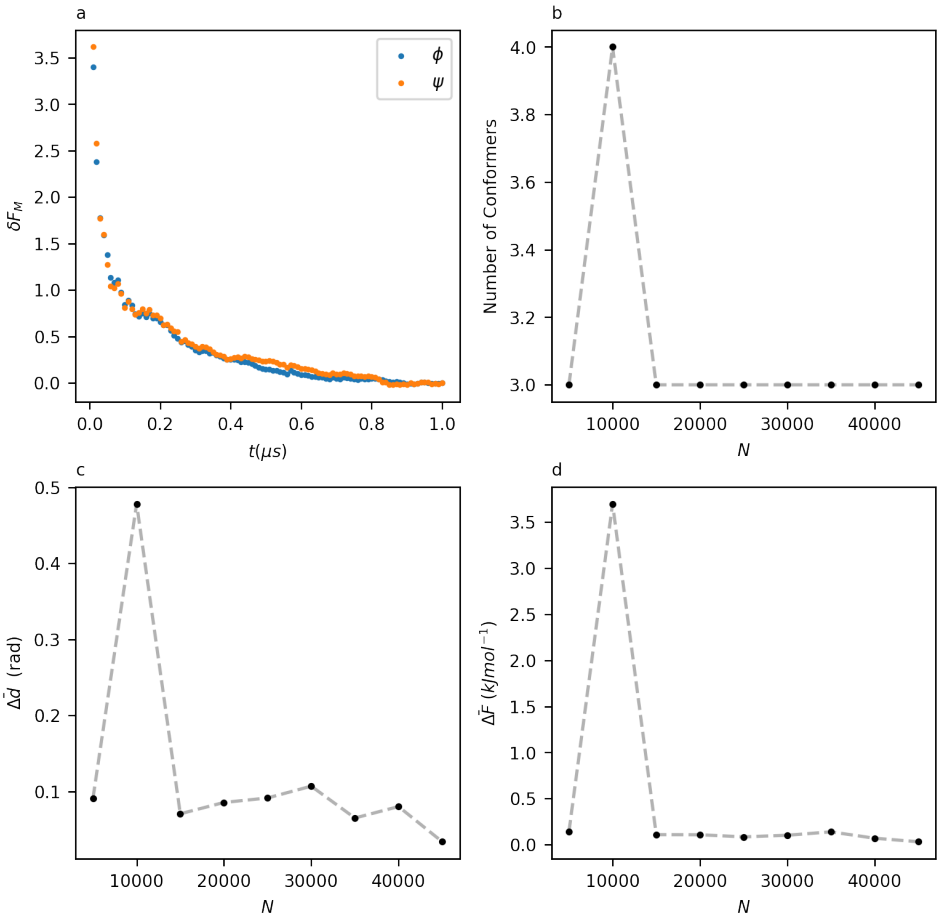


Figure 4.2: a: Evolution of the average free energy difference, $\delta F(t)$, on the marginal free energies of each torsion in alanine dipeptide. b: Number of clusters identified by clustering on datasets of size N. c: Evolution of the average positional deviation, $\Delta\bar{d}$, with N for alanine dipeptide. d: Evolution of the average free energy deviation, $\Delta\bar{F}$, with N for alanine dipeptide

and ψ are biased independently. Besides demonstrating the consistency of the free energies obtained by concurrent biasing, the comparison between Figure 4.1b and Figure 4.1c illustrates the trade-offs entailed using concurrent metadynamics. As detailed in the Methods section, concurrent metadynamics promotes the sampling of metastable states without guaranteeing an exhaustive sampling of the joint configurational probability density. Nevertheless, all relevant free energy minima are adequately sampled, and their positions and free energies agree with those obtained by standard, two-dimensional metadynamics (Fig. 4.1).

The results of the consistency analysis techniques outlined in the Methods section on the per-point FES outlined in Figure 4.1c are shown on Figure 4.2a-d. Figure

4.2a shows the evolution of $\delta F(t)$ for ϕ and ψ. The flattening of the curves shows that the sampling of the two torsions is indeed ergodic over the timescale of the simulation.

Figure 4.2c shows the mean conformer free energy difference $\bar{\Delta F}^N$ (defined in the Methods section) obtained from clustering datasets of increasing N and a reference dataset at N =50,000 configurations. Figure 4.2d shows a similar plot displaying the mean separation of the cluster centers, $\bar{\Delta d}$. Figure 4.2b shows the number of minima identified by DPA for each reduced-size dataset. The plot shows that all reduced datasets agreed that there were three conformers, with the exception of the 10,000-configuration dataset. In all other datasets, there is very good agreement on the position and free energies of the local minima, with energy differences well within 1 kJ mol^{-1} and mean separations hovering around 0.1 rad. It is surprising how little data is required to generate reasonable results in this 2-dimensional case.

4.2 Applications to higher dimensional free energy surfaces

Having demonstrated the workflow developed here on the two-dimensional case of alanine dipeptide, this section now explored higher dimensional cases, where visualization of the entire conformation space is not possible and conventional grid-based methods become unfeasible. Sulfadiazine, with a 4-dimensional conformational space, and Candidate XXXII [65, 66], with an 11-dimensional conformational space, serve as a test for the ability of the workflow to handle conformational complexity. Sketch-map [54] is used in these cases to project a 2D representation of the high dimensional per-point FES for human interpretation.

4.2.1 Sulfadiazine

Sulfadiazine is an antibiotic molecule with a 4-dimensional conformational space; its clinical relevance and intermediate complexity make it an ideal next step for the method outlined here. A 4-dimensional space is too high to allow a FES to be fully

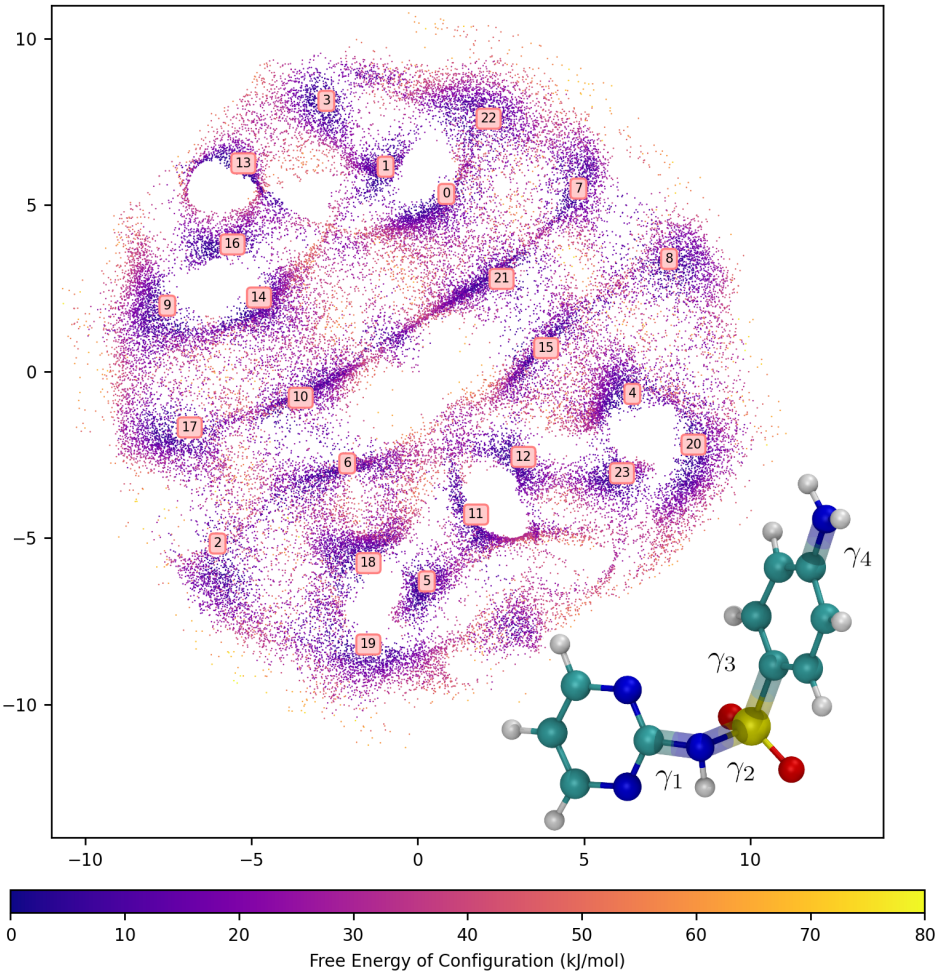


Figure 4.3: 2D Sketch-map projection of sulfadiazine's 4D conformation surface, with molecular structure of sulfadiazine inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

visualized while still being low enough that reasonable data density can be obtained (50,000 data points in a periodic 4D space results in an average density of roughly 32 configurations per rad^4). The inset in Figure 4.3 shows the 4 torsions considered in sulfadiazine. Using the same approach outlined above for alanine dipeptide, sulfadiazine's conformational FES was studied by analyzing the configurations sampled within a 1 μs single-molecule WTmetaD simulation. The resulting per-point FES cannot be fully visualized without dimensionality reduction, so the relative free energies and coordinates of each minimum are presented in Tab. 4.1. Figure 4.3 shows a 2D projection of the 4D per-point FES created using Sketch-map (cov-

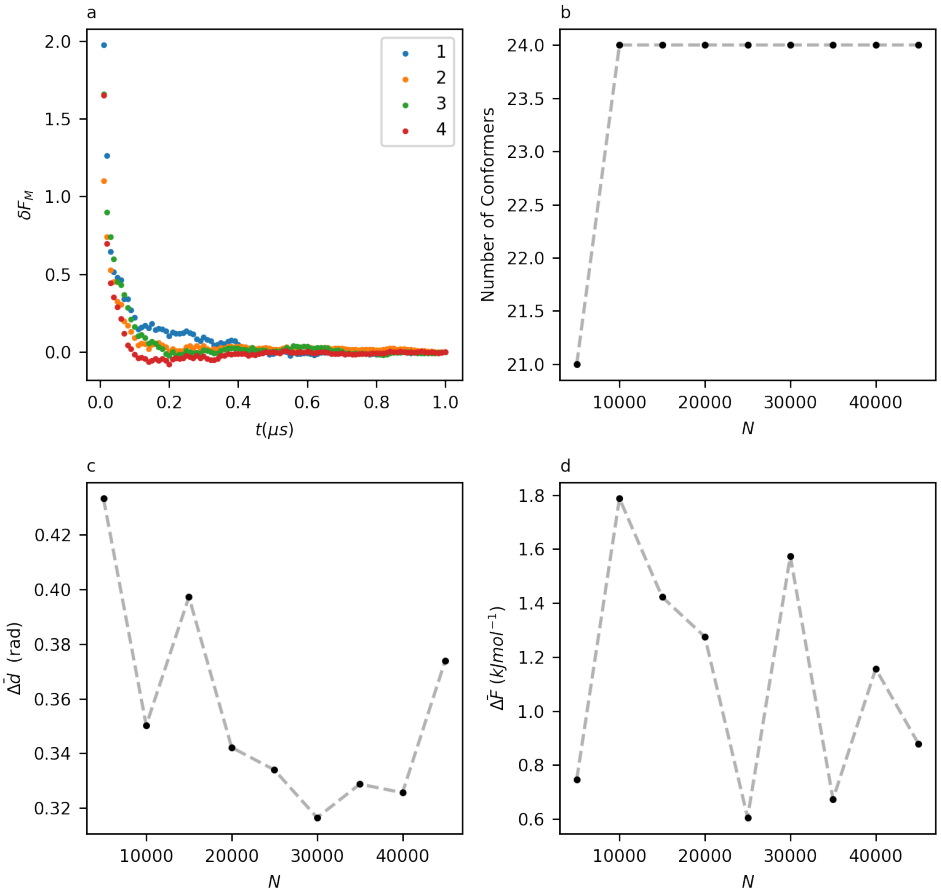


Figure 4.4: a: Evolution of the average free energy difference, $\delta F(t)$, on the marginal free energies of each torsion in sulfadiazine. b: Number of clusters identified by clustering on datasets of size N. c: Evolution of the average positional deviation, $\bar{\Delta d}$, with N for sulfadiazine. d: Evolution of the average free energy deviation, $\bar{\Delta F}$, with N for sulfadiazine

.

ered in detail in 2.4). This representation preserves the short-distance connectivity between data points, allowing the visualization of distinct free energy basins and the transition states between them, though the two axes of the new 2D projection are not physically meaningful themselves [54]. It should be emphasized that the estimation of densities and the determination of the number and coordinates of the free energy minima are determined in the full 4-dimensional conformation space and that the projection in Figure 4.3 serves only to assist in the visualization of the relationships between different conformers. It is possible to combine the 4-dimensional information presented in Table 4.1 with the 2-dimensional intuition provided by Figure 4.3.

For example, the FES in Figure 4.3 appears to be bisected by a diagonal channel, and indeed, by inspecting the torsion values of the conformer pairs (17, 2), (10, 6), (21, 15), and (7, 8), it can be determined that these conformers pairs are identical, and differ from each other in a symmetric rotation of π radians of γ_3. This example serves to show how these 2D projections may be manually interpreted and to demonstrate how symmetry elements in the molecule's conformation space can be preserved in the 2D projection.

The results of the consistency metrics for sulfadiazine are shown in Figure 4.4a-d. In comparing these results to those in Figure 4.2a-d, it is possible to evaluate the impact of doubling the dimensionality of the conformation space on the accuracy and data efficiency of the classification process. The plots of $\delta F(t)$ for the four torsions show that the four marginals in Figure 4.4a converge rapidly, providing evidence of ergodicity. Figure 4.4b shows that, with the exception of the 5000-point dataset, repeated analyses achieve a consistent number of 24 conformers. Figure 4.4c,d shows the evolution of $\bar{\Delta d}$ and $\bar{\Delta F}$ respectively, as N increases to a reference value of 50,000. Here, the differences between alanine dipeptide and sulfadiazine become apparent. The mean free energy deviation jumps from being nearly negligible to a range between 0.5 and 2 kJ mol^{-1}, and positional deviation increases from approximately 0.1 rad to between 0.3 and 0.4 rad. Sulfadiazine's energy deviation is still within 1 k_BT, and the positional deviations still correspond to very small changes in the molecular structure. However, the abrupt change following an increase in dimensionality highlights the importance of carrying out consistency checks when working with highly unintuitive results that are difficult to inspect visually. Due to the number of equivalent conformers related to one another by symmetric transformations in sulfadiazine, it is possible to compare the free energies of equivalent conformers as an assessment of the reproducibility of the free energy calculation. This is not recommended as a standard practice, as the presence of symmetrically related conformers is system dependent and not guaranteed. However, in this case, comparing the differences between equivalent conformers reveals deviations on the same order as the mean free energy deviations calculated in the

Table 4.1: Labels, free energies and CV-space coordinates of sulfadiazine's 24 conformers. The labeling convention is consistent with that of Figure 4.3

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4
0	0.72	-2.10	-1.51	-1.71	2.95
1	1.28	0.13	-1.52	-1.70	-2.98
2	0.75	-1.83	1.54	1.55	0.21
3	1.44	-0.13	-1.47	-1.62	-0.05
4	0.08	-1.94	1.62	-1.81	-2.93
5	1.05	0.09	1.58	1.64	2.86
6	0.00	1.75	1.42	1.92	-0.28
7	1.03	1.93	-1.31	-1.43	0.07
8	1.66	2.08	1.26	-1.49	0.23
9	0.67	-1.87	-1.60	1.30	-3.02
10	0.63	2.05	-1.66	1.68	0.15
11	1.56	-0.11	1.68	1.62	0.21
12	1.81	0.01	1.64	-1.55	0.17
13	0.21	-0.16	-1.65	1.62	-0.15
14	0.93	1.89	-1.53	1.85	3.13
15	1.02	-1.95	1.88	-1.76	0.31
16	1.20	0.06	-1.70	1.62	-3.08
17	1.40	-1.96	-1.77	1.40	0.15
18	1.58	1.88	1.63	1.74	-3.07
19	0.21	-2.04	1.83	1.55	2.93
20	0.80	2.00	1.54	-1.55	-2.81
21	0.47	-1.83	-1.59	-1.82	0.20
22	1.76	2.00	-1.77	-1.46	2.92
23	1.03	0.01	1.69	-1.58	2.90

smaller datasets (Figure 4.4d).

4.2.2 Target XXXII of the 7th CSP Blind Test

The final conformational FES explored in this chapter is that of Molecule XXXII, a target from the 7th CSP Blind Test [65, 66], which are a set of Crystal Structure Prediction (CSP) challenges issued by the Cambridge Crystallographic Data Centre (CCDC). As a highly flexible drug-like molecule with a conformation space defined by 11 torsions (shown inset in Figure 4.5), it is chosen here to test the limits of our method. To facilitate comparison with results collected for alanine dipep-

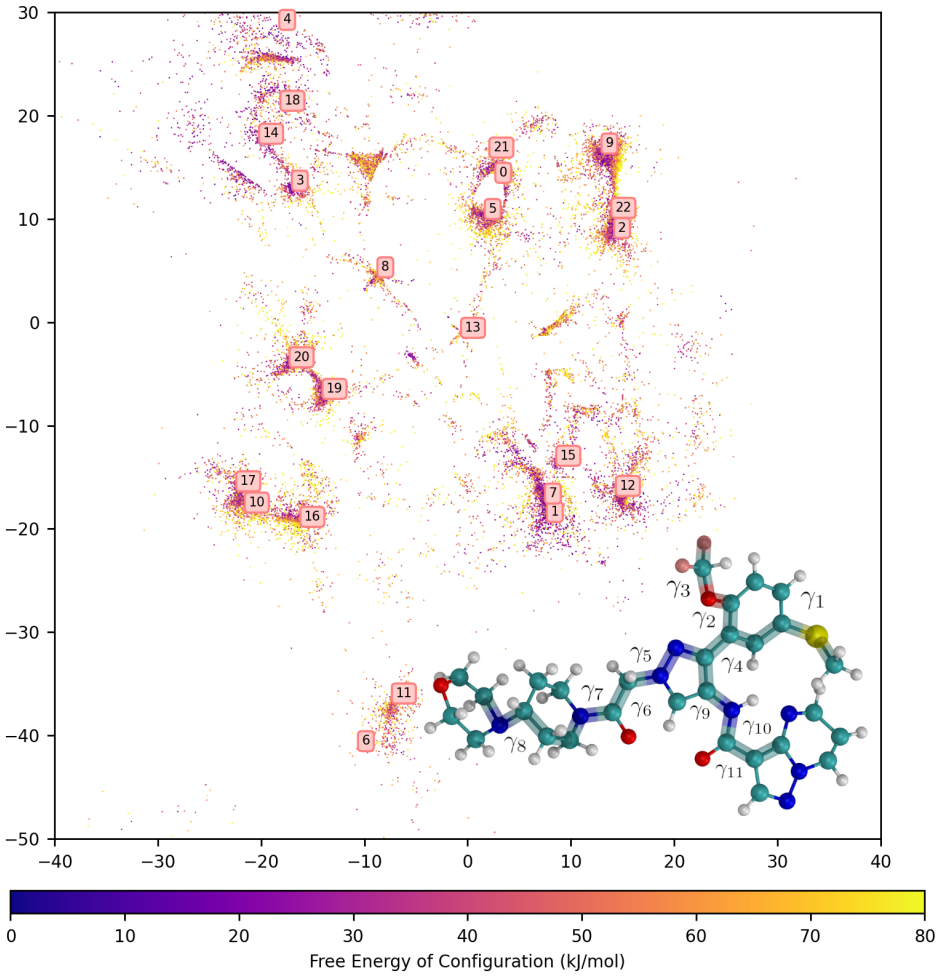


Figure 4.5: 2D Sketch-map projection of XXXII's 11D conformation surface, with molecular structure of XXXII inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

tide and sulfadiazine, the results presented here were generated using consistent simulation and analysis parameters, including the use of the same number of configurations. Using only 50,000 configurations in this high dimensional space results in an average data density of approximately 8×10^{-5} configurations per rad^{11}. Despite the extremely low data density, which is inherently linked to the complexity of the conformational space, we show that meaningful results are achievable. This is important as increasing the size of the dataset increases the cost of the analysis quadratically; increasing dataset size is thus much more expensive than increasing

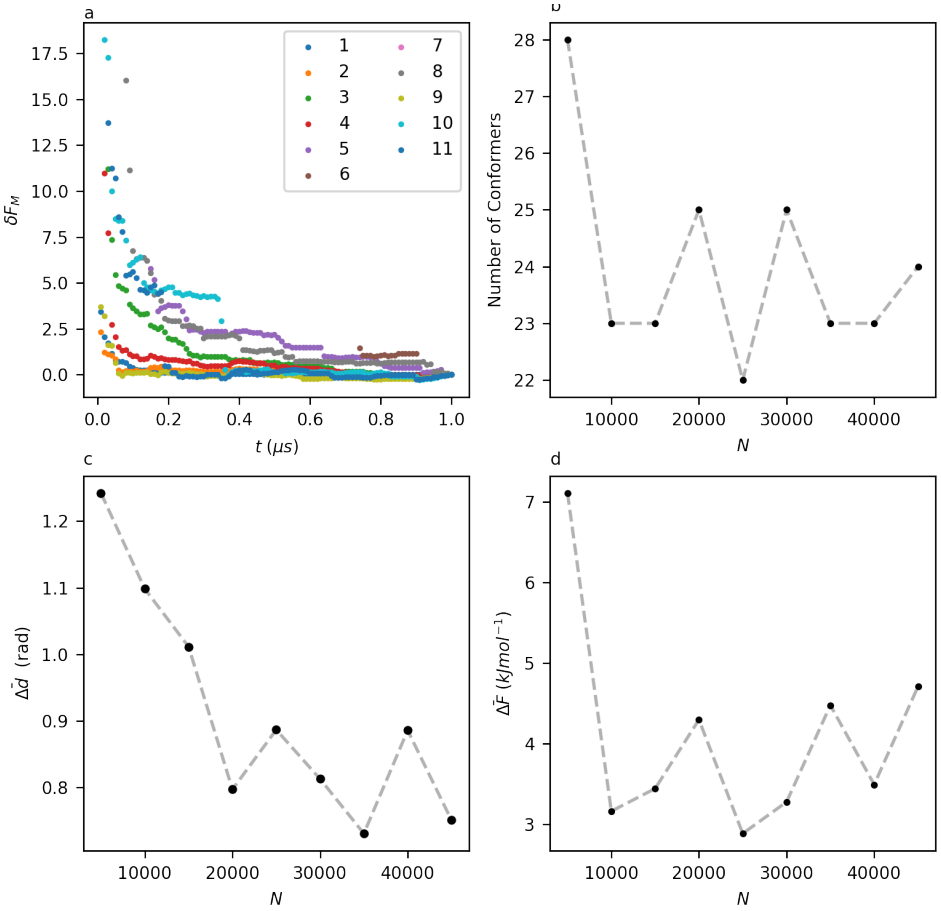


Figure 4.6: a: Evolution of the average free energy difference, $\delta F(t)$, on the marginal free energies of each torsion in XXXII. b: Number of clusters identified by clustering on datasets of size N. c: Evolution of the average positional deviation, $\Delta\bar{d}$, with N for XXXII. d: Evolution of the average free energy deviation, $\Delta\bar{F}$, with N for XXXII.

the dimensionality of the considered space. Figure 4.5 shows the projected 11-dimensional per-point FES, with cluster centers corresponding to 11-dimensional coordinates presented in Tab. 4.2. When comparing this FES to that of sulfadiazine in Figure 4.3, the features of XXXII can be seen reflected in its own FES. The relative lack of symmetrical torsions results in a less symmetrical FES, and the higher-dimensional FES is much sparser, illustrating that the computational savings arise from a more efficient, rather than more exhaustive, sampling of conformational space.

The consistency metrics in Figure 4.6a-d are, however, less reliable than those obtained for sulfadiazine. Figure 4.6a shows well-converged marginal free energies,

Table 4.2: Labels, free energies, and CV-space coordinates of XXXII's 23 conformers. The labeling convention is consistent with that of Figure 4.5

C	FE [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}	γ_{11}
0	6.09	-1.27	0.89	1.12	0.34	-2.31	1.13	0.42	1.29	2.96	3.13	0.15
1	0.00	1.84	-1.52	3.06	-2.75	-2.24	1.29	0.20	1.12	2.95	-3.13	0.04
2	8.45	1.98	1.07	1.03	0.37	-2.24	1.12	0.85	-1.19	-3.05	3.09	0.12
3	6.20	1.24	1.05	1.25	-2.93	2.43	-1.11	2.65	-1.12	3.11	3.11	-0.08
4	3.40	-1.05	1.72	1.21	2.95	2.23	-1.21	2.28	1.10	-3.09	2.93	-0.10
5	8.69	1.71	0.85	1.17	0.58	-2.06	1.21	-0.22	1.19	-3.06	-2.94	0.20
6	15.67	-0.80	-1.39	3.04	-0.41	1.06	0.93	0.43	1.42	3.07	3.13	-0.34
7	2.17	1.14	-1.60	-3.06	-2.85	-2.00	1.39	-0.18	1.00	3.10	-2.94	0.09
8	16.12	-0.62	1.61	1.07	2.67	0.91	1.01	0.26	0.94	-3.04	3.02	-0.30
9	7.62	-1.75	0.93	1.03	0.56	-2.35	1.25	0.69	-1.06	2.97	-3.08	0.18
10	12.45	-0.61	-1.43	-3.09	-0.30	2.28	-1.13	2.41	1.11	2.99	3.09	-0.18
11	21.03	1.00	-1.16	2.81	-0.36	1.04	1.00	0.51	1.14	2.92	-3.02	-0.31
12	1.43	1.39	-1.58	-3.10	-2.64	-2.34	1.17	0.64	-0.89	-3.03	3.12	0.07
13	6.92	1.63	-1.60	-2.99	-2.70	-1.02	-0.97	2.73	-0.93	2.83	-2.92	0.04
14	5.96	-0.86	1.51	1.17	2.86	1.96	-0.94	2.64	-1.04	3.12	-3.05	-0.09
15	5.72	0.71	-1.49	3.04	-2.51	-2.39	1.24	0.53	3.03	3.03	-2.98	-0.09
16	10.98	2.07	-0.91	-3.01	-0.57	2.26	-1.36	2.17	1.23	3.13	3.09	-0.16
17	8.91	-1.40	-0.90	-3.10	-0.42	2.12	-1.16	2.32	1.26	3.03	-2.94	-0.12
18	6.88	-1.62	1.61	1.18	2.66	2.10	-1.19	2.85	-0.95	-3.12	2.92	0.03
19	9.53	1.55	-1.25	3.02	-0.56	2.09	-1.20	2.69	-0.91	2.98	3.11	-0.09
20	6.89	-2.12	-1.31	3.07	-0.62	2.07	-1.17	2.86	-1.06	3.06	3.11	-0.21
21	7.35	-2.30	1.20	1.18	0.55	-2.11	0.89	0.41	1.37	2.86	-2.81	0.09
22	6.13	1.23	0.85	1.16	0.27	-2.15	1.19	0.59	-0.89	3.14	-2.93	0.16

but Figure 4.6b shows that the number of conformers identified is less consistent than in lower-dimensional cases. The number of metastable states identified as distinct conformers hovers between 22 and 25 for datasets sized 10000 and upwards. Along with a fluctuating number of conformers, larger deviations in free energies and positions are now observed, with $\bar{\Delta F}$ between cluster sets now varying by up to 5 kJmol^{-1}, and $\bar{\Delta d}$ drifting by as much as one full radian, even at large dataset sizes. Despite this drop in the quality of the results, it is still remarkable that a reasonably intuitive understanding of such a high-dimensional conformational FES can be derived from a limited amount of data in a computationally accessible way, even if its value in this instance is chiefly qualitative. To further explore the consistency of the FES in Figure 4.5, Figures A.2-A.10 in Appendix A contain the FES projection for

each of the smaller datasets used in the consistency analysis, allowing the evolution of this per-point FES to be observed. Inspection of this evolution in the FES seems to reveal that the majority of the fluctuations in $\Delta\bar{F}$ and $\Delta\bar{d}$ observed arise in higher energy conformers, with the low energy regions converging at lower N values. Although this is not rigourously proved here, it is a reasonable expectation, as lower energy regions have a high data density, resulting in free energy estimates based on a greater amount of data.

Chapter 5

Exploring the Impact of Solvent on the Conformational Landscapes of Pharmaceutical Molecules

In order to further explore the utility of the approach outlined in this project, four new compounds were selected for study that were not considered at any point during method development and present environment dependent conformational behavior. N-Phenylbenzohydroxamic acid (PBH), bicalutamide, taltirelin, and m-nisoldipine were chosen because each molecule possesses a degree of molecular flexibility and exhibits conformational polymorphism when crystallized under different solvent conditions. Of these molecules, all except PBH are pharmaceutical molecules, and all except m-nisoldipine have been observed to have their conformational distribution in the solution phase be affected by solvent choice. The main purpose of this chapter is to demonstrate a workflow for the study of solvated simulations which utilizes the high-dimensional conformational free energy landscapes and associated conformer sets generated by the approach demonstrated in Chapters 3 and 4.

Each molecule was simulated in vacuum, as well as in a pair of solution environments chosen to produce distinct conformational free energy landscapes based on experimental results. For each of these simulations, per-point free energy landscapes were generated using the method outlined in Chapter 3. The calculation of configurational free energies and conformer classification is carried out indepen-

dently for each simulation; however, to enable comparison across different environments, the Sketch-map projections generated for the solvated cases use the same a, b, and σ parameters and landmark points as determined from the post-processing of a simulation in vacuum. This ensures that the resulting two-dimensional projection is equivalent across the different environments.

As discussed in Chapter 3, the geometry of a conformer is represented by the geometry of the configuration identified by DPA as a cluster center in conformation space. This configuration is, by construction, the most representative of the geometries found within its conformational basin. Molecular geometries can be compared by evaluating their minimum root-mean square deviation (RMSD) overlap. Two conformations are overlaid, and their relative orientations are changed such that the quantity

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \tag{5.1}$$

is minimized, where δ_i is the distance between the two equivalent atoms i in each of the overlapped structures. Here, minimum RMSD is calculated using the method developed by Coutsias et al. [67] as implemented by the software package RDkit [68].

5.1 N-Phenylbenzohydroxamic acid

N-Phenylbenzohydroxamic acid (PBH) has a conformation space defined by three torsions, as shown inset in Figure 5.1a. Of key interest is the central torsion γ_2, which determines whether the molecule exists in a *cis* or *trans* conformation. Experimental studies carried out by Yamasaki et al. [69] indicate that the conformational distribution of PBH depends on its solvent environment. Specifically, through IR and H NMR spectroscopy, they identified the conformational distribution of PBH in dichloromethane as overwhelmingly favoring the *cis* conformation, while the *trans* conformer dominates in a 77:23 ratio in acetone. Furthermore, they also determined that two different conformational polymorphs of PBH are isolated when recrystallizing out of dichloromethane and acetone, with the conformations aligning with

the dominant conformation in solution. The authors report that the cause of these differences in solvated conformational distribution is unknown, making PBH a fascinating case study for the approach undertaken in this chapter.

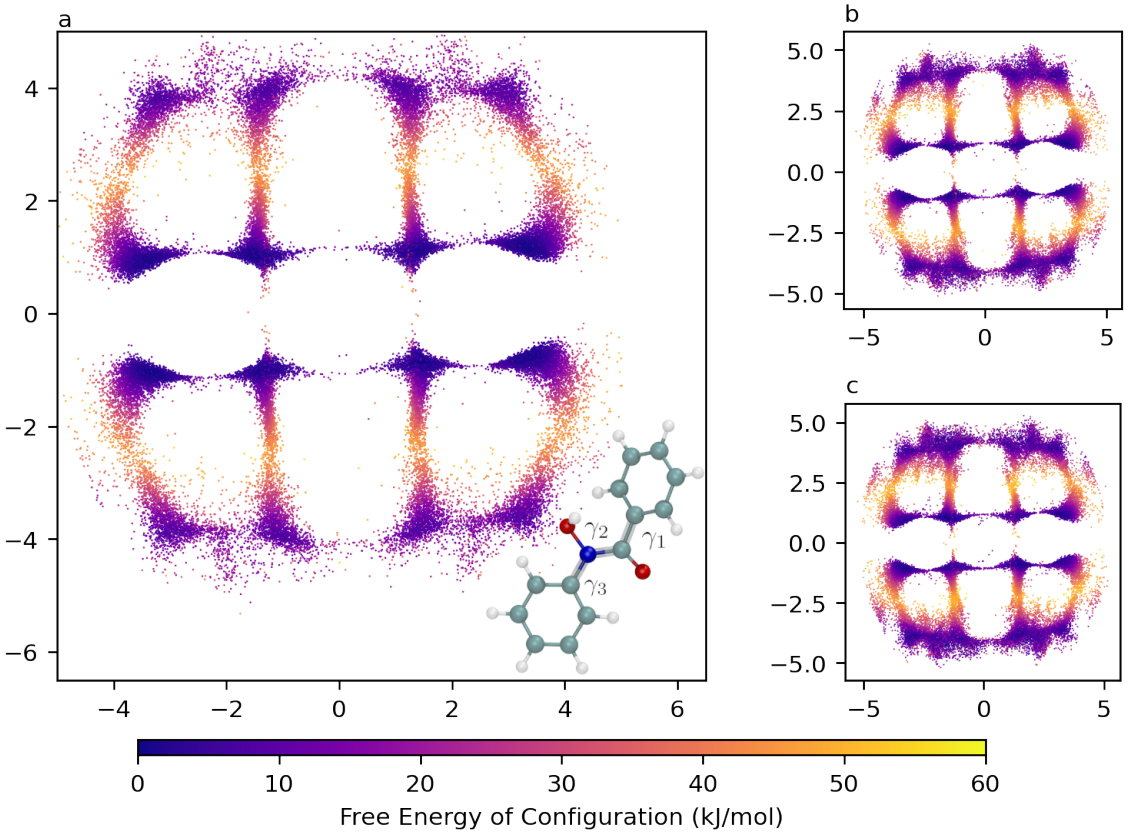


Figure 5.1: 2D Sketch-Map projection of PBH's 3D conformational free energy landscape in vacuum (a), dichloromethane(b), and acetone(c), with molecular structure of PBH inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

These experimental observations, paired with the relative simplicity of the conformation space make PBH an ideal starting point for the study of the impact of solvents on the conformational distributions of small organic molecules. Firstly, the approach detailed in Chapter 3 was applied to PBH, with a microsecond long simulation in vacuum being carried out, with concurrent metadynamic biases promoting the sampling of the three dihedrals. A total of 50,000 configurations were sampled from the second half of this MD trajectory and analyzed according to the method described in Chapter 4. The resulting three-dimensional per-point free energy landscape was projected into two dimensions using Sketch-map, with the pro-

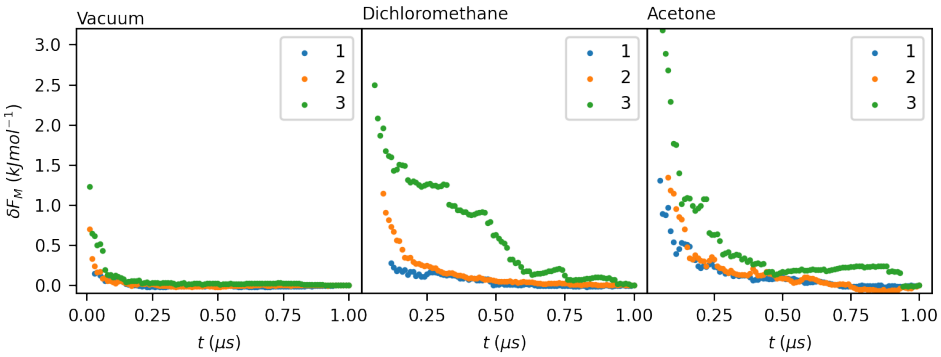


Figure 5.2: Evolution of the average free energy difference, $\delta F(t)$, on the marginal free energies of each torsion in PBH, in vacuum, dichloromethane, and acetone.

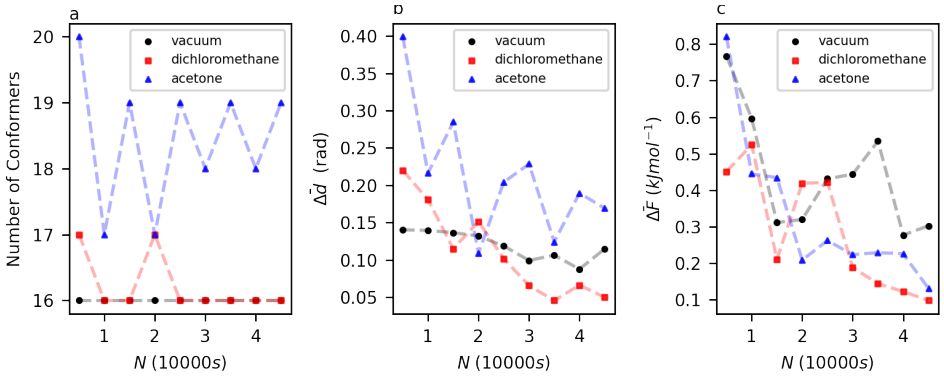


Figure 5.3: a: Number of clusters identified by clustering on datasets of size N for PBH, in vacuum, dichloromethane, and acetone. b: Evolution of the average positional deviation, $\Delta\bar{d}$, with N for PBH, in vacuum, dichloromethane, and acetone. c: evolution of the average free energy deviation, $\Delta\bar{F}$, with N for PBH, in vacuum, dichloromethane, and acetone.

jection shown in Figure 5.1a. Concurrent WTMetaD simulations of a single PBH molecule were then carried out in dichloromethane and acetone, using similar parameters as those used in vacuum. Per-point free energy landscapes were generated for the solvated simulations using the same procedure as, but completely independently from, the in-vacuum simulations. The resulting two-dimensional projections of these solvated free energy landscapes are shown in Figures 5.1b and 5.1c, for dichloromethane and acetone respectively. For each of the projections, an enlarged image featuring the locations of the free energy minima marked with a numerical label is available in Figures B.1, B.2, B.3 in Appendix B. These labels correspond to the conformer indices in the left-hand columns of Tables B.1, B.2, and B.3, also

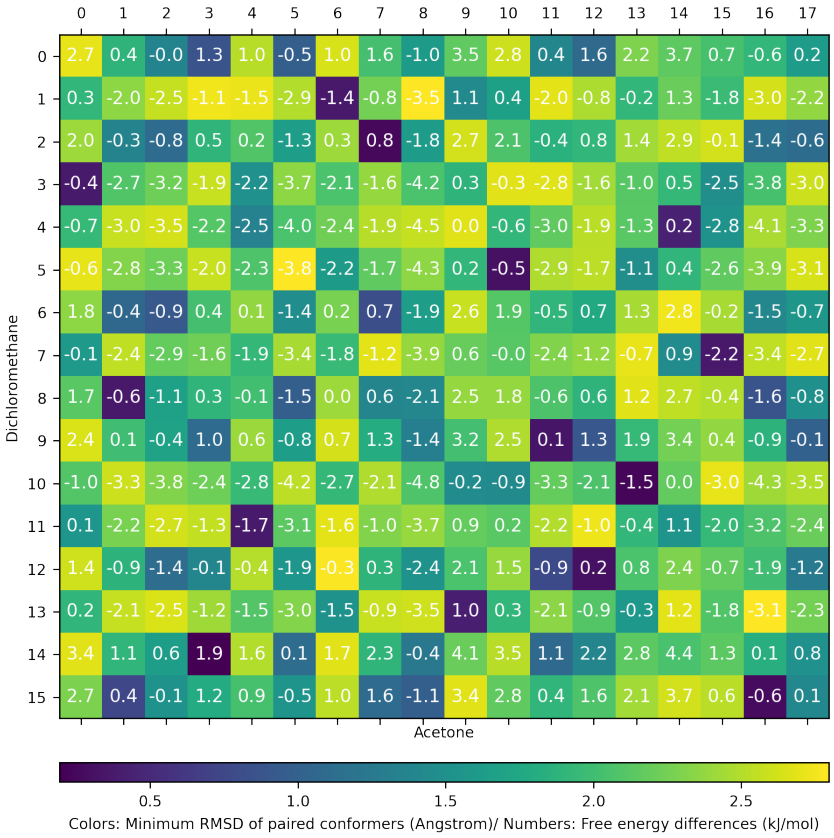


Figure 5.4: A dual matrix presenting a pairwise comparison of conformers of PBH observed in dichloromethane (rows) and acetone(columns). The color gradient indicates the minimum RMSD between atoms between the two conformers, while the number within each element indicates the stability of the conformer in dichloromethane relative to the conformer in acetone.

in Appendix B, for PBH in vacuum, dichloromethane, and acetone, respectively. These tables present the free energies of each conformer, as well as their coordinates in the three-dimensional conformation space.

To monitor the quality of the high-dimensional results collected, the same consistency procedures used in the previous chapters were applied here. Figure 5.2 shows the evolution of $\delta F_m(t)$ for each of the three torsions in PBH in each of the three environments simulated. In all cases, it is clear that marginal free energies have been sampled to convergence. The results of the high-dimensional consistency checks are shown in Figure 5.3. Figure 5.3a shows the number of conformers identified in each environment as the size of the dataset grows. While this number is consistently 16 for PBH in vacuum, there are some anomalous results deviating

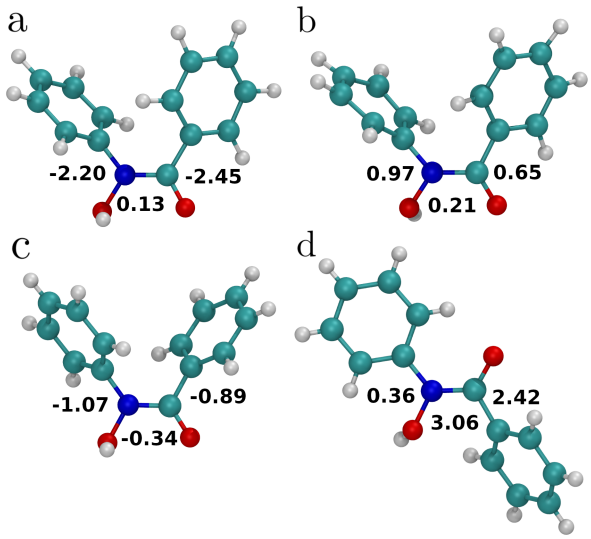
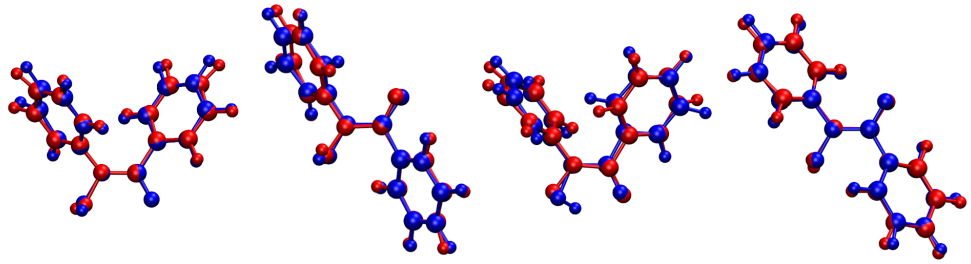


Figure 5.5: a: The lowest energy conformation of PBH in dichloromethane, index 13 in Table B.2 and Figure 5.1b. b: The lowest energy conformation of PBH in acetone, index 14 in Table B.3 and Figure 5.1c. c: The experimentally observed conformation of PBH when crystallized out of dichloromethane. d: The experimentally observed conformation of PBH when crystallized out of acetone. For all conformations, values of the torsions γ_1, γ_2, and γ_3 are indicated in radians.

from this in dichloromethane. Complete consistency in the number of clusters is never observed in acetone, though there are always more than 16 conformers identified.

This lack of consistency and increase in the number of conformers identified reflects the distortion of the *trans* region of the free energy landscape observed in the two-dimensional FES projections, as is further discussed below. Figure 5.3b shows the evolution of $\Delta\bar{d}$ in each of the three environments as the size of the clustered dataset increases. $\Delta\bar{d}$ does continue to decrease as dataset size increases, but even at small dataset sizes, $\Delta\bar{d}$ is below 0.5 radians, indicating that equivalent conformers are found at essentially the same location regardless of dataset size. Figure 5.3c shows the evolution of $\Delta\bar{F}$ in all three environments as the size of the clustered dataset grows. Again, there is a decreasing trend present in this quantity, but even at small dataset sizes, the values of $\Delta\bar{F}$ indicate a consistent free energy ranking of the conformers.

Comparing tabular results with the positions of the conformers in the projec-



DCM11-ACE4		DCM15-ACE16		DCM4-ACE14		DCM14-ACE3	
ΔF(kJmol^{-1}):	-1.68	ΔF(kJmol^{-1}):	-0.64	ΔF(kJmol^{-1}):	0.25	ΔF(kJmol^{-1}):	1.93
F_{DCM}(kJmol^{-1}):	1.08	F_{DCM}(kJmol^{-1}):	3.66	F_{DCM}(kJmol^{-1}):	0.25	F_{DCM}(kJmol^{-1}):	4.35
F_{ACE}(kJmol^{-1}):	2.76	F_{ACE}(kJmol^{-1}):	4.3	F_{ACE}(kJmol^{-1}):	0.0	F_{ACE}(kJmol^{-1}):	2.42
RMSD (angstrom):	0.38	RMSD (angstrom):	0.27	RMSD (angstrom):	0.43	RMSD (angstrom):	0.19

Figure 5.6: 4 common conformers of PBH in dichloromethane and acetone. For this system, conformers are deemed common to both solvents if their overlaid structures present a minimum RMSD deviation of less than 0.5Å. Conformers in dichloromethane are indicated with the label DCM, and have their molecular structures shown in blue. Conformers in acetone are indicated with the label ACE and their molecular structures are shown in red. The free energy in both solvents, as well as the difference in free energy is indicated for each conformer, and the conformers are ordered from those most stabilized in dichloromethane to those most stabilized in acetone.

tions, it can be deduced that the 8 conformers which form two central rows correspond to the *cis* conformers, with γ_2 values close to 0, while the outer conformers correspond to conformers in the *trans* configuration, with γ_2 values close to π. 8 *trans* conformers are observed in vacuum and in dichloromethane, and 10 are observed in acetone. This increase in the number of *trans* conformers in acetone is accompanied by what appears to be the distortion of the *trans* regions in the PBH-acetone Sketch-map. These distortions are visible at approximately (-2,4), (-2,-4), (2,-4), and (2,4) in the projected coordinates of the projections shown in Figure 5.1. These regions, which each consist of two clear free energy basins in vacuum, begin to merge in dichloromethane and the number of basins becomes even less clear in acetone. This distortion may explain why a consistent number of conformers is not observed in Figure 5.3a. However, in all three environments, *cis* conformers appear to be more thermodynamically stable, despite experimental results suggesting that this should not be the case in acetone. Due to the high quality consistency metrics

presented in Figure 5.2 and 5.3, as well as the low dimensionality of the conformation space, it is likely that this not a failure of the analysis, but rather a failure of the simulation to replicate experimental observations. Despite this, the relative simplicity of PBH's conformational space and the readily interpretable Sketch-map projection make PBH a useful case for the development and demonstration of the techniques used here to compare conformer-sets derived from simulations of the same molecule in different environments. Ultimately, the main purpose of this chapter is to demonstrate a workflow for the study of the impact of solvents on conformational free energy landscapes, and PBH serves as a useful demonstration, due to its simplicity.

Figure 5.5a and b show the two lowest-energy conformers of PBH in dichloromethane and acetone respectively. Below them, in Figure 5.5c,d, are the experimentally observed crystal structures of PBH when crystallized from dichloromethane and acetone respectively. The crystal structures in both cases are stabilized by hydrogen bonds between the carbonyl oxygen and the hydroxyl group, forming a dimer when *cis* and a chain when *trans*. Despite the experimental results reported by Yamasaki et al. [69] but as expected considering the projected landscapes, the *cis* conformation is the most stable conformer in both dichloromethane and acetone.

Figure 5.4 compares the conformer-sets of PBH in dichloromethane and acetone in a pair-wise matrix. Each element in the matrix combines two pieces of information. The color gradient indicates the RMSD between the conformer pairs, while the number indicates the stability of the conformer in dichloromethane relative to the conformer in acetone. This matrix provides an interpretable comparison of all combinations of conformers, while also rendering the relevant conformer pairs immediately apparent. The key conformer pairs to consider from this matrix are those separated by a small RMSD. These conformers are structurally very similar, indicating that their geometries arise in both solvents. Therefore, any difference in their thermodynamic stability must therefore arise due to solvent effects.

The symmetry of the rotations of γ_1 and γ_3 result in the conformational FES

consisting of 4 equivalent regions. Figure 5.6 considers the common conformers (here defined as conformers arising in both dichloromethane and acetone with a minimum RMSD separation of less than 0.5 Å) of one such region. The common conformers are arranged by the difference in free energy of the conformers in dichloromethane compared to the conformer in acetone. The free energies reported for each conformer in a specific solvent are relative to the most-stable configuration in that solvent, which has a free energy set to zero. Thus, ΔF as reported in Figure 5.6 compares the stabilities of the conformers in each solvent relative to the most stable conformer in that solvent, for both solvents. In this way, it is possible to compare the stabilizing effects of the solvent on geometries common to both environments.

Within Figure 5.6, conformers in dichloromethane are labelled with DCM, while those in acetone are labelled with ACE. DCM15-ACE16 and DCM4-ACE14 are *trans* and cis, respectively, and do not demonstrate a significant ΔF. However, DCM11-ACE4 and DCM14-ACE3, *cis* and *trans*, respectively, do demonstrate slight stabilization effects in the direction suggested by experiment. The *cis* conformer is stabilized slightly in dichloromethane while the *trans* conformer is stabilized slightly in acetone. While this does not change the fact that in both solvents, the most stable conformers overall are both *cis* (as shown in Figure 5.5), it is interesting to note that some *trans* conformers are stabilized by acetone, as suggested by experiment.

5.2 Bicalutamide

Bicalutamide is an anti-androgen compound used for the treatment of prostate cancer. It is highly flexible, with a conformational space described by the 7 dihedral angles shown inset in Figure 5.7a. This flexibility results in two conformational polymorphs being observed in the solid state, form I, shown in Figure 5.11c and form II, shown in Figure 5.11d. Form I demonstrates an open conformation, while form II adopts a more compact, closed conformation. Form I is the more thermodynamically stable form, and is the form which typically arises upon recrystallization

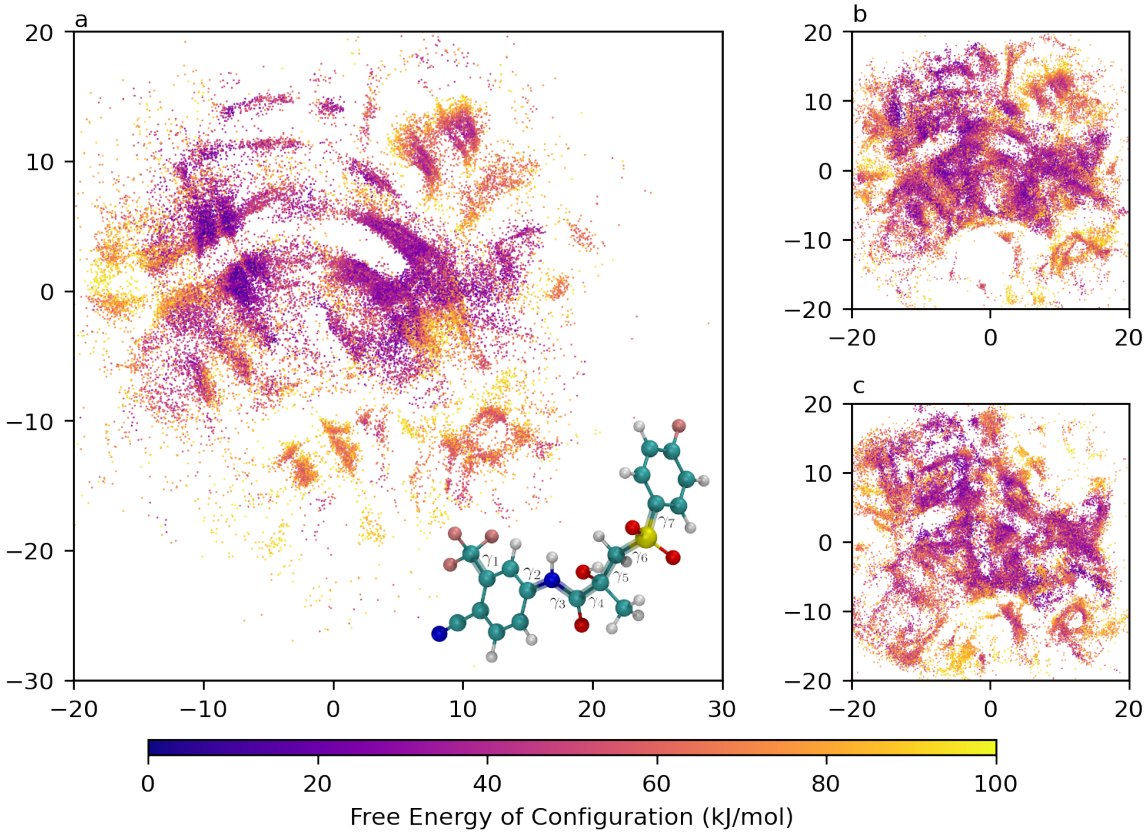


Figure 5.7: 2D Sketch-Map projection of bicalutamide's 7D conformational free energy landscape in vacuum (a), chloroform (b), and DMSO (c), with molecular structure of bicalutamide inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

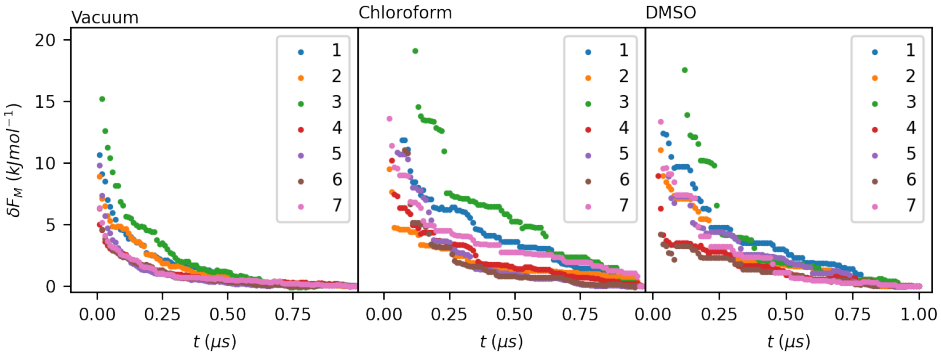


Figure 5.8: Evolution of the average free energy difference, $\delta F(t)$, on the marginal free energies of each torsion in bicalutamide, in vacuum, chloroform, and DMSO.

from most solvents [70]. It is stabilized by two intramolecular hydrogen bonds, N-H···OH and O-H···OS [70]. Form II can be obtained from a melt of form I [71]. It is also stabilized by N-H···OH, but a new hydrogen bond forms between N-H···OS,

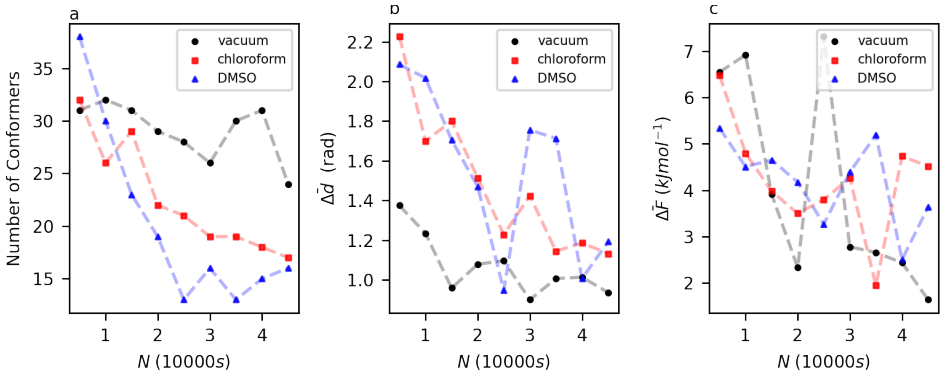


Figure 5.9: a: Number of clusters identified by clustering on datasets of size N for bicalutamide, in vacuum, chloroform, and DMSO. b: Evolution of the average positional deviation, $\Delta\bar{d}$, with N for bicalutamide, in vacuum, chloroform, and DMSO. c: evolution of the average free energy deviation, $\Delta\bar{F}$, with N for bicalutamide, in vacuum, chloroform, and DMSO.

displacing the O-H···OS bond [70]. Despite form I's tendency to recrystallize out of most solvents, Sobornova et al. discovered that solvent choice had a significant impact on bicalutamide's conformational distribution in solution [70]. Using Nuclear Overhauser Effect (NOE) spectroscopy, they demonstrated that polar solvents promote an open conformation, while non-polar solvents promote a closed conformation. Here, the conformational free energy landscapes of bicalutamide, simulated in vacuum, chloroform, and DMSO environments are explored using the gridless method developed in this project. Simulations were carried out according to the parameters defined in Chapter 3, and as usual, datasets of 50,000 configurations were used to construct the per point FESes shown here.

Figure 5.8 shows the evolution of $\delta F_m(t)$ for each of the 7 torsions of bicalutamide in vacuum, dichloromethane, and DMSO. This figure demonstrates the convergence of each of these one-dimensional marginal free energies. The results of the higher dimensional consistency checks are shown in Figure 5.9. Figure 5.9a shows that a consistent number of conformers was not arrived at in any of the environments simulated. This inconsistency in conformer number is characteristic of free energy landscapes computed in high dimensions, reflecting the trend seen with the consistency of Target XXXII shown in the previous chapter. Despite the number of conformers continuing to fluctuate as the largest dataset size is reached, the

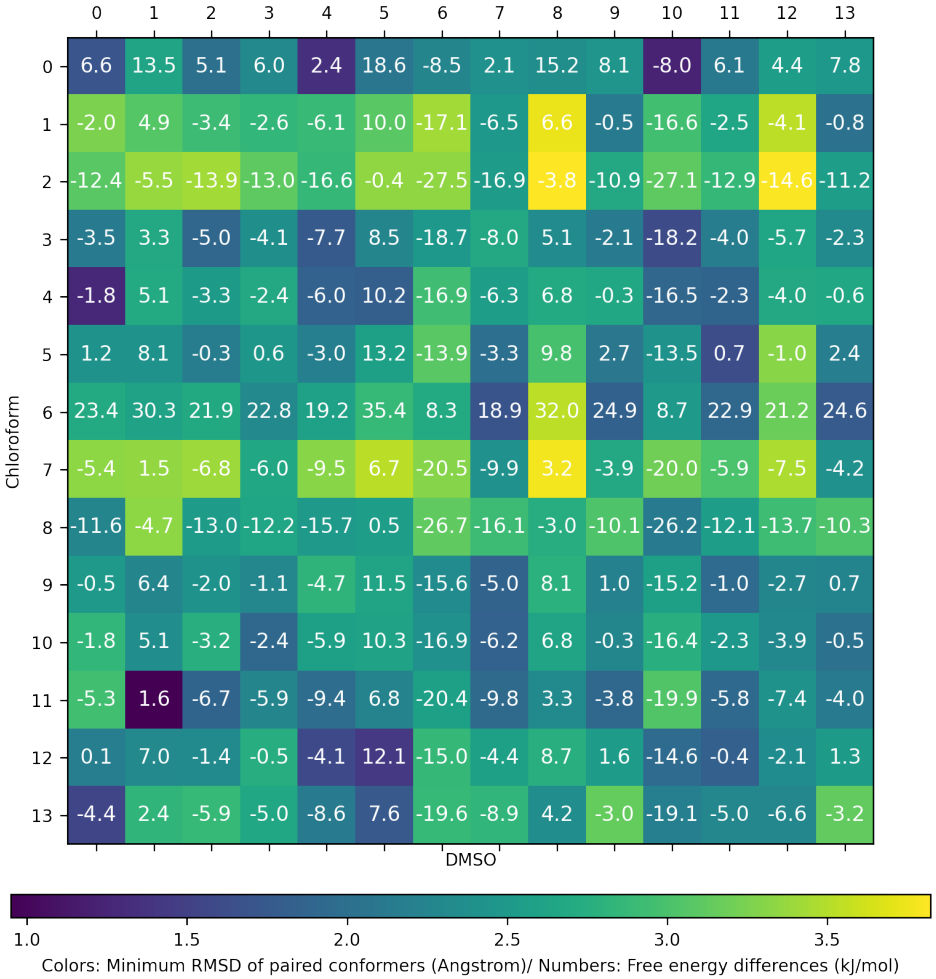


Figure 5.10: A dual matrix presenting a pairwise comparison of conformers of bicalutamide observed in chloroform (rows) and DMSO (columns). The color gradient indicates the minimum RMSD between atoms between the two conformers, while the number within each element indicates the stability of the conformer in chloroform relative to the conformer in DMSO.

average positions of equivalent conformers are fairly similar, as demonstrated by Figure 5.9b. For the landscapes in vacuum and chloroform, equivalent conformers are found on average less than 1.2 Euclidean radians from each other in a 7 dimensional space. This number is slightly less consistent in DMSO, with some $\bar{\Delta d}$ values being as high as 1.8 Euclidean radians, even towards the final dataset size. Finally, considering the difference in free energies between equivalent conformers, demonstrated in Figure 5.9c, it can be seen that in all three environments, conformers deemed equivalent are within, on average, 5 kJmol^{-1} of each other. This not as good an agreement as observed in the lower dimensional cases, however, it will

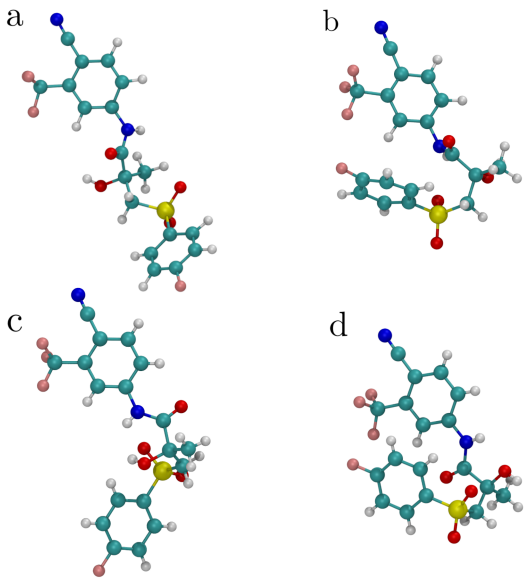


Figure 5.11: a: The lowest energy conformation of bicalutamide in DMSO, index 5 in Table B.5 and Figure 5.7b. b: The lowest energy conformation of bicalutamide in chloroform, index 2 in Table B.6 and Figure 5.7c. c: The experimentally observed conformation of bicalutamide form I. d: The experimentally observed conformation of bicalutamide from II.

be sufficient when comparing the free energies of conformers that differ by more than this amount. The inconsistency in results for this system is reflective of the dimensionality of the conformation space, and the difference in these metrics between PBH and bicalutamide are comparable in scale to the difference between the consistency metrics observed during the in-vacuo studies of alanine dipeptide and target XXXII from the previous chapter.

The per-point free energy landscapes generated through the gridless analysis are projected into two dimensions using Sketch-map. As in PBH, the parameters and landmarks identified in vacuum were used to create all three projections. These projections in vacuum, chloroform, and DMSO are shown in Figures 5.7a, 5.7b, 5.7c, respectively. For each of the projections, an enlarged image featuring the locations of the free energy minima marked with a numerical label is available in Figures B.4, B.5, B.6 in Appendix B. These labels correspond to the conformer indices in the left-hand columns of Tables B.4, B.5, and B.6, also in Appendix B, for bicalutamide in vacuum, chloroform, and DMSO, respectively. Some of the structure

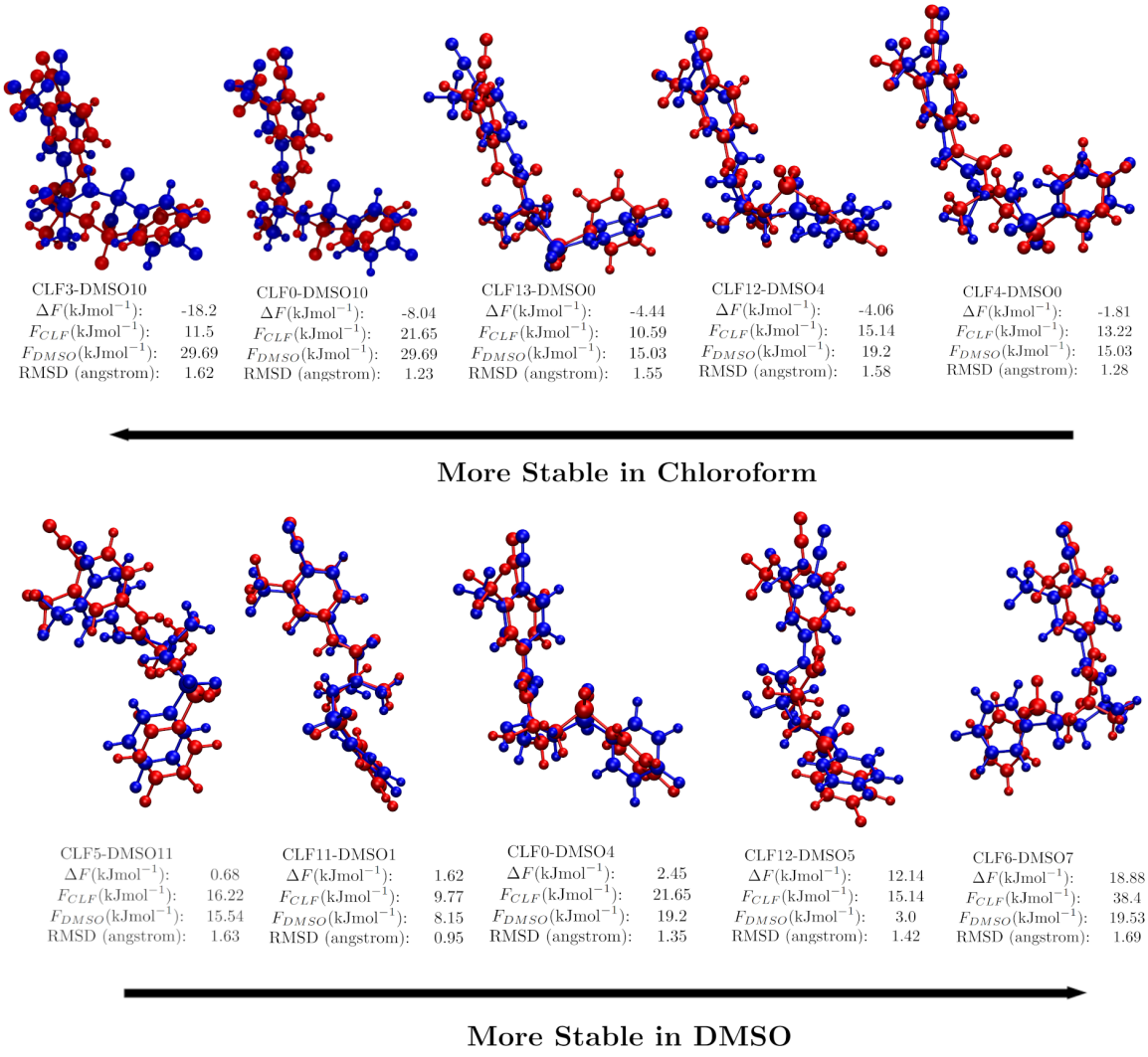


Figure 5.12: 10 common conformers of bicalutamide in chloroform and DMSO. For this system, conformers are deemed common to be both solvents if their overlaid structures present a minimum RMSD deviation of less than 1.7Å . Conformers in chloroform are indicated with the label CLF, and have their molecular structures shown in blue. Conformers in DMSO are indicated with the label DMSO and their molecular structures are shown in red. The free energy in both solvents, as well as the difference in free energy is indicated for each conformer, and the conformers are ordered from those most stabilized in chloroform to those most stabilized in DMSO.

present in the vacuum projection seems to be preserved in the chloroform projection, with the left-right gulf having been narrowed slightly and conformers being

distributed more diffusely. The DMSO projection, however, appears significantly different, with a new network of interconnected conformers appearing to be shown. The quantitative utility of these projections is an interesting question in such high dimensions; it is not possible to completely replicate the high-dimensional arrangement of every sampled configuration in a two-dimensional projection, and with a dimensional reduction from 7 to 2, it is likely that a large amount of information is lost. Therefore, the utility of these extremely reduced projections may be limited to providing a visual source of intuition for the degree to which conformational landscapes change in different environments.

Figure 5.11 compares the conformational free energy minima in each solvent environment with the experimentally determined crystal structures. Figure 5.11a shows the most stable conformer in DMSO, which has an open conformation, as observed experimentally by Sobornova et al [70]. Despite this being an open conformation, it is distinct from the conformation observed in form I of bicalutamide, illustrated in Figure 5.11c. The most stable conformer in chloroform is shown in Figure 5.11b and demonstrates a closed conformation, again matching the experimental observation of Sobornova et al[70]. Additionally, the conformation adopted in chloroform corresponds closely with the conformation adopted in crystal form II, shown in Figure 5.11d.

With confirmation that the most stable conformers in each environment line up with experimental observation, a more extensive analysis on the relationships between distinct conformers and solvent environment can be carried out. Figure 5.10 shows a double matrix comparing all bicalutamide conformers in chloroform to all bicalutamide conformers in DMSO. Each element's number corresponds to the difference in free energy between the two conformers (relative to the most stable conformers in their own environment). The color gradient indicates the similarity of the structures, measured by their minimum RMSD separation, considering all atom positions. Comparing Figure 5.10 to Figure 5.4, the equivalent figure for PBH, reveals that in PBH, there is approximately one set of conformer pairs with extremely low RMSD separation for each row and column of the matrix, indicating that al-

most every conformer in dichloromethane was also present in acetone. This not the case in Figure 5.10, where there are relatively few conformers common to both solvent environments. In order to study these common conformers more closely, we define the set of common conformers to be those conformer pairs which exhibit a minimum RMSD of less than 1.7 Å. There are 10 of these common conformers for bicalutamide in chloroform and DMSO, and they are shown in Figure 5.12. The overlapped conformers are shown in blue for chloroform and red for DMSO. It is interesting to note that the fully closed conformation observed as the most stable form in chloroform and in bicalutamide's crystalline form II, shown in Figure 5.11b,d, is not present as a common conformer, meaning it does not arise at all in DMSO. The majority of the common conformers shown in Figure 5.12 seem to exhibit a semi-open 'L'-shaped conformation, rather than the fully open and closed conformations seen as the most stable conformers in DMSO and chloroform in Figure 5.11a,b. The conformers in Figure 5.12 as ordered by ΔF where

$$\Delta F = F_{CLF} - F_{DMSO},$$

where F_{CLF} and F_{DMSO} are the free energies of the conformers in chloroform and DMSO respectively. Note that, as before, these individual free energies are themselves relative to the lowest energy configuration within the free energy landscape.

Figure 5.12 thus seems to show that these 'L'-shaped conformers tend to be stabilized in chloroform, the same environment that promotes the fully closed conformer, and that common conformers with a greater open character tend to be stabilized by DMSO.

It is also interesting to consider the common conformer labelled CLF5-DMSO11, corresponding to the conformer 5 in the chloroform landscape and 11 in the DMSO landscape. This conformation closely resembles bicalutamide crystal form I, as shown on Figure 5.11c. This common conformer has a low ΔF of 0.68 kJmol^{-1}, indicating that is equally stabilized by both solvents, but has an F_{CLF} of 16.22 kJmol^{-1} and an F_{DMSO} of 15.54 kJmol^{-1}, making it far from the most sta-

ble conformer in either solvent. From this it can be inferred that while the form I conformation is metastable in solution, packing effects are responsible for the conformational rearrangements leading to the observed conformational polymorph.

5.3 Taltirelin

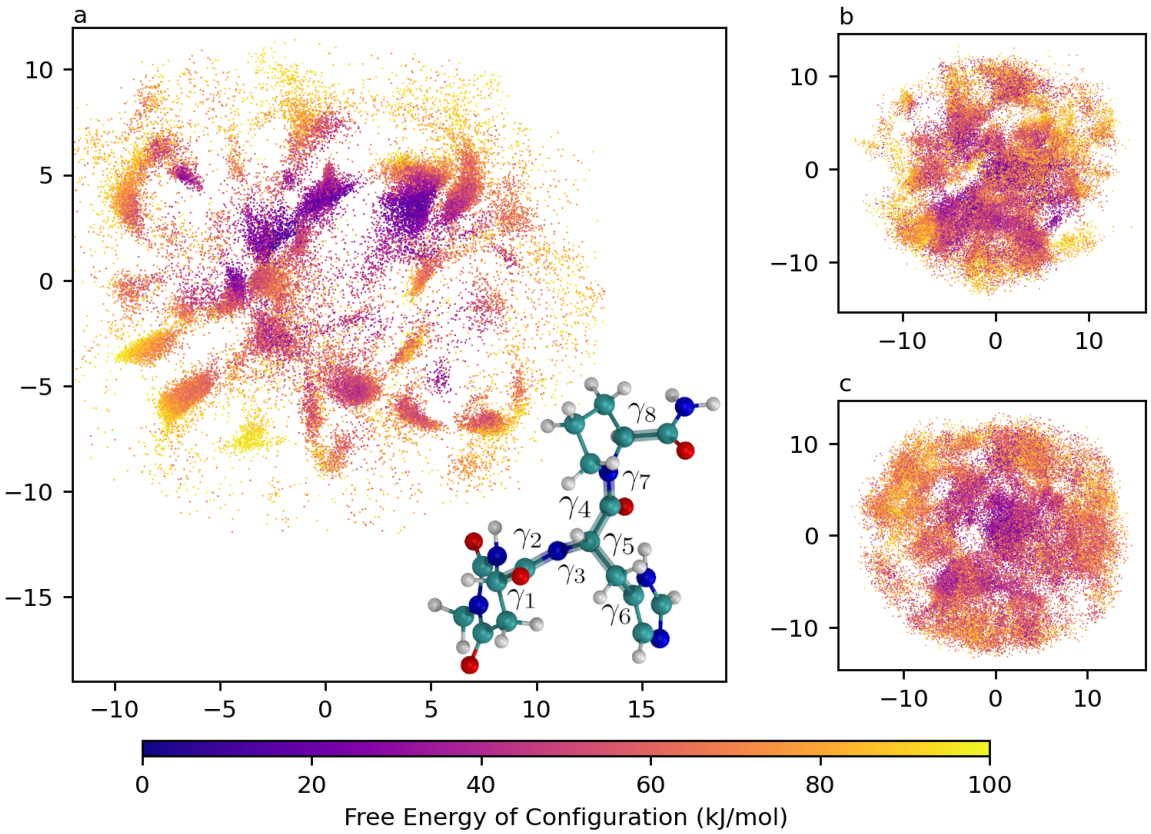


Figure 5.13: 2D Sketch-Map projection of taltirelin's 8D conformational free energy landscape in vacuum (a), water (b), and a water-methanol mixture (c), with molecular structure of taltirelin inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

Taltirelin is a TRH analog used to treat spinocerebellar ataxia [72]. It has an 8-dimensional conformation space with configurations defined by the 8 torsions shown inset on Figure 5.13a. Experimentally, it exhibits two conformationally polymorphic crystal structures, as reported by Maruyama et al. [72]. The two conformers are characterized by the distance between the 1-methyl carbon on the diazinane ring and the 4 carbon in the imidazole group, as shown in Figure 5.17. This charac-

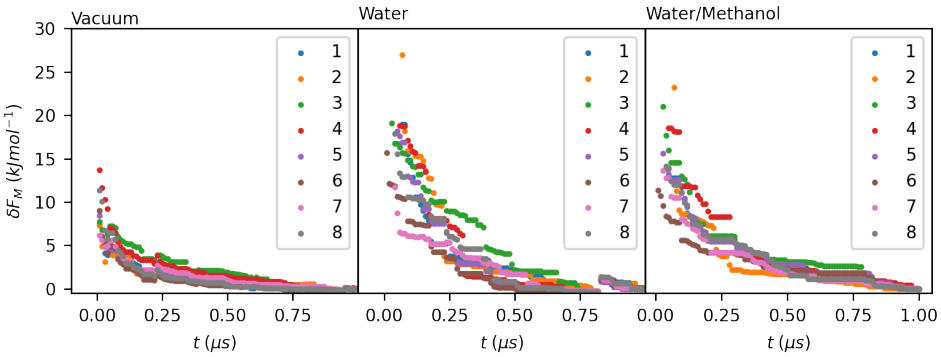


Figure 5.14: Evolution of the average free energy difference, $\delta F(t)$, on the marginal free energies of each torsion in taltirelin, in vacuum, water, and a water/methanol mixture.

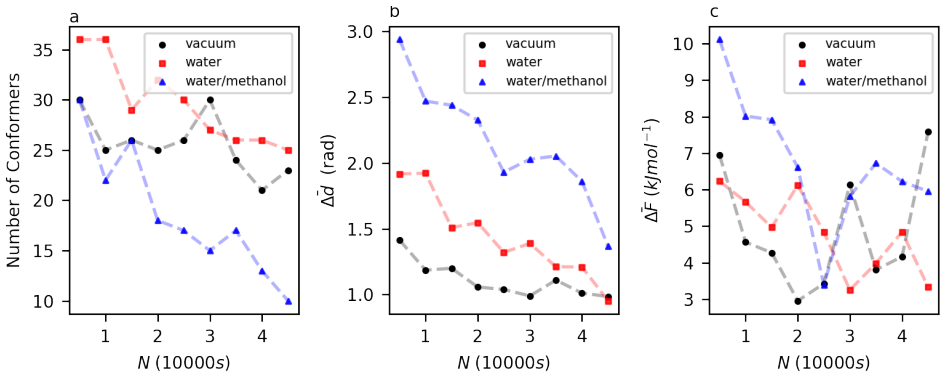


Figure 5.15: a: Number of clusters identified by clustering on datasets of size N for taltirelin, in vacuum, water, and a water/methanol mixture. b: Evolution of the average positional deviation, $\Delta\bar{d}$, with N for taltirelin, in vacuum, water, and a water/methanol mixture. c: evolution of the average free energy deviation, $\Delta\bar{F}$, with N for taltirelin, in vacuum, water, and a water/methanol mixture.

teristic interatomic distance is reported as being 2.9 Å in form I and 9.9 Å in form II [1]. Both forms are stabilized by an intramolecular hydrogen bond N-H···O between the primary amide group and the donor carbonyl on the diazinane ring. Form I is further stabilized by an N-H···O hydrogen bond between the secondary amine in the diazinane ring and the donor carbonyl on the tertiary amide. Form II is instead stabilized by a hydrogen bond N-H···O between the secondary amide and the donor carbonyl on the primary amide [72]. Form I can be recrystallized from pure water, while the addition of even a small amount of methanol will result in form II being obtained [72]. In a further study [1], Maruyama et al. used NOE spectroscopy to determine that methanol effects the conformational distribution of taltirelin in solu-

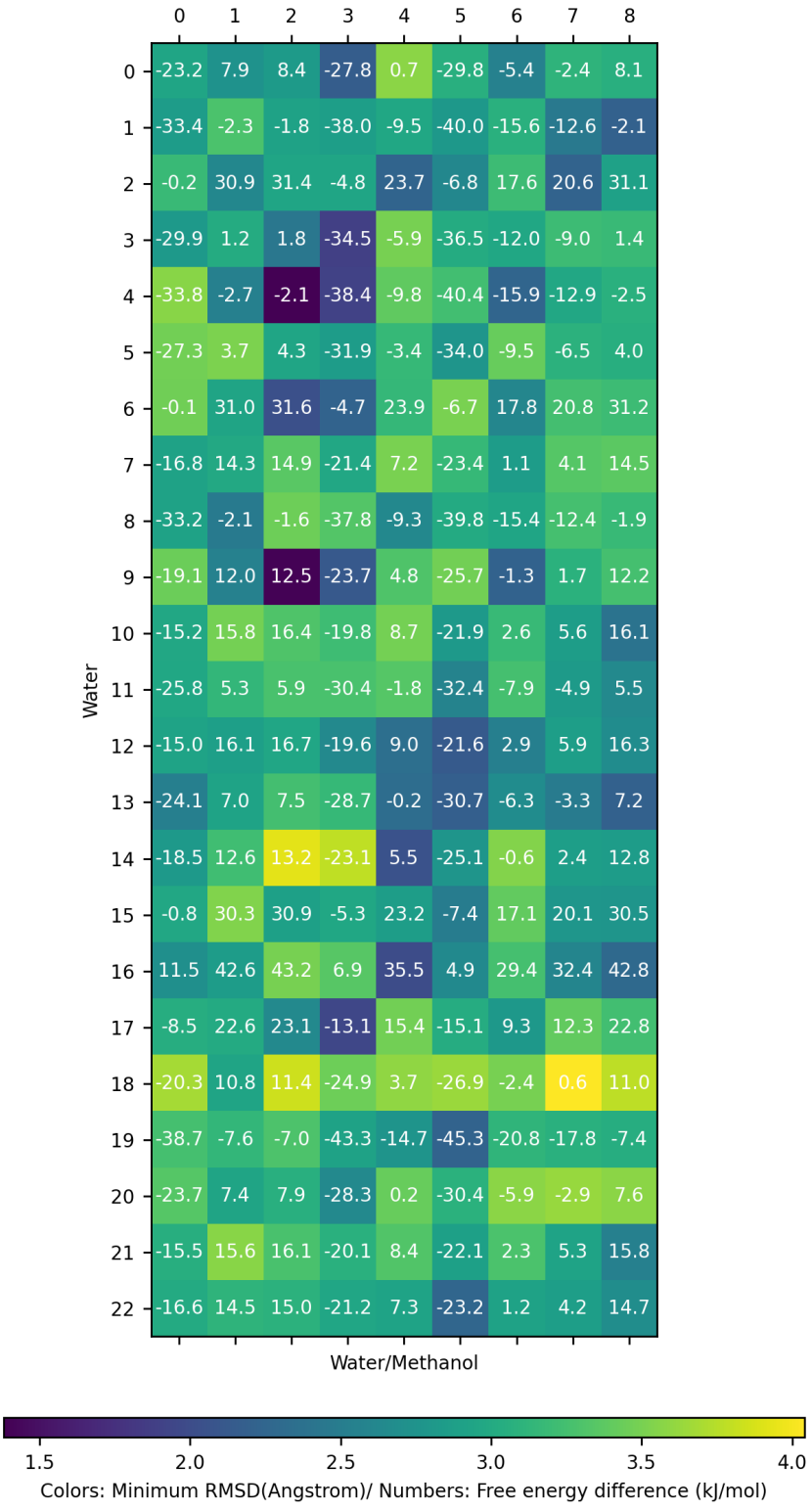


Figure 5.16: A dual matrix presenting a pairwise comparison of conformers of taltirelin observed in water (rows) and a water/methanol mixture(columns). The color gradient indicates the minimum RMSD between atoms between the two conformers, while the number within each element indicates the stability of the conformer in water relative to the conformer in the water/methanol mixture.

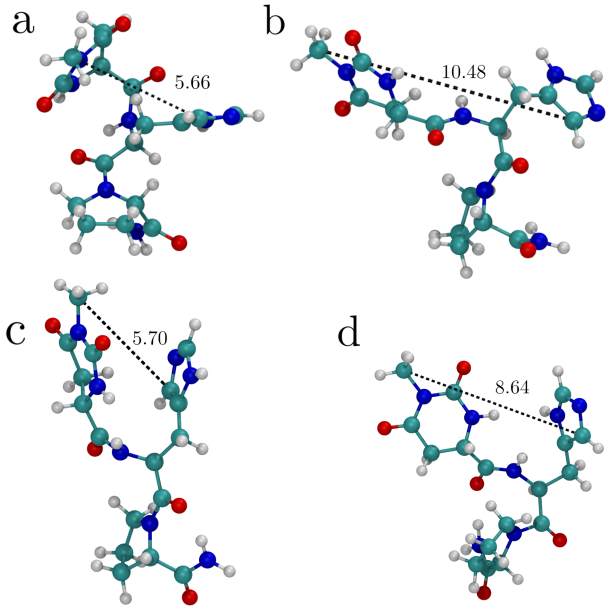


Figure 5.17: a: The lowest energy conformation of taltirelin in water, index 19 in Table B.8 and Figure 5.13b. b: One of the lowest energy conformations of taltirelin in water/methanol, index 1 in Table B.9 and Figure 5.13c. c: One of the lowest energy conformations of taltirelin in water/methanol, index 2 in Table B.9 and Figure 5.13c. d: One of the lowest energy conformations of taltirelin in water/methanol, index 8 in Table B.9 and Figure 5.13c. The free energies of conformers b-d are within 0.5 kJ/mol of one another, thus all three can be equally considered to be global free energy minima. All structures have the characteristic distance between ring groups defined by Maruyama et al. [1] shown.

tion, stabilizing form II. This stabilization effect could be in part responsible for the eventual isolated polymorph.

As with the other molecules studied in this chapter, simulations of taltirelin were carried out in vacuum as well as two distinct solvent environments, in this case pure water, and an approximately 80:20 mixture of water and methanol. Simulations were carried out as outlined in the Methods Chapter, and a gridless analysis was carried out on 50,000 configurations sampled from each simulation.

Figure 5.14 shows that the 1-dimensional marginal free energies converge within the simulation timeframe for all torsions. Higher dimensional consistency metrics are plotted in Figure 5.15. Like in the other high-dimensional conformation spaces studied in this work, the analysis does not assign a consistent number of conformers to taltirelin in any environment, as shown by Figure 5.15a. The fluctuation

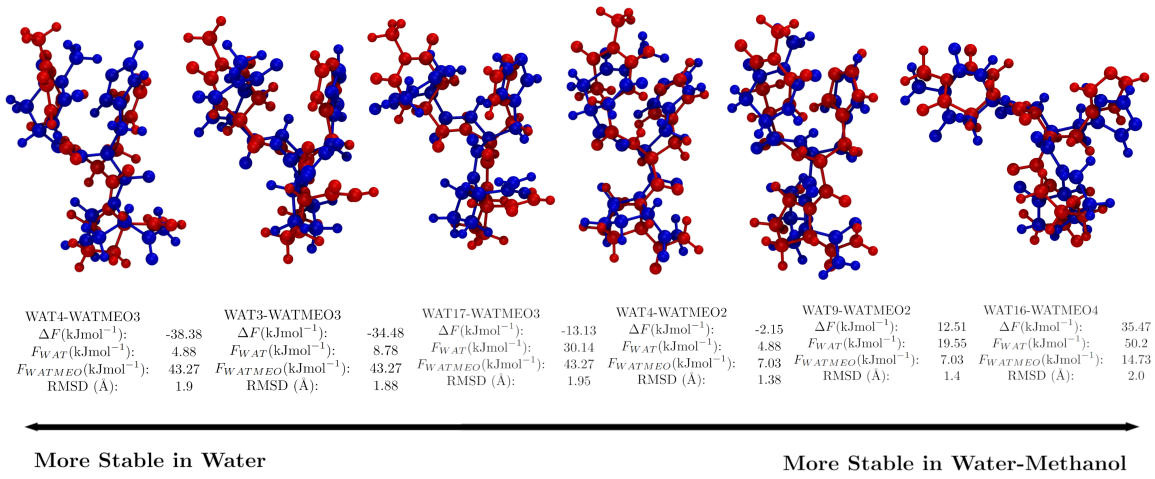


Figure 5.18: 6 common conformers of taltirelin in water and a water/methanol mixture. For this system, conformers are deemed common to be both solvents if their overlaid structures present a minimum RMSD deviation of less than 2.0Å. Conformers in water are indicated with the label WAT, and have their molecular structures shown in blue. Conformers in the water/methanol mixture are indicated with the label WATMEO and their molecular structures are shown in red. The free energy in both solvents, as well as the difference in free energy is indicated for each conformer, and the conformers are ordered from those most stabilized in water to those most stabilized in the water/methanol mixture.

seems to become minor for the three largest dataset sizes in vacuum and water, but continues to drop steeply in water/methanol. It is interesting to note in this plot that a consistently smaller number of conformers are observed in water/methanol compared to pure water, across all dataset sizes.

Figure 5.15b,c shows how $\bar{\Delta d}$ and $\bar{\Delta F}$ evolve with dataset size, respectively. The trend in $\bar{\Delta d}$ shows that the positions of equivalent conformers drift by between 1 and 1.5 Euclidean radians. The trends in $\bar{\Delta F}$ are particularly inconsistent, with energies potentially deviating by up to 8 kJmol^{-1}. As with bicalutamide and other previously studied high-dimensional molecules, the high dimensionality and low data-density result in these inconsistencies in conformer position and free energy. Despite the low resolution of this approach in these cases, a meaningful interpretation of these conformational free energy landscapes is still possible, due to the wide range in free energies obtained for the different structures. If increased consistency is required, it is likely that the expensive step of increasing dataset size would

need to be taken, demanding the use of more powerful hardware. This is further discussed in Chapter 6.

As with bicalutamide, 2-dimensional projections of taltirelin's 8-dimensional per-point free energy landscapes were created with Sketch-map, using the parameter and landmarks used in the in-vacuo case. They are shown in Figures 5.13a, 5.13b, 5.13c, for taltirelin in vacuum, water, and water/methanol respectively. For each of the projections, an enlarged image featuring the locations of the free energy minima marked with a numerical label is available in Figures B.7, B.8, B.9 in Appendix B. These labels correspond to the conformer indices in the left-hand columns of Tables B.7, B.8, and B.9, also in Appendix B, for taltirelin in vacuum, water, and water/methanol, respectively.

Examination of the three Sketch-map projections reveals a clearly defined landscape in vacuum, that becomes distorted when considering taltirelin in water, and more distorted still when considering taltirelin in water/methanol. Furthermore, there are regions of what appear to be low free energy in the water/methanol projection which have not been identified as conformers.

These are likely effects resulting from the decision to use the same Sketch-map landmarks and parameters as those determined to be ideal for vacuum for all environments, and not a failure of DPA to identify density peaks. Should the molecule's occupation of the conformation space in solution be sufficiently different from its occupation in vacuum, it likely that the projection will be of poorer quality. This suggests that the impact of solvent on taltirelin's conformational distribution is extensive, with solvent extensively flattening the landscape. An interesting contrast to this case are the Sketch-map projections of m-nisoldipine analyzed in the next section, which change little across different environments.

Figure 5.17 shows the most stable conformers of taltirelin in both solvent environments. Figure 5.17a shows the most stable form of taltirelin in water. Taltirelin has three stable conformers within 0.55 kJmol^{-1} of each other in water/methanol, thus all three conformers can be considered to equally the most stable. These are shown in Figure 5.17b,c,d. For each of these structures, the characteristic inter-

atomic distance[1] is shown. The short interatomic distance of 5.66 Å in water results in a similar molecular geometry as the distance of 2.9 Å observed in the crystal form I. The three stable conformers in water/methanol demonstrate a diverse range of values of the characteristic interatomic distance, ranging from the 10.48 Å shown in Figure 5.17b (compare with the distance of 9.9 Å observed in the crystal form II) to the 5.7 Å seen in Figure 5.17c.

These observations support the experimental observation that the addition of methanol to an aqueous solution of taltirelin promotes the population of conformational states resembling that of crystal form II. As the solvent environment is still predominantly aqueous, the conformers which dominate in pure water continue to exist in water/methanol, but now coexist with forms specific to water/methanol.

Figure 5.16 shows a matrix comparing all the conformers observed in water with all the conformers observed in water/methanol. As in Figures 5.10 and 5.4, the numbers in each element correspond to ΔF and the color gradient represents the minimum RMSD between the two structures. Despite the presence of three almost equally favored conformers in the water/methanol environment, this environment presents far fewer conformers overall. It is therefore of interest to visualize the shapes of the common conformers which feature in both water and water/methanol. Figure 5.18 shows the 6 common conformers of taltirelin in water and water/methanol, where conformers have been deemed to be common if their minimum RMSD was smaller than 2 Å. These common conformers are ordered based on whether their geometries are stabilized in pure water, or in the water/methanol mixture, according to their ΔF. It is immediately apparent that the majority of the conformers stable in both solvents correspond to the form I conformer characterized by the short interatomic distance. The only common conformer (WAT16-WATMEO4) presenting an elongated distance is far more stable in water/methanol than in pure water. This in agreement with the observation that the conformation characteristic of form II is stabilized by the addition of methanol to aqueous solution, while the conformation characteristic of form I is found in any aqueous solvent environment.

5.4 m-Nisoldipine

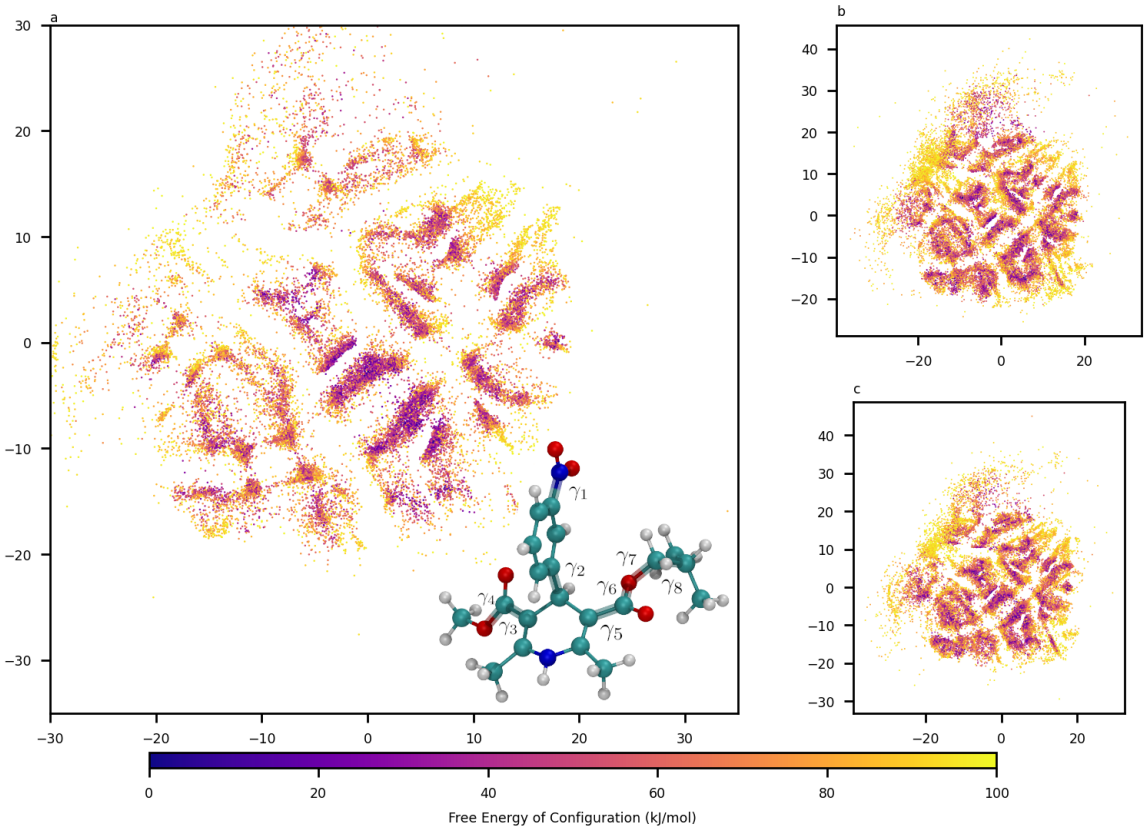


Figure 5.19: 2D Sketch-Map projection of m-nisoldipine's 8D conformational free energy landscape in vacuum (a), an acetone-ethanol mixture(b), and an ethyl acetate - hexane mixture(c), with molecular structure of m-nisoldipine inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

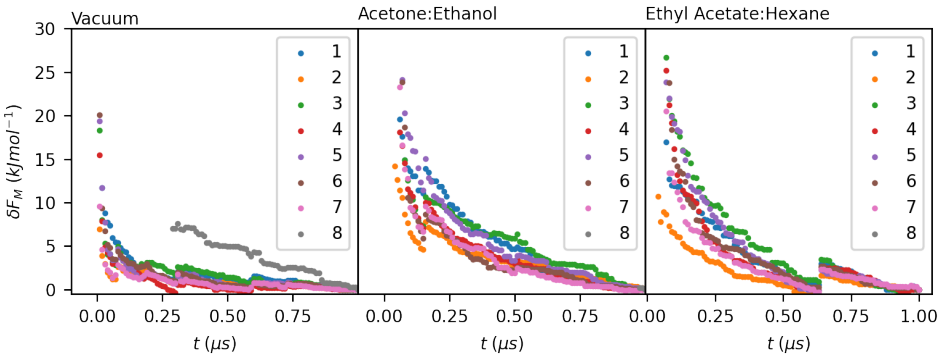


Figure 5.20: Evolution of the average free energy difference, $\delta F(t)$, on the marginal free energies of each torsion in m-nisoldipine, in vacuum, an acetone/ethanol mixture, and an ethyl acetate/hexane mixture.

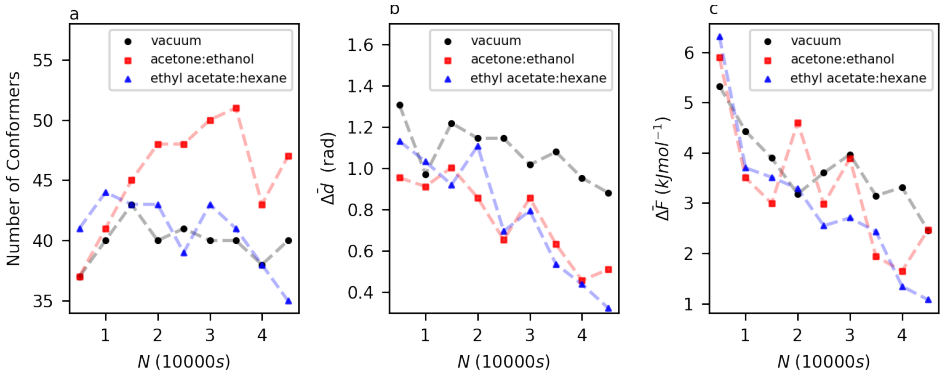


Figure 5.21: a: Number of clusters identified by clustering on datasets of size N for m-nisoldipine, in vacuum, an acetone/ethanol mixture, and an ethyl acetate/hexane mixture. b: Evolution of the average positional deviation, $\Delta\bar{d}$, with N for m-nisoldipine, in vacuum, an acetone/ethanol mixture, and an ethyl acetate/hexane mixture. c: evolution of the average free energy deviation, $\Delta\bar{F}$, with N for m-nisoldipine, in vacuum, an acetone/ethanol mixture, and an ethyl acetate/hexane mixture.

m-Nisoldipine is a calcium ion agonist used to treat high blood pressure. Like taltirelin, it has an 8-dimensional conformation space. This space is defined by the 8 torsions shown inset in Figure 5.19a. Unlike the other molecules in this chapter, there are no experimental results regarding m-nisoldipine's conformational distribution in the solution phase. However, it does exhibit solvent-mediated conformational polymorphism in the solid state, with form A obtained through recrystallization of a 1:1 mixture of acetone and ethanol, and form B obtained through recrystallization of a 1:1 mixture of ethyl acetate and hexane. The conformers observed in form A and B are shown in Figure 5.23c and d respectively. Both crystal forms are primarily stabilized by an intermolecular hydrogen bond N-H···O between the amine group and the carbonyl on the tert-butyl substituted ester[73].

As with the other molecules studied within this chapter, concurrent WTMetaD simulations were run in vacuum and in both solvent environments, followed by the creation of gridless free energy landscapes from 50,000 configurations distributed within the conformational space.

Figure 5.20 shows the convergence of all 1-dimensional marginal FESs of all 8 torsions in the three simulation environments. Convergence occurs as expected for all marginals within the simulation timescale. Higher dimensional consistency

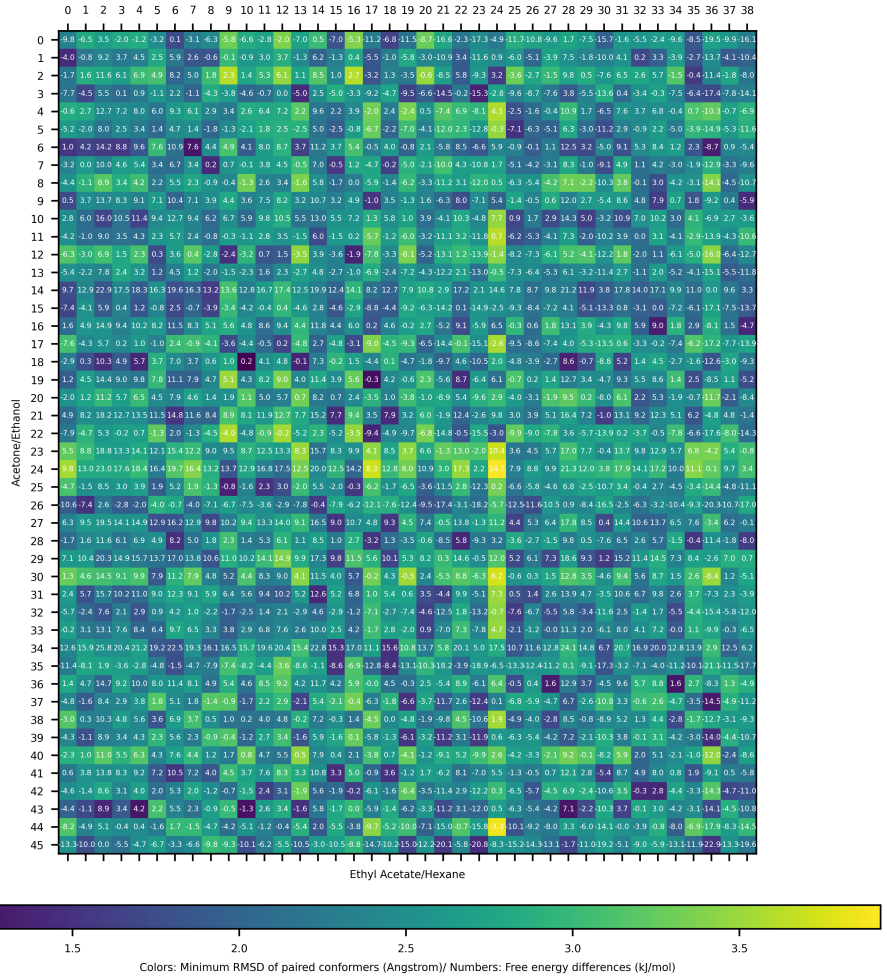


Figure 5.22: A dual matrix presenting a pairwise comparison of conformers of m-nisoldipine observed in an acetone/ethanol mixture (rows) and a ethyl acetate/hexane mixture(columns). The color gradient indicates the minimum RMSD between atoms between the two conformers, while the number within each element indicates the stability of the conformer in the acetone/ethanol mixture relative to the conformer in the ethyl acetate/hexane mixture. For the sake of legibility, this figure is available in a larger size in Figure B.13, in Appendix B.

checks are shown in Figure 5.21. Figure 5.21a shows how the number of conformers identified varies with dataset size, and as with other high-dimensional systems, complete consistency is not observed in any of the environments. Results in vacuum and ethyl acetate/hexane appear more consistent than those in acetone/ethanol, where swings of up to 8 conformers are present even towards the largest dataset sizes employed. The large number of conformers in all environments is noteworthy, when compared to landscapes previously studied in this chapter. Figure 5.21b,c

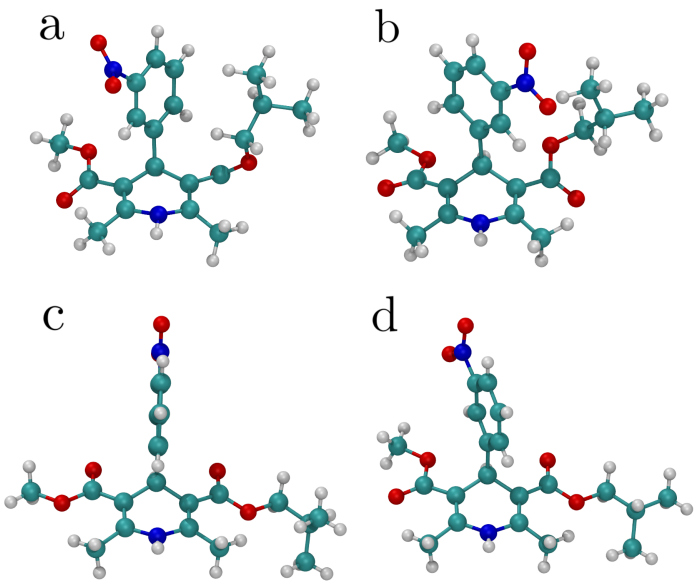


Figure 5.23: a: The lowest energy conformation of m-nisoldipine in acetone/ethanol, index 45 in Table B.11 and Figure 5.19b. b: The lowest energy conformation of m-nisoldipine in ethyl acetate/hexane, index 2 in Table B.12 and Figure 5.19c. c: The experimentally observed conformation of m-nisoldipine form I. d: The experimentally observed conformation of m-nisoldipine form II.

show the evolution of $\bar{\Delta d}$ and $\bar{\Delta F}$ respectively. These plots demonstrate that conformers are consistently placed with 0.6 Euclidean radians of separation from one another for the two solvated environments, and that these conformers are within 4 kJmol^{-1} of one another. Considering that m-nisoldipine's conformation space has the same dimensionality as that of taltirelin, it is interesting to note that these consistency metrics are considerably improved over those observed in the previously considered molecule.

Figures 5.19a, 5.19b, 5.19c show the two-dimensional projection of the 8-dimensional per-point conformational free energy landscapes in the three simulation environments. For each of the projections, an enlarged image featuring the locations of the free energy minima marked with a numerical label is available in Figures B.10, B.11, B.12 in Appendix B. These labels correspond to the conformer indices in the left-hand columns of Tables B.10, B.11, and B.12, also in Appendix

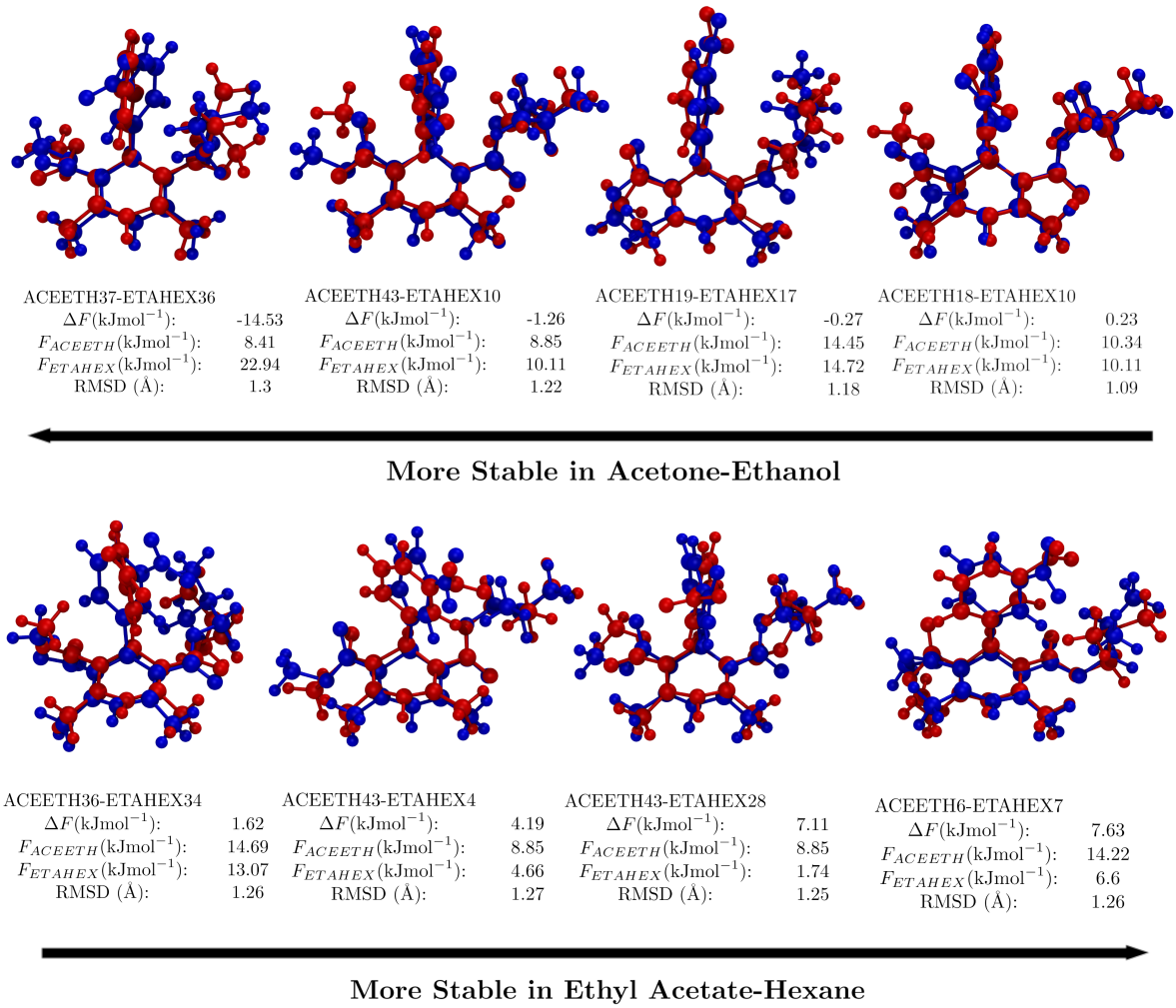


Figure 5.24: 6 common conformers of m-nisoldipine in acetone/ethanol and a ethyl acetate/hexane mixtures. For this system, conformers are deemed common to both solvents if their overlaid structures present a minimum RMSD deviation of less than 1.3Å. Conformers in acetone/ethanol are indicated with the label ACEETH, and have their molecular structures shown in blue. Conformers in the ethyl acetate/hexane mixture are indicated with the label ETAHEX and their molecular structures are shown in red. The free energy in both solvents, as well as the difference in free energy is indicated for each conformer, and the conformers are ordered from those most stabilized in acetone/ethanol to those most stabilized in the ethyl acetate/hexane mixture.

B, for m-nisoldipine in vacuum, acetone/ethanol, and ethyl acetate/hexane, respectively. Unlike the projections observed for bicalutamide and taltirelin, the impact of solvent on the projection appears to be minimal. The same distribution of states is found across all three projections, suggesting that the conformational distribution of m-nisoldipine is not largely affected by the bulk solvent environment. Furthermore,

the character of the landscape is distinctly different from those seen in taltirelin and bicalutamide, with a large number of clearly separated basins, all occupying similar energy levels. Examination of Tables B.11 and B.12 shows that there are 4 conformers within 5 kJmol^{-1} of the minimum for m-nisoldipine in acetone/ethanol and 5 conformers within 5 kJmol^{-1} of the minimum for m-nisoldipine in ethyl acetate/hexane. This number grows to 21 in acetone/ethanol and 19 and ethyl acetate/hexane for conformers within 10 kJmol^{-1} of the minimum, confirming this observation gleaned from the projections.

Figure 5.23 compares the two most stable conformers in the solvent environments to crystal forms A and B. Unlike bicalutamide and taltirelin, this comparison seems to reveal that both the two conformers in solution have more in common with each other than the crystal forms to which their environments are experimentally linked. Despite the fact that acetone/ethanol (most stable conformer shown in Figure 5.23a) promotes crystal form A (shown in Figure 5.23c), and that ethyl acetate/hexane (most stable conformer shown in Figure 5.23b) promotes crystal form B (shown in Figure 5.23d), the orientation of the nitrobenzene ring and the isopropyl ester groups have preferred orientations in solvent and crystal phases, regardless of solvent and crystal form. The methyl ester group also appears to favor its orientation in form B in both solvent environments.

Figure 5.22 shows a double matrix comparing the difference in conformer free energy ΔF and minimum RMSD of every combination of conformers in both acetone/ethanol and ethyl acetate/hexane. Compared to the equivalent matrices presented previously for PBH, bicalutamide and taltirelin, the large number of conformers in both of these environments makes this matrix cumbersome to interpret. It does appear that many of the elements showing a low RMSD exhibit a relatively small magnitude of ΔF, though to examine this more carefully, all conformers with an RMSD separation of less than 1.3 Å are presented in Figure 5.24.

Consideration of these common conformers reveals that none of these structures, in either acetone/ethanol or ethyl acetate/hexane, exhibit the orientation of the isopropyl ester group seen in the two crystal forms. The magnitude of ΔF at the

extremes is also smaller than those observed in bicalutamide and taltirelin, with 4 of the 8 structures shown have a ΔF magnitude smaller than 2 kJmol^{-1} and 7 of the 8 have a ΔF magnitude smaller than 8 kJmol^{-1}. Despite this, there is diversity in the positions of the methyl ester, as well as the nitrobenzene ring. This consistent with the observation made that the free energy landscapes shown in Figures 5.19a, 5.19b, 5.19c consist of many basins of similar free energy and that these landscapes do not seem to be dramatically affected by solvent environment. From this, it must be surmised that the impact of solvent choice on crystallization does arise in the bulk solution phase.

5.5 Conclusion

In this chapter, the utility of the gridless per-point conformational free energy landscapes generated by the workflow developed in this project was demonstrated on four molecular systems, studied in a pair of solvent environments associated with distinct conformational behavior, either in solution, in the crystal phase or both. MD simulations of the first molecule, PBH, did not replicate the experimentally observed behavior, and the conformational polymorphism of m-nisoldipine was not found to be linked to the conformational distribution of m-nisoldipine in solution. However, results for bicalutamide and taltirelin seemed to support experimental observations. For all molecules, a combination of Sketch-map projections, energy-RMSD matrices, and renders of selected molecular structures allowed some degree of analysis and understanding of the high-dimensional conformational free energy landscapes. Despite this, the problem of inconsistency in conformer number, position, and free energy in high-dimensional molecules, initially observed in Target XXXII in the previous chapter, persists in bicalutamide, taltirelin, and m-nisoldipine. With a dataset size of 50,000 configurations representing the upper limit possible using currently available hardware, the relationship between dataset size, consistency, and conformation space dimensionality is still poorly understood. It is likely that increasing the size of the datasets further would improve the quality of results, however, the difference in consistency metrics between taltirelin and

m-nisoldipine (both with 8 dimensional conformation spaces), suggests that there is also a system-dependent effect on consistency.

Chapter 6

Research Outlook

A new analysis method, based on the use of DPA clustering, creates high-dimensional conformational FESes in a gridless, computationally accessible way, allowing the conformational ensembles of highly flexible molecules to be characterized in a systematic and efficient way which scales better than the conventional grid-based approach. Pairing DPA's density estimation tool with a dataset of configurations generated through concurrent well-tempered metadynamics simulations allows for quantitative per-point FES generation through a simple Zwanzig-based reweighing scheme. DPA's classification approach further allows for automatic, high-dimensional interpretation of these FESes. This approach has been initially demonstrated for systems with 2, 4, and 11-dimensional conformation spaces in vacuum, before being used in a study on the relationship between solvent and conformational distribution for 4 conformationally complex molecules. The performance of this method was tracked using a set of consistency metrics that enable its application in realistic cases. The approach is entirely simulation agnostic, operating solely on the coordinates of configurations in conformation space and their corresponding biases. We thus envision applications for simulations performed at a broad range of theory levels and across various physical environments.

A key area for future study is the relationship between the dimensionality of the conformation space, the size of the configurational dataset, and the quality of the resulting consistency metrics. In this work, conformer-sets identified with the method for molecules with a complex conformation space (7+ torsions) have failed

to produce a consistent number of conformers and demonstrate inconsistencies in the free energies and positions of the conformers identified. The majority of the landscapes presented in this work are the result of clustering on datasets of 50,000 configurations. As the dimensionality increases, the density of data in conformation space decreases exponentially, a scenario demonstrated in the example of target XXXII, where the dataset possessed a data density of 8×10^{-5} configurations per rad^{11} in the 11-dimensional conformation space. To increase the consistency observed in these cases, it is likely that increasing the size of the dataset will be required, however, this is not a trivial step to take. As discussed in the Methods chapter, the most computationally expensive stage of the process is computing the pairwise Euclidean distance matrix for the conformational dataset. Thus, while the increase in cost associated with increasing dimensionality scales linearly, the cost associated with increasing dataset size scales quadratically. It should be stressed that this cost is still much smaller than the exponential costs incurred in the construction of high-dimensional grids, and that all the analysis carried out within this project occurred on a desktop workstation, where dataset sizes could not realistically be pushed higher than the 50,000 configurations used here. There exists more powerful hardware which could be used to explore this relationship between dataset size and consistency. As briefly discussed in Chapter 5, the difference in the consistency metrics for m-nisoldipine and taltirelin, both of which are 8-dimensional, suggests that there are system-specific effects on the consistency of the free energy landscape.

One potential solution to the inconsistencies observed in high-dimensional conformation spaces lies in the work of Olehnovics et al. [74] [75]; using machine-learned invertible maps [76], they calculate the relative free energies of states sampled in separate simulations without the sampling of intermediate states to create overlaps in the probability distribution. In the context of this work, this would be equivalent running independent unbiased simulations in each of the conformational basins detected, without any requirement for the simulation to explore the conformation space. A potential workflow would involve using the free energy landscapes

here as a rough roadmap, which then informs a set of unbiased simulations, each of which explores one of conformer basins identified by the approach here. Due to limitations in the training process, this approach would only be feasible for studies of molecules in vacuum. That said, the potential to increase the precision of the free energy estimates is high.

Another interesting direction of future research would be the study of conformation space in a greater range of environments. Marinova et al. [11] produced conformational FESes of ibuprofen along a range of stages of crystal growth, from the solution bulk, to partial embedding in the crystal surface at the solution interface, to a molecule embedded fully in the crystal bulk. For each of the transitions, the reduction in ibuprofen's flexibility is plain to see in as the fraction of the FES accessed shrinks. This work was based on a two-dimensional conformation space, with the FESes generated on a grid. Observing the constriction of higher-dimensional conformation spaces as a molecule transitions from the solution to solid phase could offer detailed mechanistic insight into the process of crystal growth.

The analysis method presented in this work has been demonstrated here on a handful of molecules, some in vacuum, and some in solvated environments. However, the ultimate aim of this project was to create a tool with broad applicability, suitable for a range of molecular conformation spaces analyzed by simulations at any level of theory and in any environment. In this spirit, the code developed for the analyses carried out here is publicly available and open source. Included in this code is a tutorial notebook which facilitates its application. A record of this notebook can be found in Appendix C. This tutorial should make the creation and analysis of the 'per-point' FESes outlined here widely accessible.

The code developed here is and available from
https://github.com/ucecvan/Twister.

Bibliography

[1] Shoji Maruyama, Hiroshi Ooshima, and Jyoji Kato. Crystal structures and solvent-mediated transformation of taltireline polymorphs. *Chemical Engineering Journal*, 75, 1999.

[2] Daniel E Koshland. Application of a theory of enzyme specificity to protein synthesis*. *Proceedings of the National Academy of Sciences*, 44:98–104, 2 1958. doi: 10.1073/pnas.44.2.98.

[3] Hans Frauenfelder, Stephen G. Sligar, and Peter G. Wolynes. The energy landscapes and motions of proteins. *Science (New York, N.Y.)*, 254:1598–1603, 1991.

[4] Aurora J Cruz-Cabeza and Joel Bernstein. Conformational polymorphism. *Chemical Reviews*, 114:2170–2191, 2 2014. doi: 10.1021/cr400249d.

[5] Joel Bernstein and Arnold T Hagler. Conformational polymorphism. the influence of crystal structure on molecular conformation. *Journal of the American Chemical Society*, 100:673–681, 2 1978. doi: 10.1021/ja00471a001.

[6] Stephen Byrn, Ralph Pfeiffer, Michael Ganey, Charles Hoiberg, and Guirag Poochikian. Pharmaceutical solids: A strategic approach to regulatory considerations. *Pharmaceutical Research: An Official Journal of the American Association of Pharmaceutical Scientists*, 12:945–954, 1995.

[7] Paolo Lucaioli, Elisa Nauha, Ilaria Gimondi, Louise S Price, Rui Guo, Luca Iuzzolino, Ishwar Singh, Matteo Salvalaglio, Sarah L Price, and Nicholas Blagden. Serendipitous isolation of a disappearing conformational polymorph

of succinic acid challenges computational polymorph prediction. *CrystEngComm*, 20:3971–3977, 2018.

[8] Berend Smit and Daan Frenkel. Understanding molecular simulation, second edition (computational science series, vol 1). Technical report, 2002.

[9] Jos P.M. Lommerse, W. D. Sam Motherwell, Herman L. Ammon, Jack D. Dunitz, Angelo Gavezzotti, Detlef W.M. Hofmann, Frank J.J. Leusen, Wijnand T.M. Mooij, Sarah L. Price, Bernd Schweizer, Martin U. Schmidt, Bouke P. Van Eijck, Paul Verwer, and Donald E. Williams. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica Section B: Structural Science*, 56:697–714, 8 2000.

[10] Rabia Bushra and Nousheen Aslam. An overview of clinical pharmacology of ibuprofen. *Oman Medical Journal*, 25:155, 7 2010.

[11] Veselina Marinova, Geoffrey P F Wood, Ivan Marziano, and Matteo Salvalaglio. Dynamics and thermodynamics of ibuprofen conformational isomerism at the crystal/solution interface. *Journal of Chemical Theory and Computation*, 14:6484–6494, 12 2018. doi: 10.1021/acs.jctc.8b00702.

[12] Veselina Marinova, Geoffrey P.F. Wood, Ivan Marziano, and Matteo Salvalaglio. Solvent dynamics and thermodynamics at the crystal-solution interface of ibuprofen. *Crystal Growth and Design*, 19:6534–6541, 11 2019.

[13] Veselina Marinova, Laurence Dodd, Song-Jun Lee, Geoffrey P F Wood, Ivan Marziano, and Matteo Salvalaglio. Identifying conformational isomers of organic molecules in solution via unsupervised clustering. *Journal of Chemical Information and Modeling*, 61:2263–2273, 5 2021. doi: 10.1021/acs.jcim.0c01387.

[14] Irwin Goldstein, Tom F. Lue, Harin Padma-Nathan, Raymond C. Rosen, William D. Steers, and Pierre A. Wicker. Oral sildenafil in the treatment of erectile dysfunction. sildenafil study group. *The New England journal of medicine*, 338:1397–1404, 5 1998.

[15] Daniel F. Veber, Stephen R. Johnson, Hung Yuan Cheng, Brian R. Smith, Keith W. Ward, and Kenneth D. Kopple. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, 45:2615–2623, 6 2002.

[16] David Ferro-Costas, Irea Mosquera-Lois, and Antonio Fernández-Ramos. Torsiflex: an automatic generator of torsional conformers. application to the twenty proteinogenic amino acids. *Journal of Cheminformatics*, 13, 2021.

[17] Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22:7169–7192, 4 2020.

[18] Bernardo de Souza. Goat: A global optimization algorithm for molecules and atomic clusters. *Angewandte Chemie International Edition*, 3 2025.

[19] Christopher Zurek, Ruslan A. Mallaev, Alexander C. Paul, Nils van Staalduinen, Philipp Pracht, Roman Ellerbrock, and Christoph Bannwarth. Tensor train optimization for conformational sampling of organic molecules. *Journal of Chemical Theory and Computation*, 2 2025.

[20] Philipp Pracht, Stefan Grimme, Christoph Bannwarth, Fabian Bohle, Sebastian Ehlert, Gereon Feldmann, Johannes Gorges, Marcel Müller, Tim Neudecker, Christoph Plett, Sebastian Spicher, Pit Steinbach, Patryk A. Wesołowski, and Felix Zeller. Crest—a program for the exploration of low-energy molecular chemical space. *Journal of Chemical Physics*, 160:114110, 3 2024.

[21] Robert W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22:1420, 12 2004.

[22] Erik R. Lindahl. *Molecular Modeling of Proteins*. Humana Press, first edition, 2008.

[23] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. C. Berendsen. Gromacs: Fast, flexible, and free. *Journal of Computational Chemistry*, 26:1701–1718, 12 2005.

[24] Haskell B. Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2:258–261, 1944.

[25] Peter Atkins and Julio De Paula. *Physical Chemistry*. Oxford University Press, ninth edition, 2010.

[26] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159:98, 7 1967.

[27] Michael P. Allen and Dominic J. Tildesley. *Molecular Simulations of Liquids*. Oxford University Press, 2nd edition, 2017.

[28] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25:1157–1174, 7 2004.

[29] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 6 1983.

[30] Peter Schwerdtfeger and David J. Wales. 100 years of the lennard-jones potential. *Journal of Chemical Theory and Computation*, 20:3379–3405, 5 2024.

[31] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092, 6 1953.

[32] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103:8577–8593, 11 1995.

[33] Berni J. Alder and Thomas. E. Wainwright. Studies in molecular dynamics. i. general method. *The Journal of Chemical Physics*, 31:459–466, 8 1959.

[34] Lindahl, Abraham, Hess, and van der Spoel. Gromacs 2019.3 manual. 6 2019.

[35] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18:1463–1472, 9 1997.

[36] Herman J C Berendsen, James P M Postma, Wilfred F van Gunsteren, Alberto DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684–3690, 10 1984.

[37] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81:511–519, 7 1984.

[38] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31:1695–1697, 3 1985.

[39] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126:014101, 1 2007.

[40] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1:826–843, 9 2011.

[41] Gerald Geudtner and Andreas M. Köster. First principles global optimization method from parallel tempering molecular dynamics. *Journal of Computational Chemistry*, 46:e70057, 3 2025.

[42] Stefano Piana and Alessandro Laio. A bias-exchange approach to protein folding. *The Journal of Physical Chemistry B*, 111:4553–4559, 5 2007. doi: 10.1021/jp067873l.

[43] Glen. M. Torrie and John. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23:187–199, 2 1977.

[44] Jagdish Suresh Patel, Anna Berteotti, Simone Ronsisvalle, Walter Rocchia, and Andrea Cavalli. Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5. *Journal of Chemical Information and Modeling*, 54:470–480, 2 2014.

[45] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100:020603, 1 2008.

[46] Massimiliano Bonomi, Alessandro Barducci, and Michele Parrinello. Reconstructing the equilibrium boltzmann distribution from well-tempered metadynamics. *Journal of Computational Chemistry*, 30:1615–1621, 8 2009.

[47] Estivill-CastroVladimir. Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, 4:65–75, 6 2002.

[48] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[49] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344:1492–1496, 6 2014.

[50] Maria d'Errico, Elena Facco, Alessandro Laio, and Alex Rodriguez. Automatic topography of high-dimensional data sets by non-parametric density peak clustering. *Information Sciences*, 560:476–492, 2021.

[51] Alex Rodriguez, Maria d'Errico, Elena Facco, and Alessandro Laio. Computing the free energy without collective variables. *Journal of Chemical Theory and Computation*, 14:1206–1215, 3 2018. doi: 10.1021/acs.jctc.7b00916.

[52] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.

[53] Jerzy Neyman, Egon Sharpe Pearson, and Karl Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1 1933. doi: 10.1098/rsta.1933.0009.

[54] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108:13023–13028, 8 2011. doi: 10.1073/pnas.1108486108.

[55] Alexandre Ferreira, Rui Guo, Ivan Marziano, and Matteo Salvalaglio. A gridless approach to sampling and classifying high-dimensional conformational landscapes of active pharmaceutical ingredients (preprint). *ChemRxiv*, 2 2025. doi = 10.26434/CHEMRXIV-2025-J2N83.

[56] Alejandro Gil-Ley and Giovanni Bussi. Correction to enhanced conformational sampling using replica exchange with collective-variable tempering. *Journal of Chemical Theory and Computation*, 11:5554, 11 2015. doi: 10.1021/acs.jctc.5b00981.

[57] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25:247–260, 10 2006.

[58] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185:604–613, 2014.

[59] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, 91(24):6269–6271, 1987.

[60] M. Scott Shell, Athanassios Panagiotopoulos, and Andrew Pohorille. *Free Energy Calculations*. Springer, 2007.

[61] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics 2020 2:4*, 2:200–212, 3 2020.

[62] Ilaria Gimondi, Gareth A Tribello, and Matteo Salvalaglio. Building maps in collective variable space. *The Journal of Chemical Physics*, 149:104104, 9 2018.

[63] Veselina Marinova and Matteo Salvalaglio. Time-independent free energies from metadynamics via mean force integration. *The Journal of Chemical Physics*, 151:164115, 10 2019.

[64] Gopalasamudram N Ramachandran, Chandrasekharan Ramakrishnan, and Viswanathan Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.

[65] Lily M Hunnisett, Jonas Nyman, Nicholas Francia, Nathan S Abraham, Claire S Adjiman, Srinivasulu Aitipamula, Tamador Alkhidir, Mubarak Almehairbi, Andrea Anelli, Dylan M Anstine, John E Anthony, Joseph E Arnold, Faezeh Bahrami, Michael A Bellucci, Rajni M Bhardwaj, Imanuel Bier, Joanna A Bis, A Daniel Boese, David H Bowskill, James Bramley, Jan Gerit Brandenburg, Doris E Braun, Patrick W V Butler, Joseph Cadden, Stephen Carino, Eric J Chan, Chao Chang, Bingqing Cheng, Sarah M Clarke, Simon J Coles, Richard I Cooper, Ricky Couch, Ramon Cuadrado, Tom Darden, Graeme M Day, Hanno Dietrich, Yiming Ding, Antonio DiPasquale, Bhausaheb Dhokale, Bouke P van Eijck, Mark R J Elsegood, Dzmitry Firaha, Wenbo Fu, Kaori Fukuzawa, Joseph Glover, Hitoshi Goto, Chandler Greenwell, Rui Guo, Jürgen Harter, Julian Helfferich, Detlef W M Hofmann, Johannes Hoja, John Hone, Richard Hong, Geoffrey Hutchison, Yasuhiro Ikabata, Olexandr Isayev, Ommair Ishaque, Varsha Jain, Yingdi Jin, Aling Jing, Erin R Johnson, Ian Jones, K V Jovan Jose, Elena A Kabova,

Adam Keates, Paul F Kelly, Dmitry Khakimov, Stefanos Konstantinopoulos, Liudmila N Kuleshova, He Li, Xiaolu Lin, Alexander List, Congcong Liu, Yifei Michelle Liu, Zenghui Liu, Zhi-Pan Liu, Joseph W Lubach, Noa Marom, Alexander A Maryewski, Hiroyuki Matsui, Alessandra Mattei, R Alex Mayo, John W Melkumov, Sharmarke Mohamed, Zahrasadat Momenzadeh Abardeh, Hari S Muddana, Naofumi Nakayama, Kamal Singh Nayal, Marcus A Neumann, Rahul Nikhar, Shigeaki Obata, Dana O'Connor, Artem R Oganov, Koji Okuwaki, Alberto Otero de-la Roza, Constantinos C Pantelides, Sean Parkin, Chris J Pickard, Luca Pilia, Tatyana Pivina, Rafał Podeszwa, Alastair J A Price, Louise S Price, Sarah L Price, Michael R Probert, Angeles Pulido, Gunjan Rajendra Ramteke, Atta Ur Rehman, Susan M Reutzel-Edens, Jutta Rogal, Marta J Ross, Adrian F Rumson, Ghazala Sadiq, Zeinab M Saeed, Alireza Salimi, Matteo Salvalaglio, Leticia de Almada, Kiran Sasikumar, Sivakumar Sekharan, Cheng Shang, Kenneth Shankland, Kotaro Shinohara, Baimei Shi, Xuekun Shi, A Geoffrey Skillman, Hongxing Song, Nina Strasser, Jacco van de Streek, Isaac J Sugden, Guangxu Sun, Krzysztof Szalewicz, Benjamin I Tan, Lu Tan, Frank Tarczynski, Christopher R Taylor, Alexandre Tkatchenko, Rithwik Tom, Mark E Tuckerman, Yohei Utsumi, Leslie Vogt-Maranto, Jake Weatherston, Luke J Wilkinson, Robert D Willacy, Lukasz Wojtas, Grahame R Woollam, Zhuocen Yang, Etsuo Yonemochi, Xin Yue, Qun Zeng, Yizu Zhang, Tian Zhou, Yunfei Zhou, Roman Zubatyuk, and Jason C Cole. The seventh blind test of crystal structure prediction: structure generation methods. *Acta Crystallographica Section B*, 80, 12 2024.

[66] Lily M Hunnisett, Nicholas Francia, Jonas Nyman, Nathan S Abraham, Srinivasulu Aitipamula, Tamador Alkhidir, Mubarak Almehairbi, Andrea Anelli, Dylan M Anstine, John E Anthony, Joseph E Arnold, Faezeh Bahrami, Michael A Bellucci, Gregory J O Beran, Rajni M Bhardwaj, Raffaello Bianco, Joanna A Bis, A Daniel Boese, James Bramley, Doris E Braun, Patrick W V Butler, Joseph Cadden, Stephen Carino, Ctirad Červinka, Eric J Chan, Chao Chang, Sarah M Clarke, Simon J Coles, Cameron J Cook, Richard I

Cooper, Tom Darden, Graeme M Day, Wenda Deng, Hanno Dietrich, Antonio DiPasquale, Bhausaheb Dhokale, Bouke P van Eijck, Mark R J Elsegood, Dzmitry Firaha, Wenbo Fu, Kaori Fukuzawa, Nikolaos Galanakis, Hitoshi Goto, Chandler Greenwell, Rui Guo, Jürgen Harter, Julian Helfferich, Johannes Hoja, John Hone, Richard Hong, Michal Hušák, Yasuhiro Ikabata, Olexandr Isayev, Ommair Ishaque, Varsha Jain, Yingdi Jin, Aling Jing, Erin R Johnson, Ian Jones, K V Jovan Jose, Elena A Kabova, Adam Keates, Paul F Kelly, JiříKlimeš, Veronika Kostková, He Li, Xiaolu Lin, Alexander List, Congcong Liu, Yifei Michelle Liu, Zenghui Liu, Ivor Lončarić, Joseph W Lubach, Jan Lud´, Alexander A Maryewski, Noa Marom, Hiroyuki Matsui, Alessandra Mattei, R Alex Mayo, John W Melkumov, Bruno Mladineo, Sharmarke Mohamed, Zahrasadat Momenzadeh Abardeh, Hari S Muddana, Naofumi Nakayama, Kamal Singh Nayal, Marcus A Neumann, Rahul Nikhar, Shigeaki Obata, Dana O'Connor, Artem R Oganov, Koji Okuwaki, Alberto Otero de-la Roza, Sean Parkin, Antonio Parunov, Rafał Podeszwa, Alastair J A Price, Louise S Price, Sarah L Price, Michael R Probert, Angeles Pulido, Gunjan Rajendra Ramteke, Atta Ur Rehman, Susan M Reutzel-Edens, Jutta Rogal, Marta J Ross, Adrian F Rumson, Ghazala Sadiq, Zeinab M Saeed, Alireza Salimi, Kiran Sasikumar, Sivakumar Sekharan, Kenneth Shankland, Baimei Shi, Xuekun Shi, Kotaro Shinohara, A Geoffrey Skillman, Hongxing Song, Nina Strasser, Jacco van de Streek, Isaac J Sugden, Guangxu Sun, Krzysztof Szalewicz, Lu Tan, Kehan Tang, Frank Tarczynski, Christopher R Taylor, Alexandre Tkatchenko, Petr Touš, Mark E Tuckerman, Pablo A Unzueta, Yohei Utsumi, Leslie Vogt-Maranto, Jake Weatherston, Luke J Wilkinson, Robert D Willacy, Lukasz Wojtas, Grahame R Woollam, Yi Yang, Zhuocen Yang, Etsuo Yonemochi, Xin Yue, Qun Zeng, Tian Zhou, Yunfei Zhou, Roman Zubatyuk, and Jason C Cole. The seventh blind test of crystal structure prediction: structure ranking methods. *Acta Crystallographica Section B*, 80, 12 2024.

[67] Evangelos A Coutsias, Chaok Seok, and Ken A Dill. Using quaternions to calculate rmsd. *Journal of Computational Chemistry*, 25:1849–1857, 11 2004.

[68] Greg Landrum. Rdkit, open source cheminformatics, 2022.

[69] Ryu Yamasaki, Aya Tanatani, Isao Azumaya, Hyuma Masu, Kentaro Yamaguchi, and Hiroyuki Kagechika. Solvent-dependent conformational switching of n-phenylhydroxamic acid and its application in crystal engineering. *Crystal Growth & Design*, 6:2007–2010, 9 2006. doi: 10.1021/cg060151z.

[70] Valentina V Sobornova, Konstantin Belov, Mikhail Krest'yaninov, and Ilya Khodov. Influence of solvent polarity on the conformer ratio of bicalutamide in saturated solutions: Insights from noesy nmr analysis and quantum-chemical calculations. *International Journal of Molecular Sciences*, 25:8254, 7 2024.

[71] Daniel R Vega, Griselda Polla, Andrea Martinez, Elsa Mendioroz, and María Reinoso. Conformational polymorphism in bicalutamide. *International Journal of Pharmaceutics*, 328, 2007.

[72] Shoji Maruyama, Hiroshi Ooshima, and Jyoji Kato. Crystal structures and solvent-mediated transformation of taltireline polymorphs. *Chemical Engineering Journal*, 75:193–200, 1999.

[73] Caiqin Yang, Zhenwei Zhang, Yanli Zeng, Jing Wang, Yongli Wang, and Baoqing Ma. Structures and characterization of m-nisoldipine polymorphs. *CrystEngComm*, 14:2589–2594, 2012.

[74] Edgar Olehnovics, Yifei Michelle Liu, Nada Mehio, Ahmad Y Sheikh, Michael R Shirts, and Matteo Salvalaglio. Assessing the accuracy and efficiency of free energy differences obtained from reweighted flow-based probabilistic generative models. *Journal of Chemical Theory and Computation*, 20:5913–5922, 7 2024. doi: 10.1021/acs.jctc.4c00520.

[75] Edgar Olehnovics, Yifei Michelle Liu, Nada Mehio, Ahmad Y Sheikh, Michael R Shirts, and Matteo Salvalaglio. Accurate lattice free energies of packing polymorphs from probabilistic generative models. *Jour-*

nal of Chemical Theory and Computation, 21:2244–2255, 3 2025. doi: 10.1021/acs.jctc.4c01612.

[76] Xinqiang Ding and Bin Zhang. Computing absolute free energy with deep generative models. *The Journal of Physical Chemistry B*, 124:10166–10172, 11 2020. doi: 10.1021/acs.jpcb.0c08645.

Appendix A

Supplementary Materials for Chapter 4: Exploring Conformations in the Gas Phase

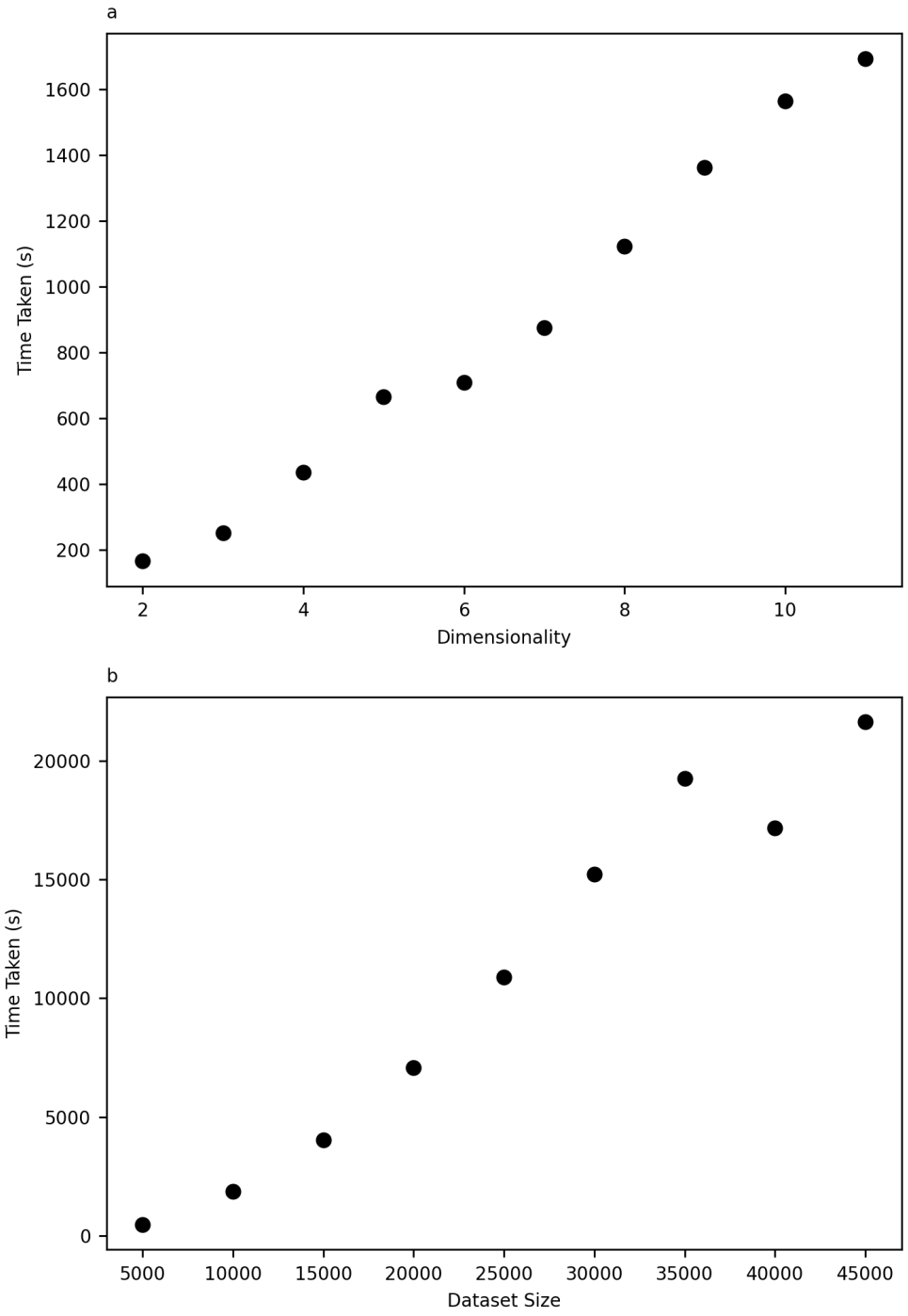


Figure A.1: Computational cost of the analysis procedure as it increases with (a): the dimensionality of conformation space, on a dataset with a constant size of 10,000 molecular configurations, and (b): the size of 11-dimensional configurational datasets. Cost is shown as the time taken to carry out the analysis on a standard desktop workstation. The cost of the simulation used to generate the configurations is not shown. Configurations of Target XXXII are used in both cases. As expected, the cost increases linearly with dimensionality, and with the square of dataset size.

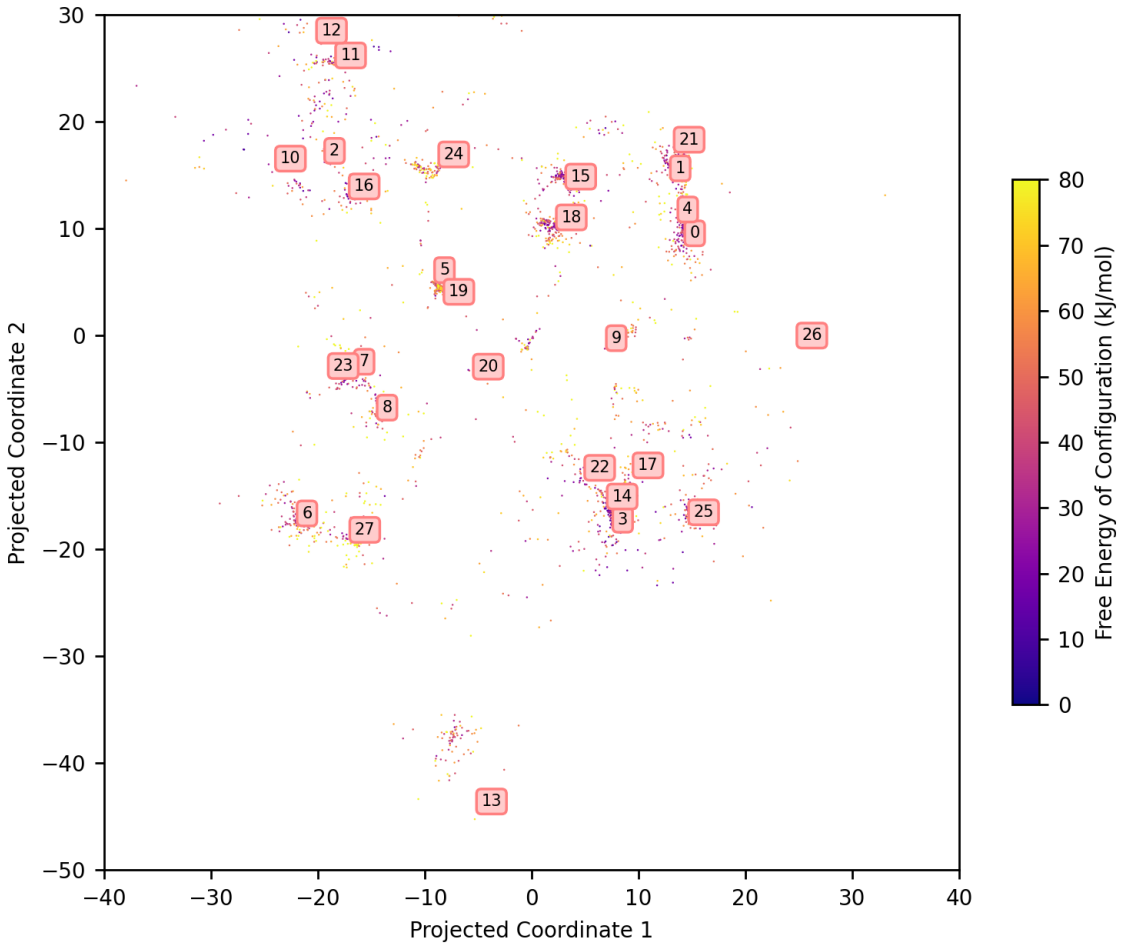


Figure A.2: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 5000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

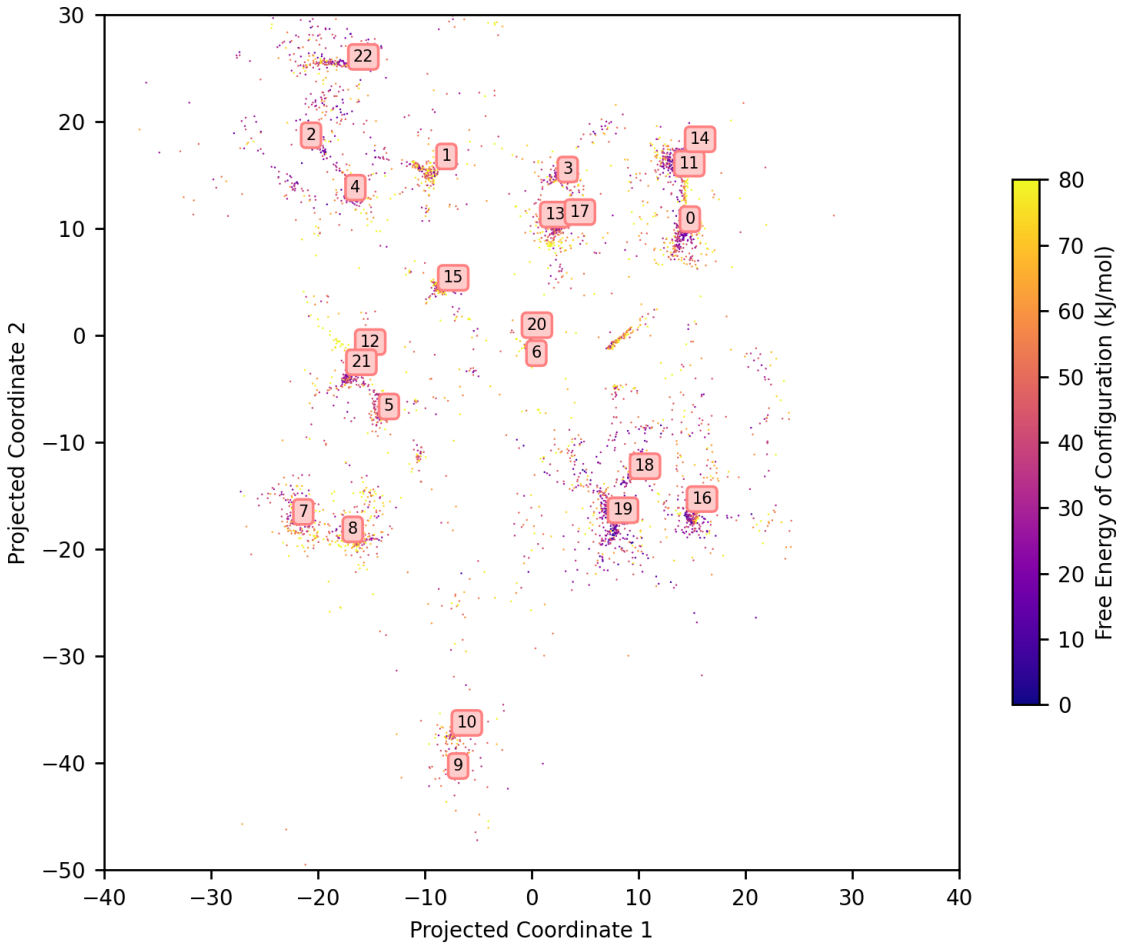


Figure A.3: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 10000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

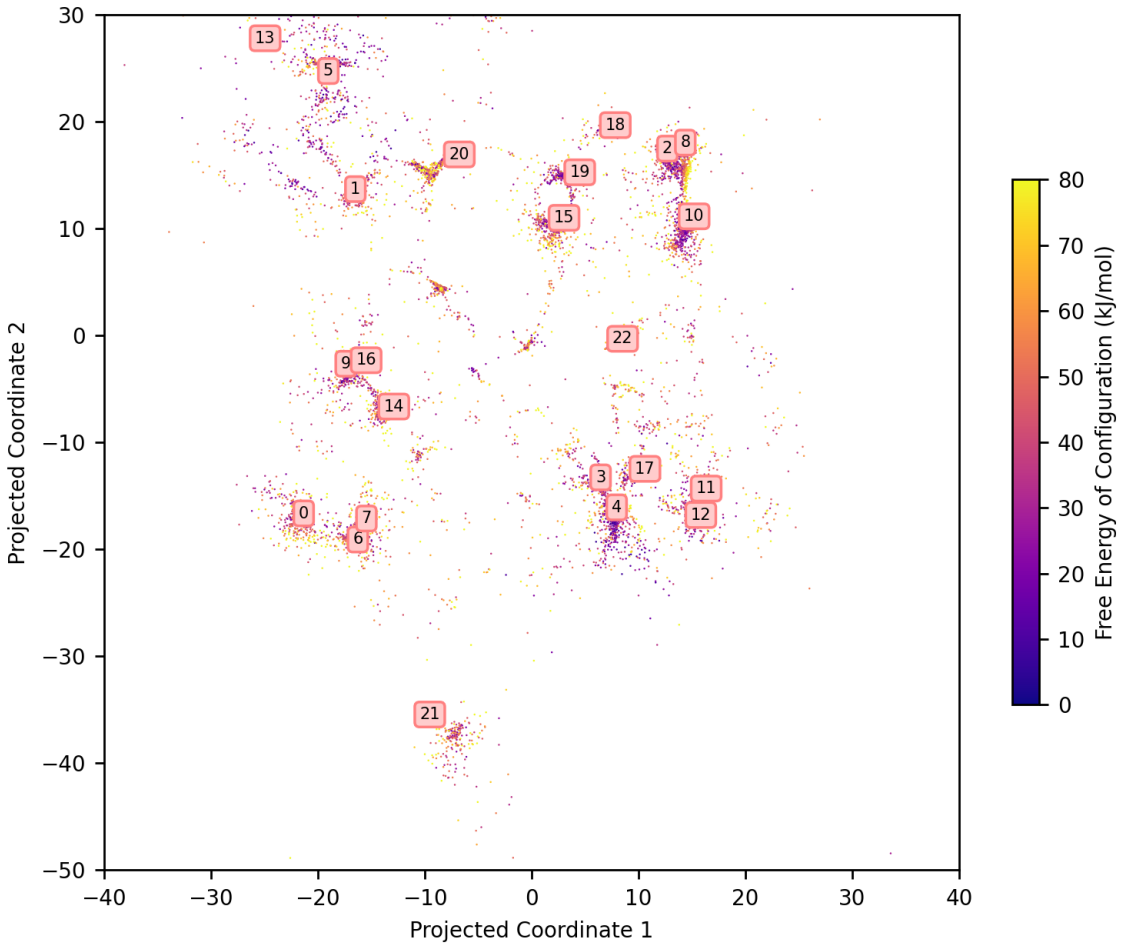


Figure A.4: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 15000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

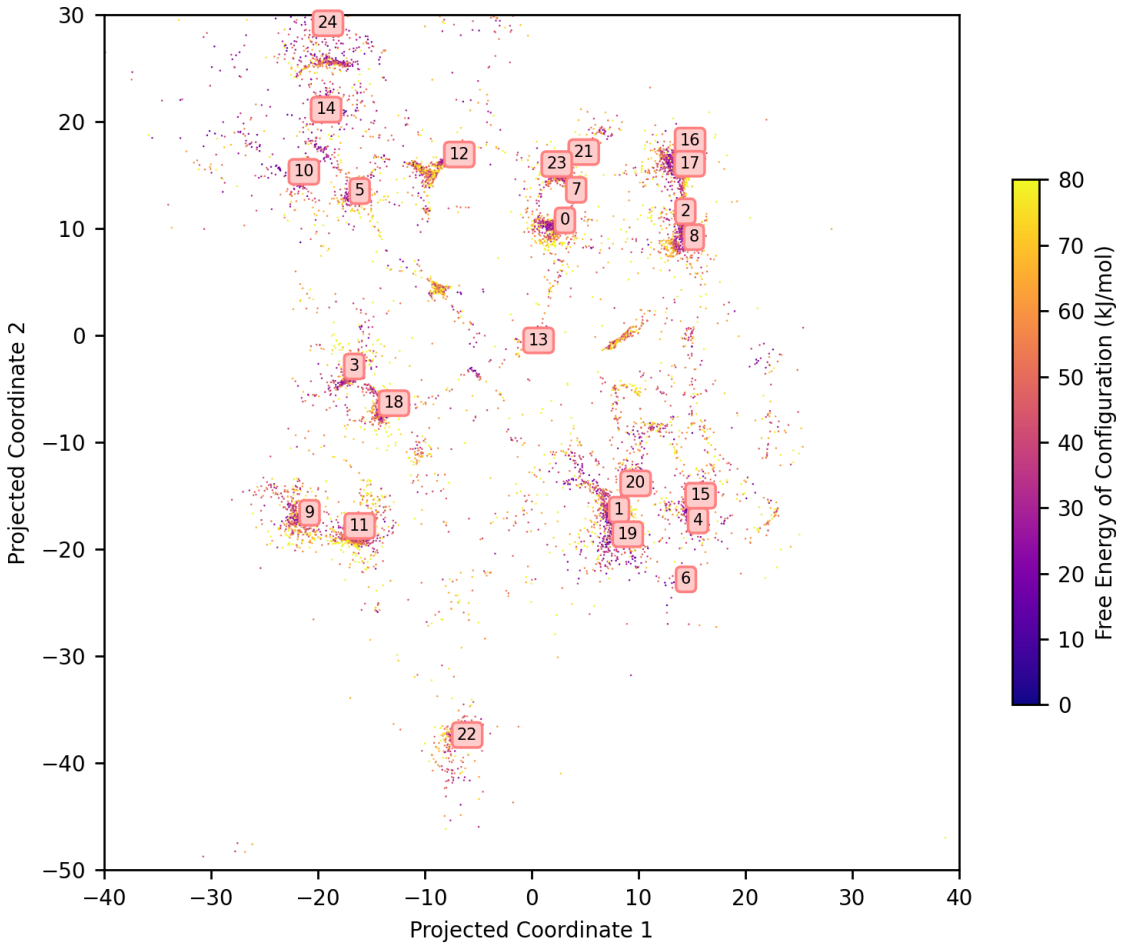


Figure A.5: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 20000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

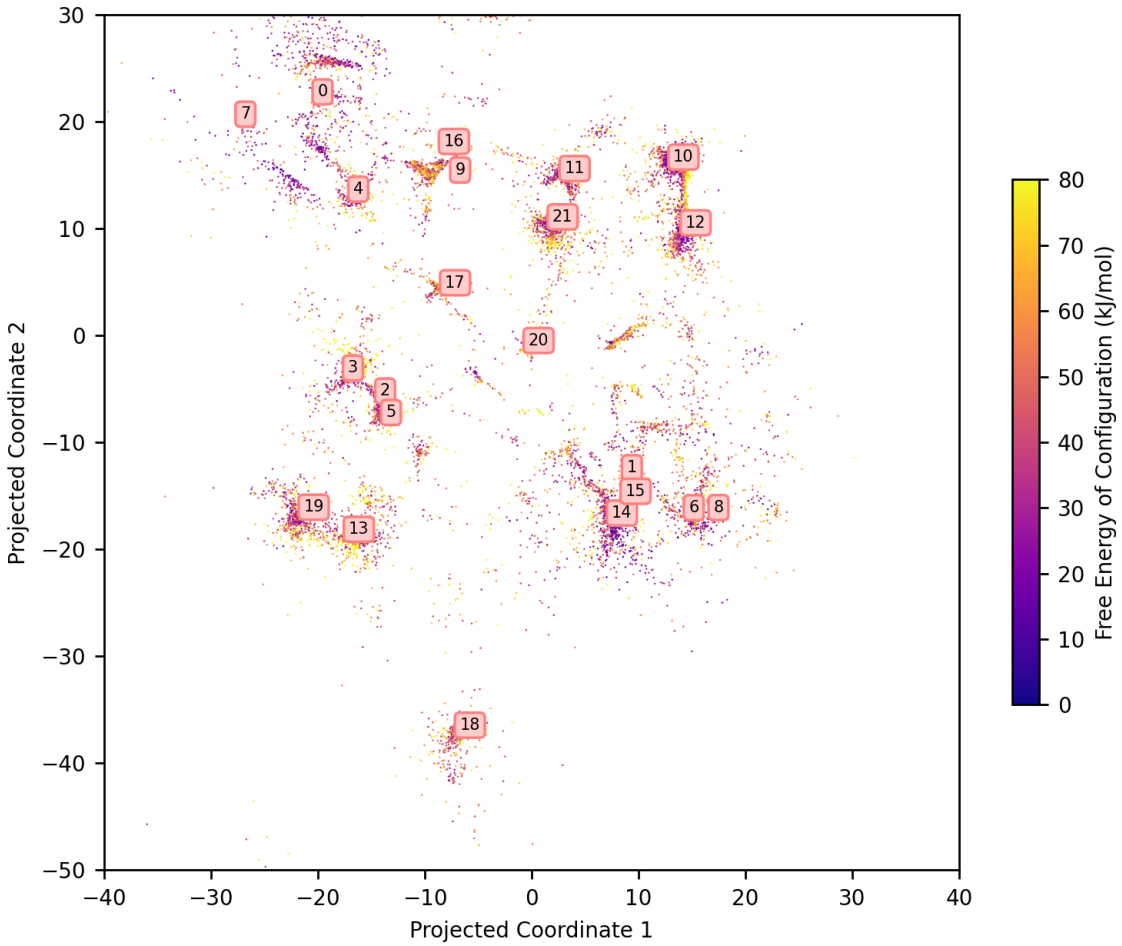


Figure A.6: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 25000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

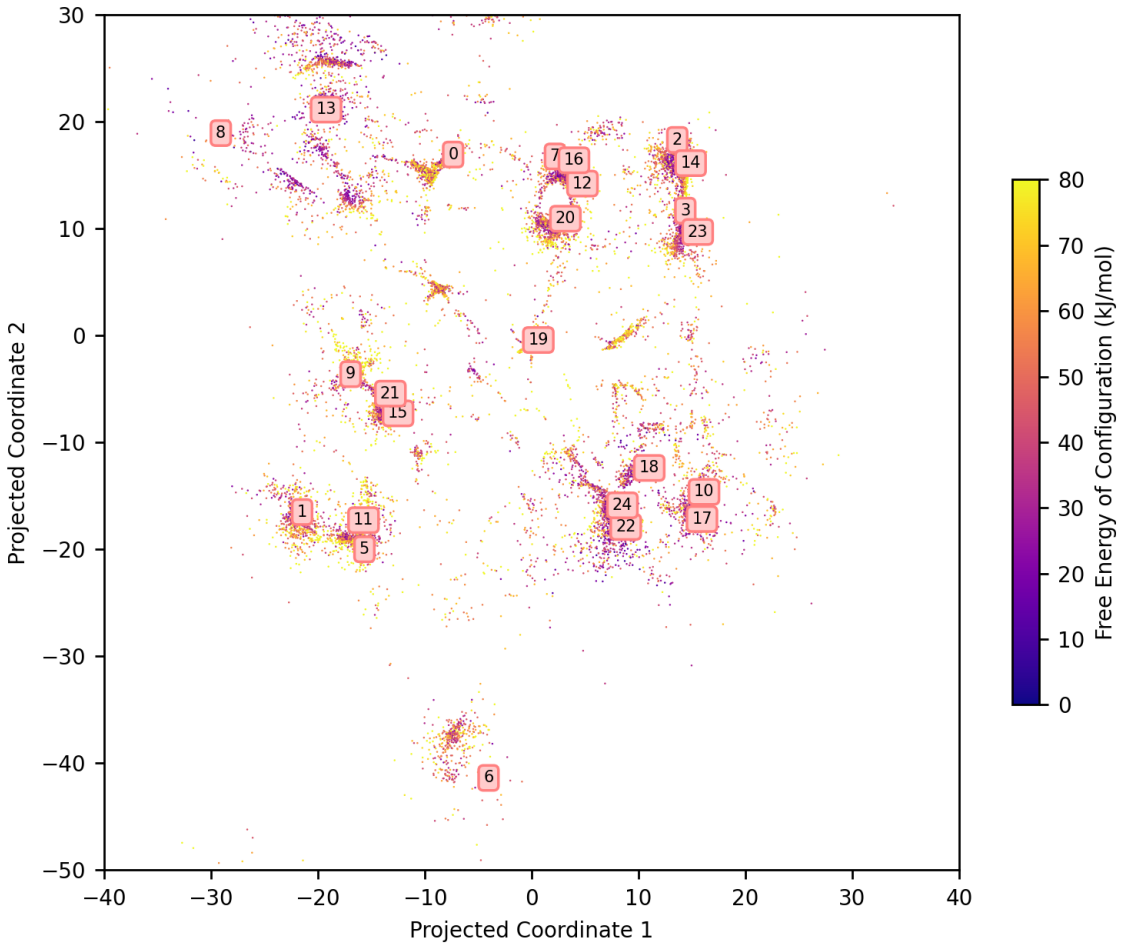


Figure A.7: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 30000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

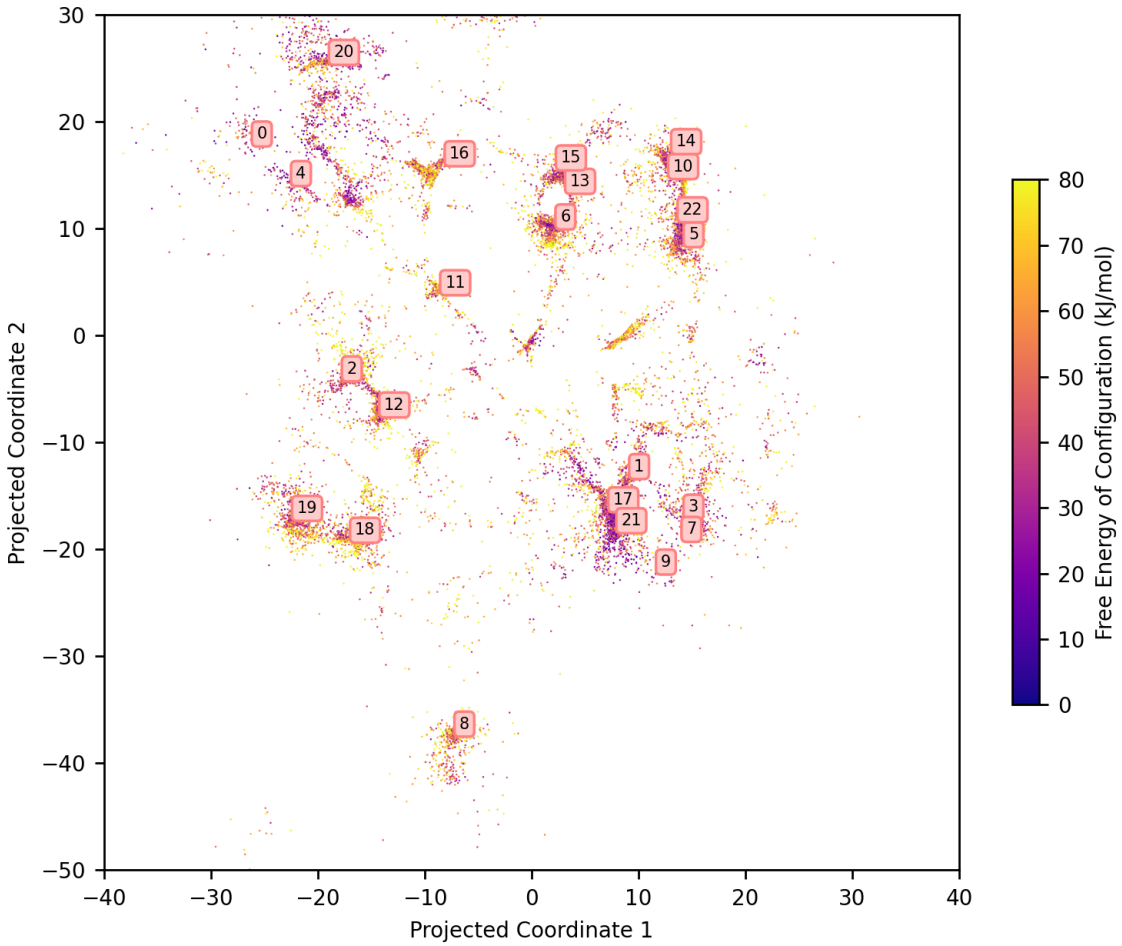


Figure A.8: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 35000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

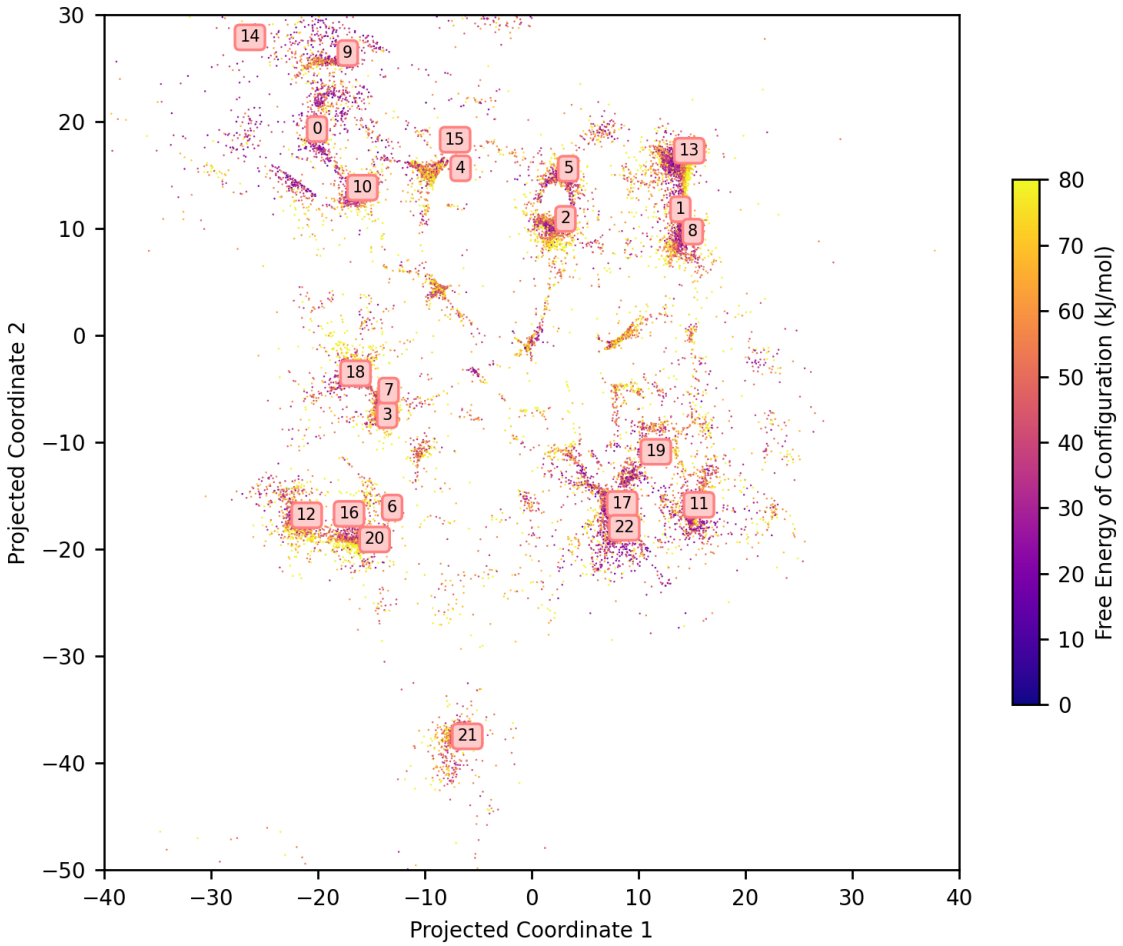


Figure A.9: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 45000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

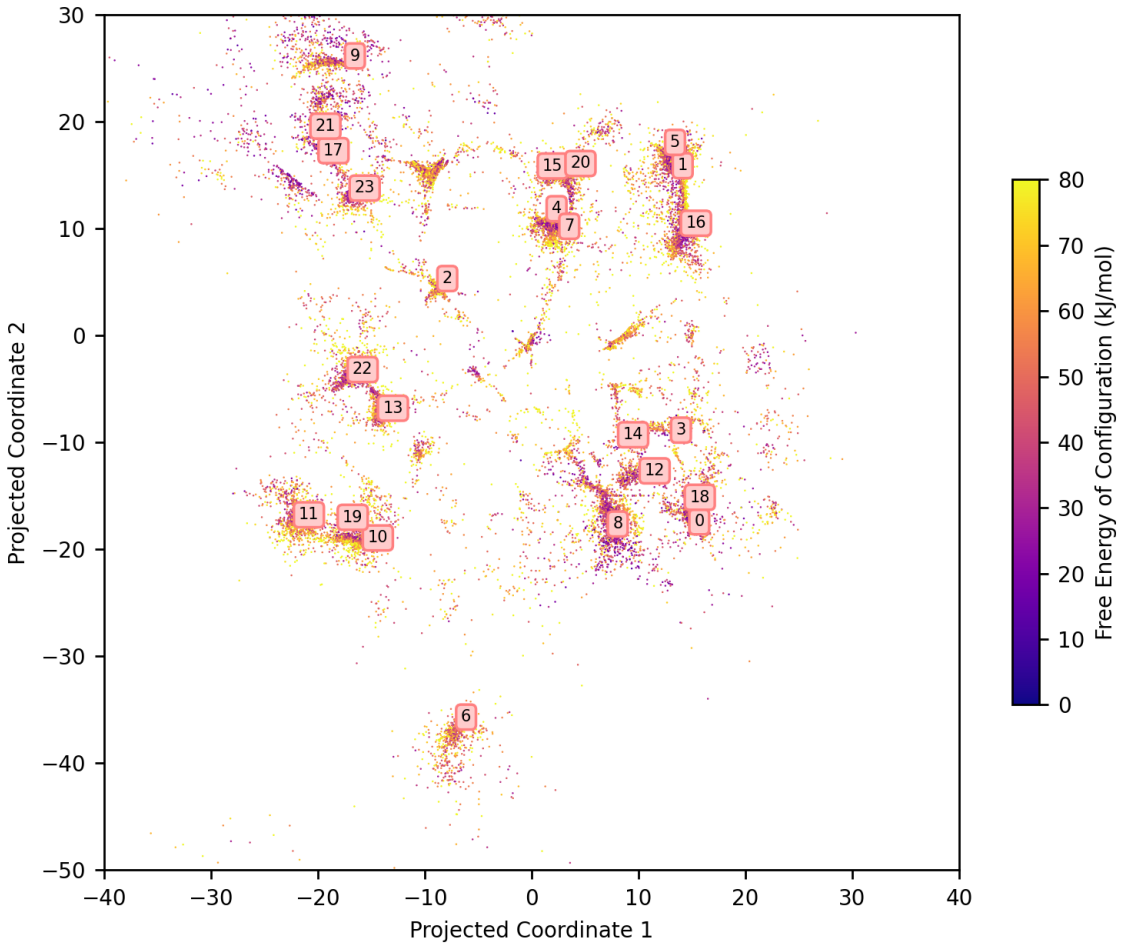


Figure A.10: 2D Sketch-map projection of the 11D per-point FES of Target XXXII generated from a dataset of 45000 configurations. Shown here alongside other projections of Target XXXII's FES generated with smaller datasets, to illustrate the evolution of features of the FES as the size of the dataset increases.

Appendix B

Supplementary Materials for Chapter 5: Exploring the Impact of Solvent on the Conformational Landscapes of Pharmaceutical Molecules

Table B.1: Labels, free energies and CV-space coordinates of PBH's conformers in vacuum. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3
0	0.51	-0.64	-0.28	2.25
1	6.65	-0.70	3.06	0.93
2	6.41	0.80	-3.07	2.24
3	0.79	2.57	-0.29	2.29
4	0.00	-2.55	0.30	0.86
5	7.33	-0.80	3.08	-2.38
6	0.44	-0.68	-0.26	-0.89
7	7.07	-2.43	-3.12	-0.84
8	7.82	2.53	3.03	1.00
9	6.86	-2.43	-3.12	2.18
10	0.21	-2.50	0.23	-2.28
11	7.32	2.51	3.03	-2.29
12	0.66	2.54	-0.27	-1.01
13	0.40	0.55	0.38	0.79
14	6.29	0.78	-3.10	-0.85
15	0.00	0.66	0.25	-2.27

Table B.2: Labels, free energies and CV-space coordinates of PBH's conformers in dichloromethane. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3
0	3.71	-2.52	-3.03	2.24
1	1.29	2.54	-0.20	-0.96
2	2.94	-0.82	-3.10	-2.22
3	0.54	-2.56	0.26	0.97
4	0.25	0.67	0.15	1.06
5	0.43	0.65	0.20	-2.14
6	2.83	-2.50	-3.04	-0.94
7	0.86	2.46	-0.19	2.26
8	2.68	2.50	-3.14	0.95
9	3.40	0.78	-3.10	-0.97
10	0.00	-0.60	-0.27	-0.89
11	1.08	-0.68	-0.17	2.15
12	2.36	2.36	3.07	-2.21
13	1.21	-2.45	0.13	-2.20
14	4.35	-0.83	3.10	1.01
15	3.66	0.75	-3.12	2.28

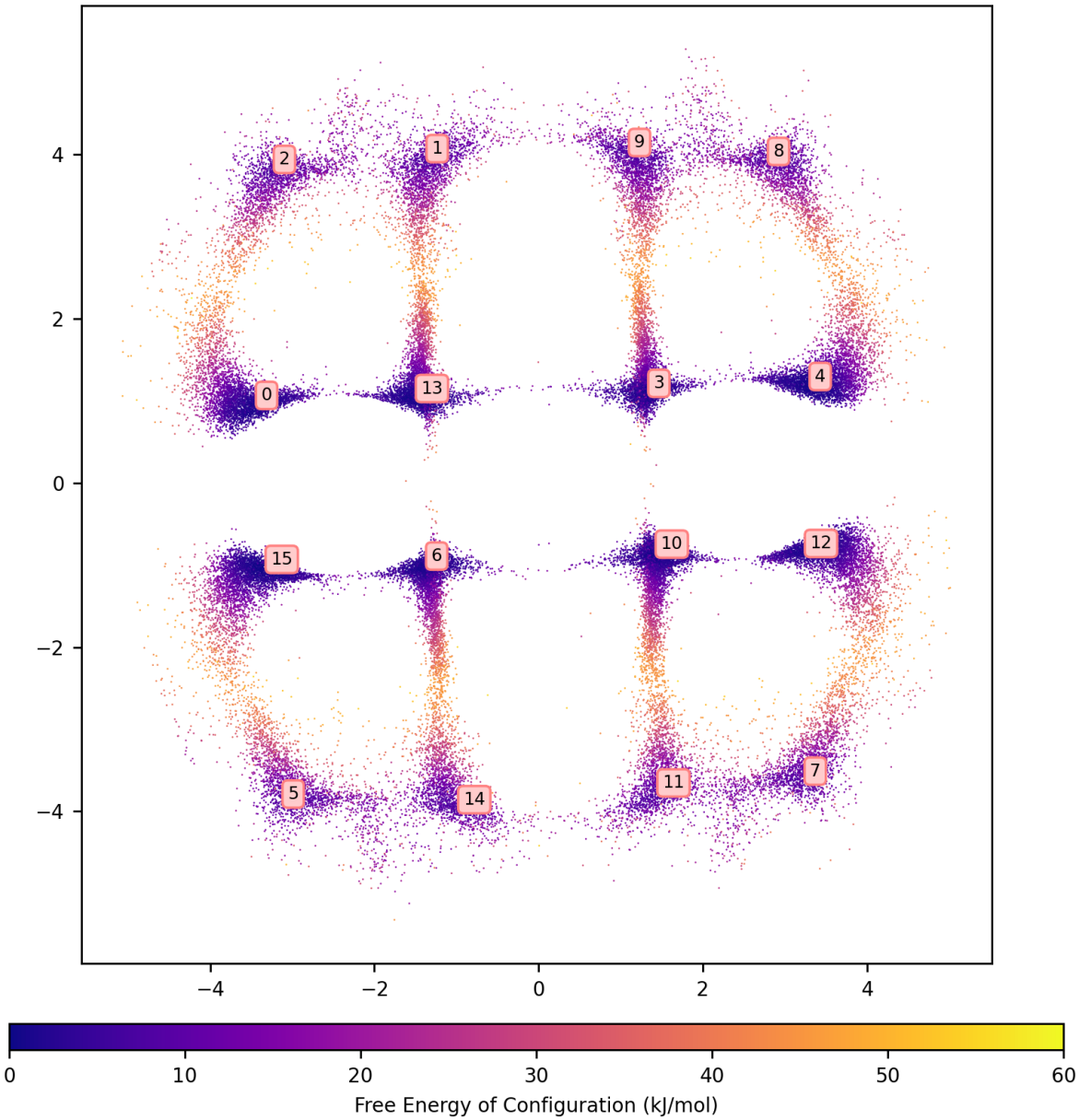


Figure B.1: 2D Sketch-map projection of PBH's 3D conformational free energy landscape in vacuum, with molecular structure of PBH inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

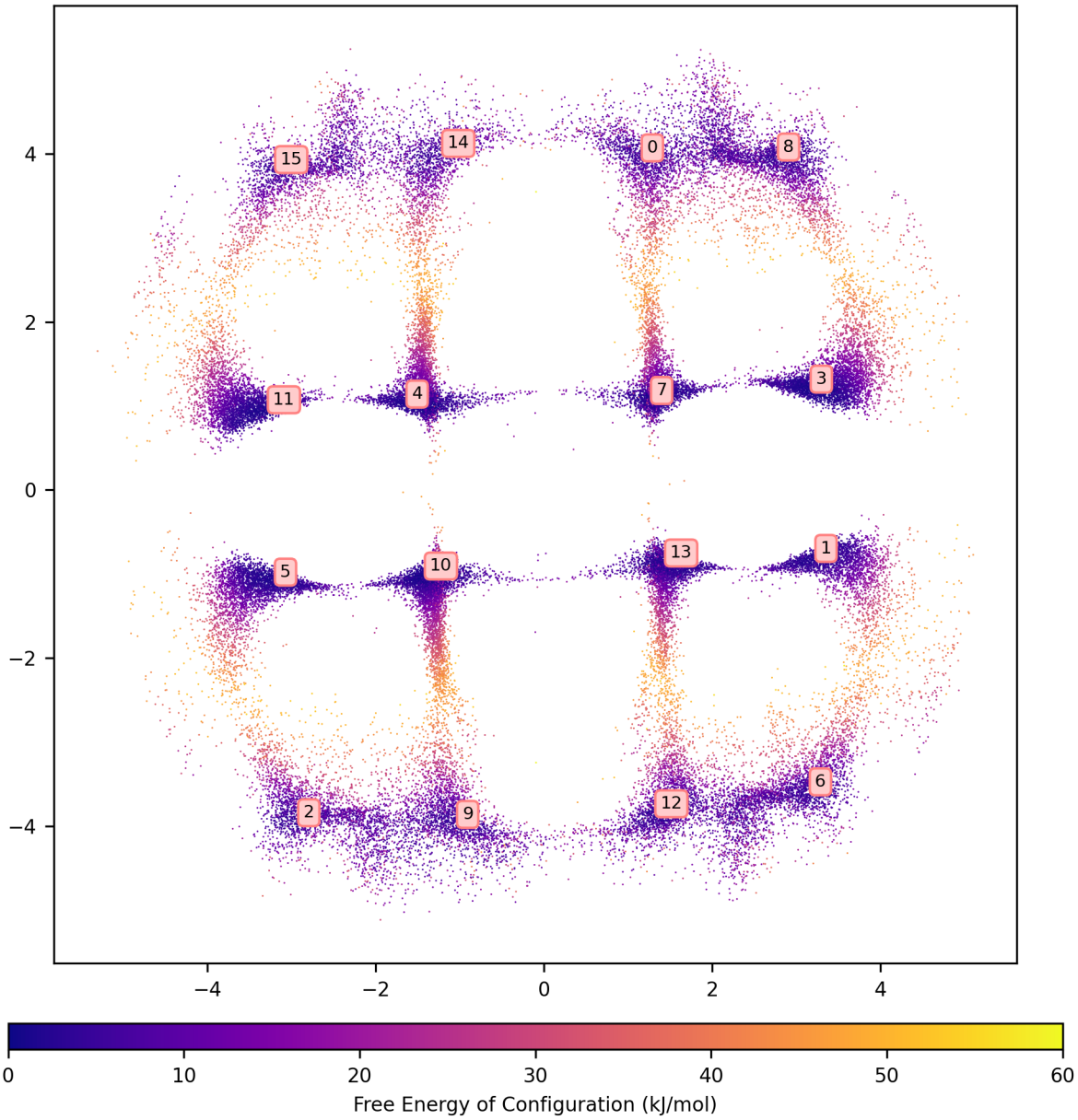


Figure B.2: 2D Sketch-map projection of PBH's 3D conformational free energy landscape in dichloromethane. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

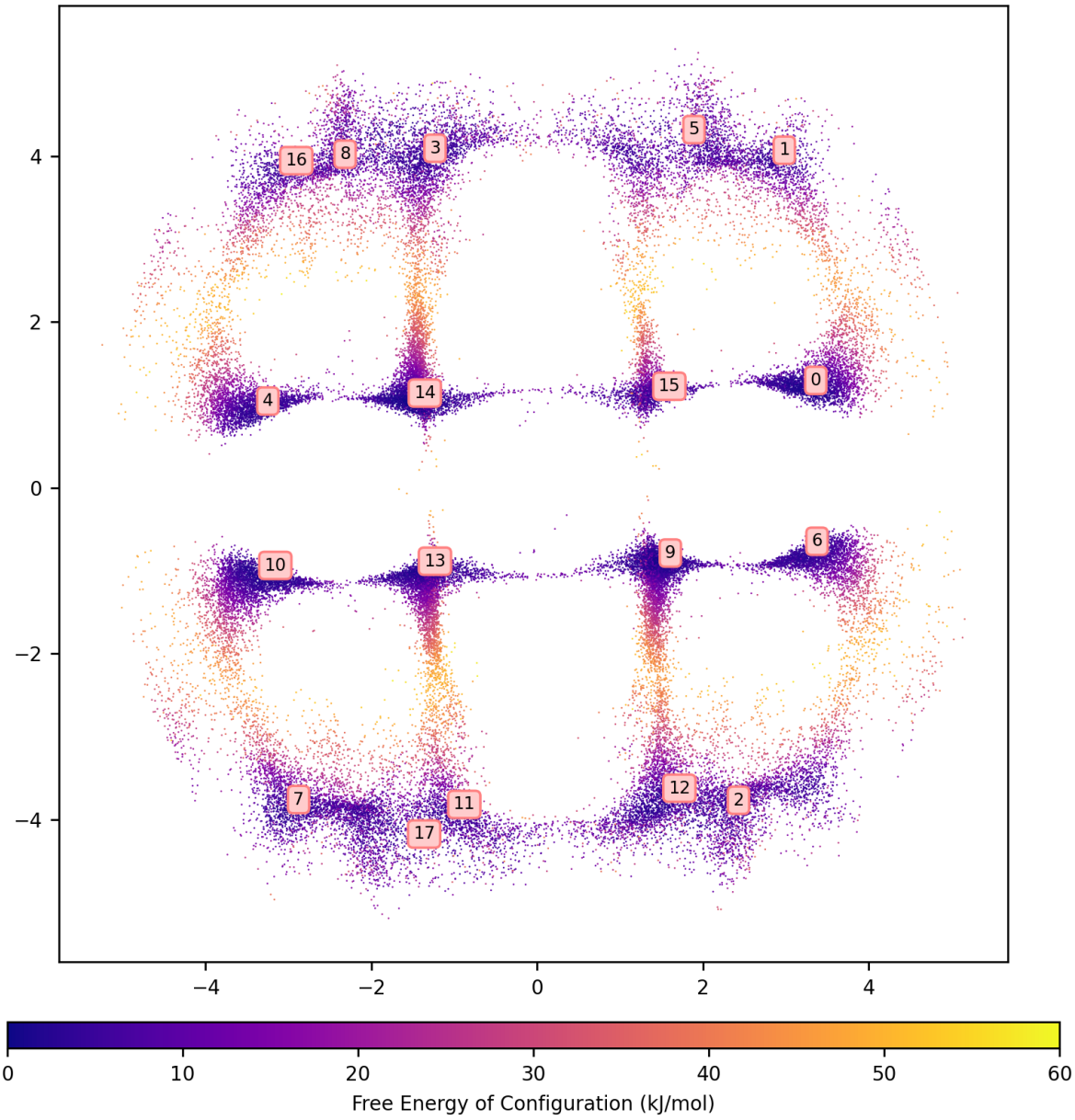


Figure B.3: 2D Sketch-map projection of PBH's 3D conformational free energy landscape in acetone. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

Table B.3: Labels, free energies and CV-space coordinates of PBH's conformers in acetone. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3
0	0.99	-2.55	0.26	0.91
1	3.28	2.43	3.04	0.97
2	3.76	-2.40	-3.08	-2.28
3	2.42	-0.68	3.10	0.81
4	2.76	-0.76	-0.16	2.22
5	4.21	-2.46	-3.14	0.96
6	2.67	2.46	-0.12	-0.92
7	2.10	-0.73	3.08	-2.28
8	4.76	0.66	-3.04	1.08
9	0.22	-2.56	0.27	-2.22
10	0.88	0.65	0.27	-2.28
11	3.29	0.70	-3.06	-0.86
12	2.11	2.57	3.02	-2.23
13	1.53	-0.79	0.02	-1.12
14	0.00	0.65	0.21	0.97
15	3.04	2.52	-0.19	2.17
16	4.30	0.81	-3.06	2.09
17	3.53	0.77	-3.05	-2.28

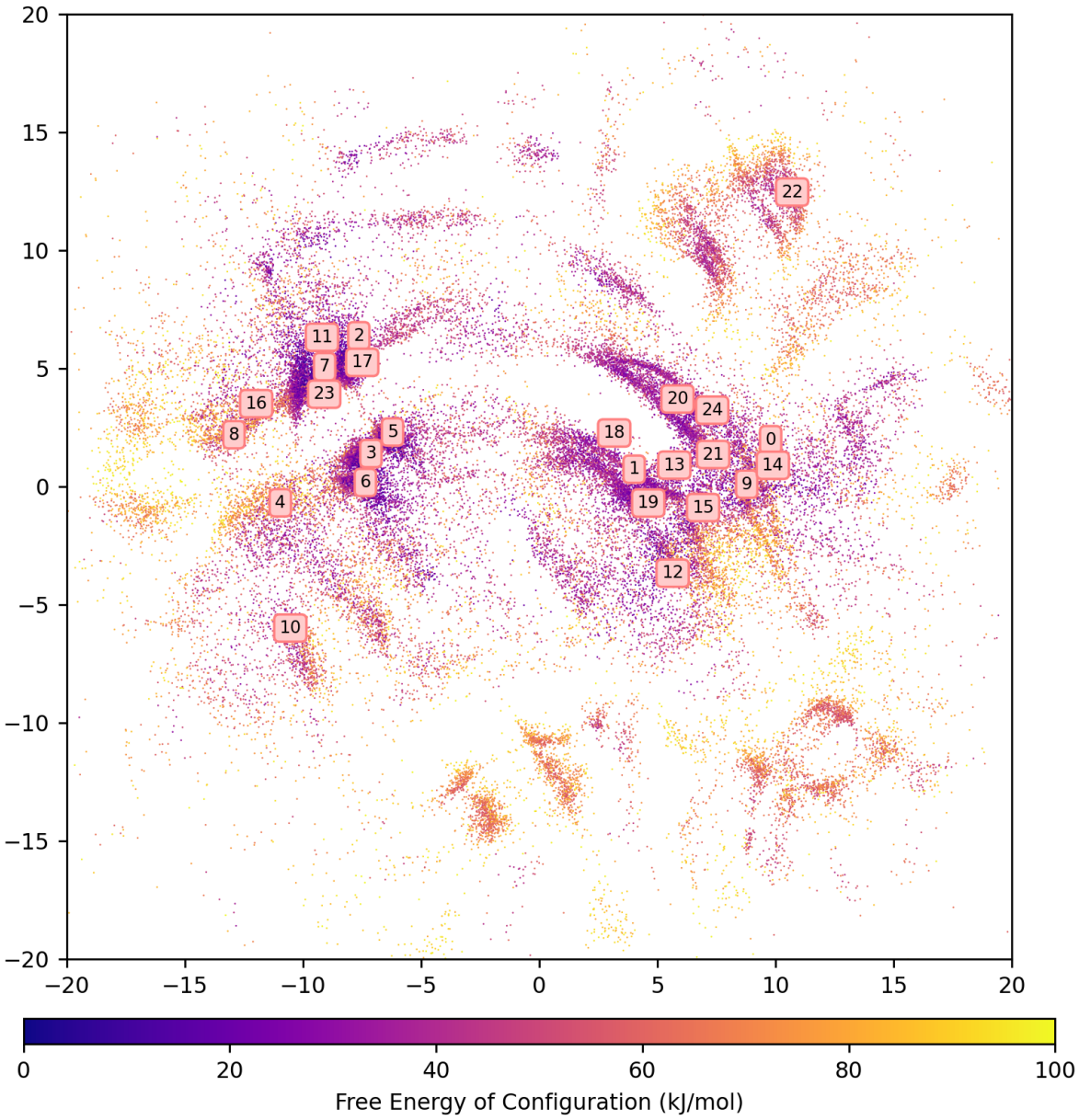


Figure B.4: 2D Sketch-map projection of bicalutamide's 3D conformational free energy landscape in vacuum, with molecular structure of bicalutamide inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

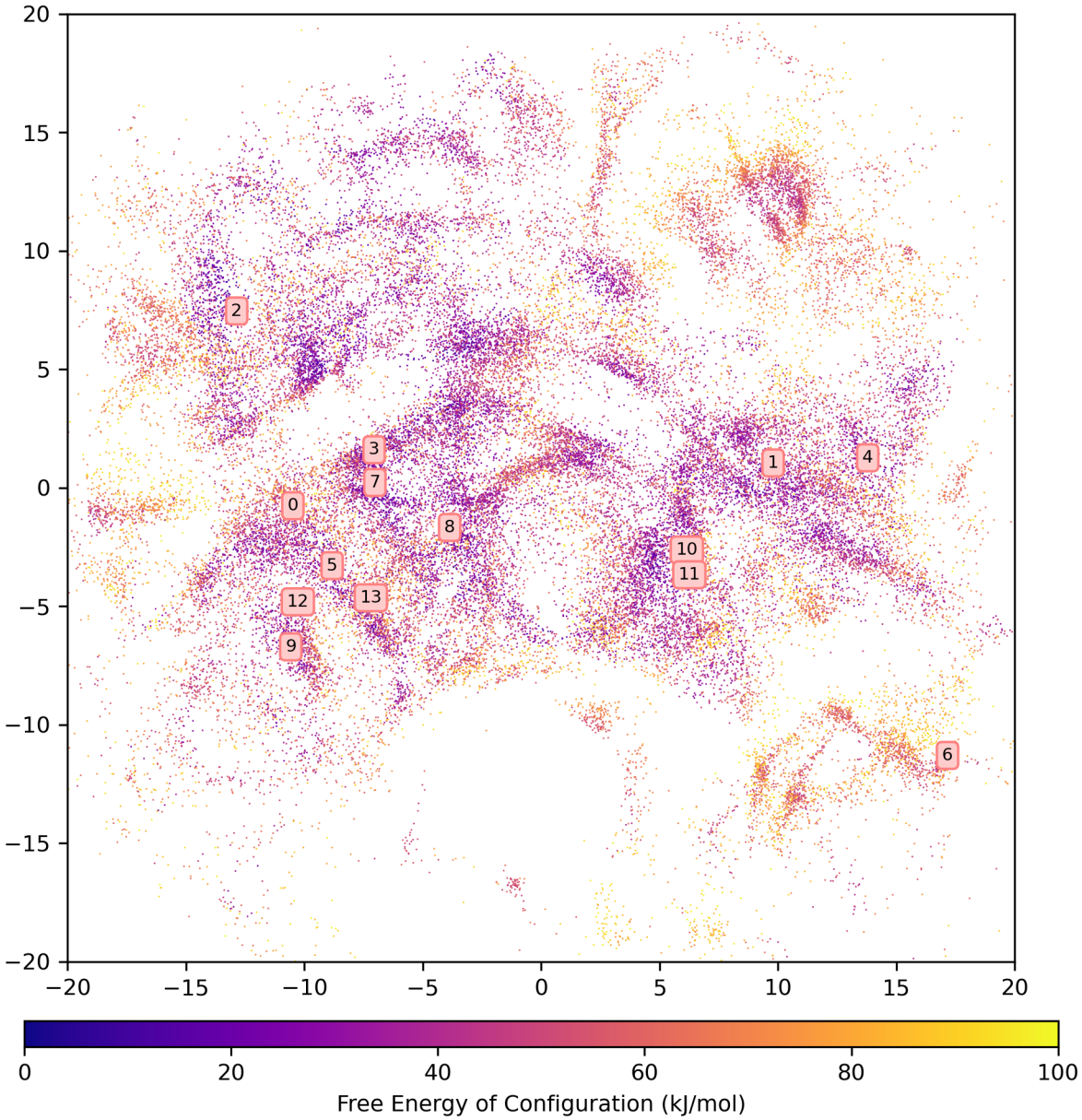


Figure B.5: 2D Sketch-map projection of bicalutamide's 3D conformational free energy landscape in chloroform. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

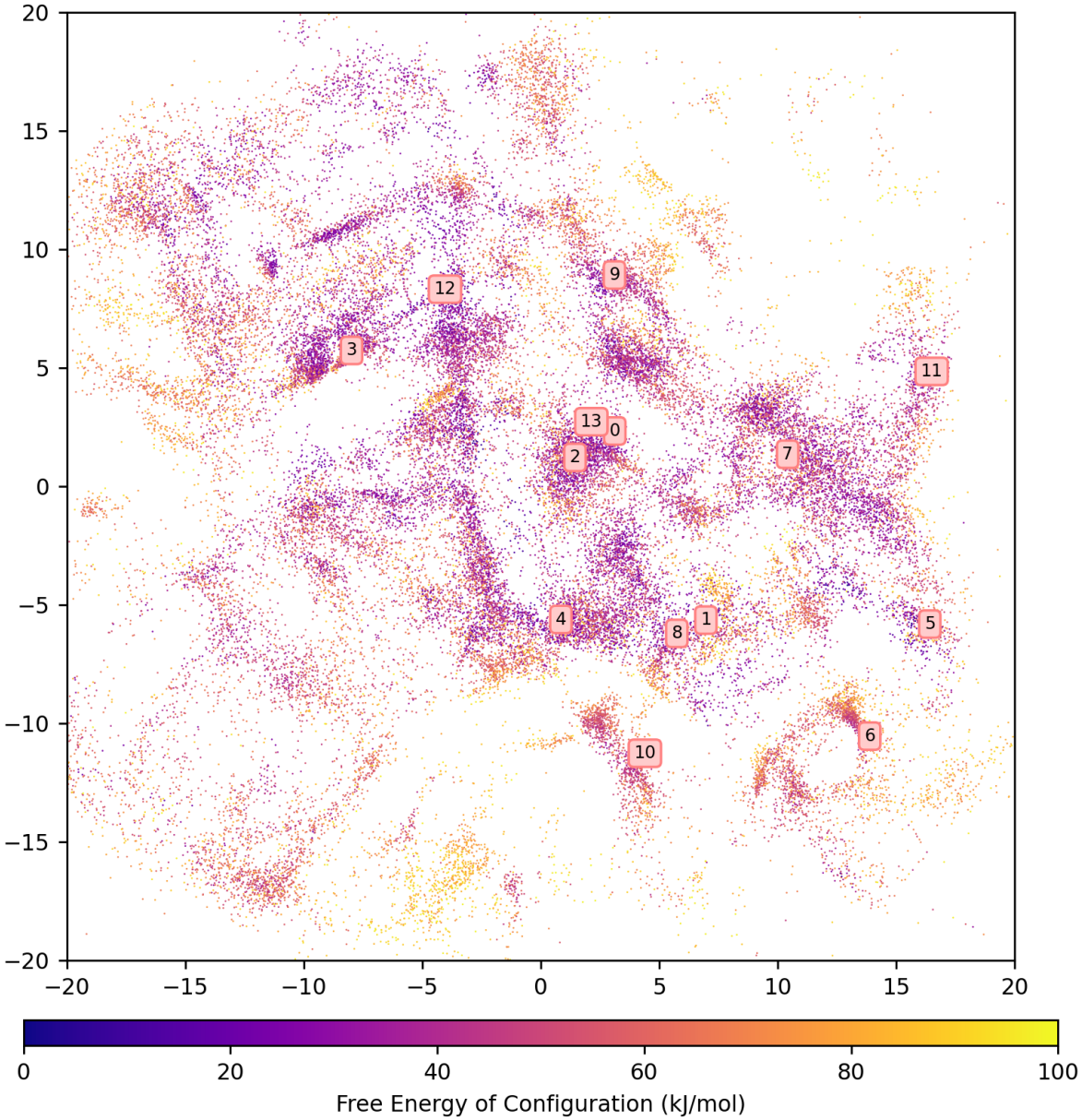


Figure B.6: 2D Sketch-map projection of bicalutamide's 3D conformational free energy landscape in DMSO. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

Table B.4: Labels, free energies and CV-space coordinates of bicalutamide's conformers in vacuum. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
0	8.16	-1.90	-3.06	-3.13	1.13	-1.38	2.96	1.45
1	15.83	-2.00	2.54	-3.06	1.50	-0.64	1.30	-2.36
2	4.09	2.37	-2.47	3.06	-2.23	1.02	-1.11	2.03
3	0.00	-0.02	-2.42	-3.11	-2.16	1.22	-1.23	-1.17
4	19.00	0.04	-2.25	2.91	-2.18	-3.03	-2.82	-1.72
5	0.52	2.11	-2.41	2.79	-2.23	1.23	-1.22	-1.07
6	1.17	-1.84	-2.41	2.86	-2.25	1.23	-1.15	-1.03
7	1.28	0.04	-2.65	2.85	-2.25	1.23	-1.05	1.95
8	21.50	-2.07	-2.64	2.87	-2.30	-3.01	-2.83	1.39
9	14.63	0.14	2.21	-2.92	1.14	-1.58	2.88	1.25
10	23.18	1.83	2.94	3.07	-2.30	-3.05	-2.75	1.39
11	2.23	-2.16	-2.63	3.09	-2.29	1.21	-1.09	2.05
12	12.55	2.16	2.49	-3.03	1.40	-1.08	-2.47	-1.42
13	14.18	0.15	2.58	-2.91	1.72	-0.80	1.53	-2.09
14	11.29	2.12	2.66	-3.02	1.15	-1.46	2.85	1.56
15	11.60	-0.21	2.36	-3.05	1.00	-1.51	2.86	-1.76
16	25.23	-0.15	-2.53	2.85	-2.09	-3.01	-2.51	1.46
17	7.03	1.81	-2.58	2.94	-2.02	1.01	-1.23	2.41
18	15.88	0.04	2.47	-3.13	0.80	0.70	0.78	-2.02
19	14.32	1.91	2.58	3.07	1.58	-0.82	1.40	-2.20
20	17.52	-0.01	2.65	-3.02	1.49	-0.55	1.11	0.74
21	9.81	-2.13	2.68	3.09	1.29	-1.38	2.85	-1.14
22	28.67	0.14	1.68	-0.12	1.36	0.68	0.95	-1.90
23	3.07	-2.08	-2.43	2.77	-2.08	0.99	-1.36	2.08
24	15.28	1.94	2.52	3.11	1.73	-0.82	1.45	0.92

Table B.5: Labels, free energies and CV-space coordinates of bicalutamide's conformers in chloroform. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
0	21.65	-0.46	-2.17	2.77	-2.16	3.09	-3.02	-1.24
1	13.05	-2.12	2.77	2.93	1.45	-0.79	-3.03	1.91
2	2.60	2.39	2.82	2.93	-2.50	1.11	2.95	1.56
3	11.50	0.11	-2.66	2.83	-2.51	1.36	-0.75	-1.33
4	13.22	-2.29	-1.15	-3.12	1.23	-0.57	-3.08	1.80
5	16.22	-1.98	2.64	3.11	-1.97	-2.90	-2.71	-1.96
6	38.40	2.21	-1.34	-0.22	-1.10	-0.80	-0.92	1.76
7	9.66	-1.86	-2.42	-3.00	-2.27	1.24	-0.98	-1.07
8	3.47	-2.21	-2.75	-3.10	-2.70	0.98	2.53	-1.54
9	14.52	1.86	2.69	3.09	-1.96	-3.11	-2.90	1.72
10	13.28	0.26	-2.61	3.05	1.07	-1.64	2.88	-1.73
11	9.77	-2.23	2.54	3.03	1.52	-0.77	-2.79	-1.23
12	15.14	-2.23	-2.82	-2.99	-2.21	-3.10	-2.97	1.48
13	10.59	2.19	2.57	-3.01	-2.40	-2.97	-2.82	-1.59

Table B.6: Labels, free energies and CV-space coordinates of bicalutamide's conformers in DMSO. The labeling convention is consistent with that of Figure 5.7c

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
0	15.03	0.02	2.55	-3.07	0.88	0.89	0.90	-1.75
1	8.15	-0.24	-2.66	-2.90	-1.09	-0.82	-3.02	-1.64
2	16.50	2.21	-2.34	-3.01	1.00	0.83	0.88	-1.85
3	15.63	2.04	-2.59	2.93	-2.46	1.29	-0.92	2.09
4	19.20	0.10	-2.53	-3.08	2.77	-1.14	-1.69	-1.29
5	3.00	2.14	-0.61	3.04	-1.04	-1.02	2.95	1.14
6	30.15	-0.65	-1.47	0.17	-1.44	-0.90	3.08	1.45
7	19.53	-0.01	2.65	2.98	3.08	-1.38	2.57	1.67
8	6.43	-2.25	-2.44	3.13	-0.99	-0.90	-2.95	-1.76
9	13.55	-2.09	-0.93	-3.08	1.16	0.70	0.88	1.34
10	29.69	2.09	1.84	0.27	-1.29	-0.82	3.11	-1.61
11	15.54	0.03	-0.81	-3.07	2.47	-1.16	2.58	-1.27
12	17.20	1.86	-2.77	-3.13	3.11	-1.36	-1.60	2.21
13	13.81	-0.12	-2.58	-3.10	0.98	0.84	0.91	-1.63

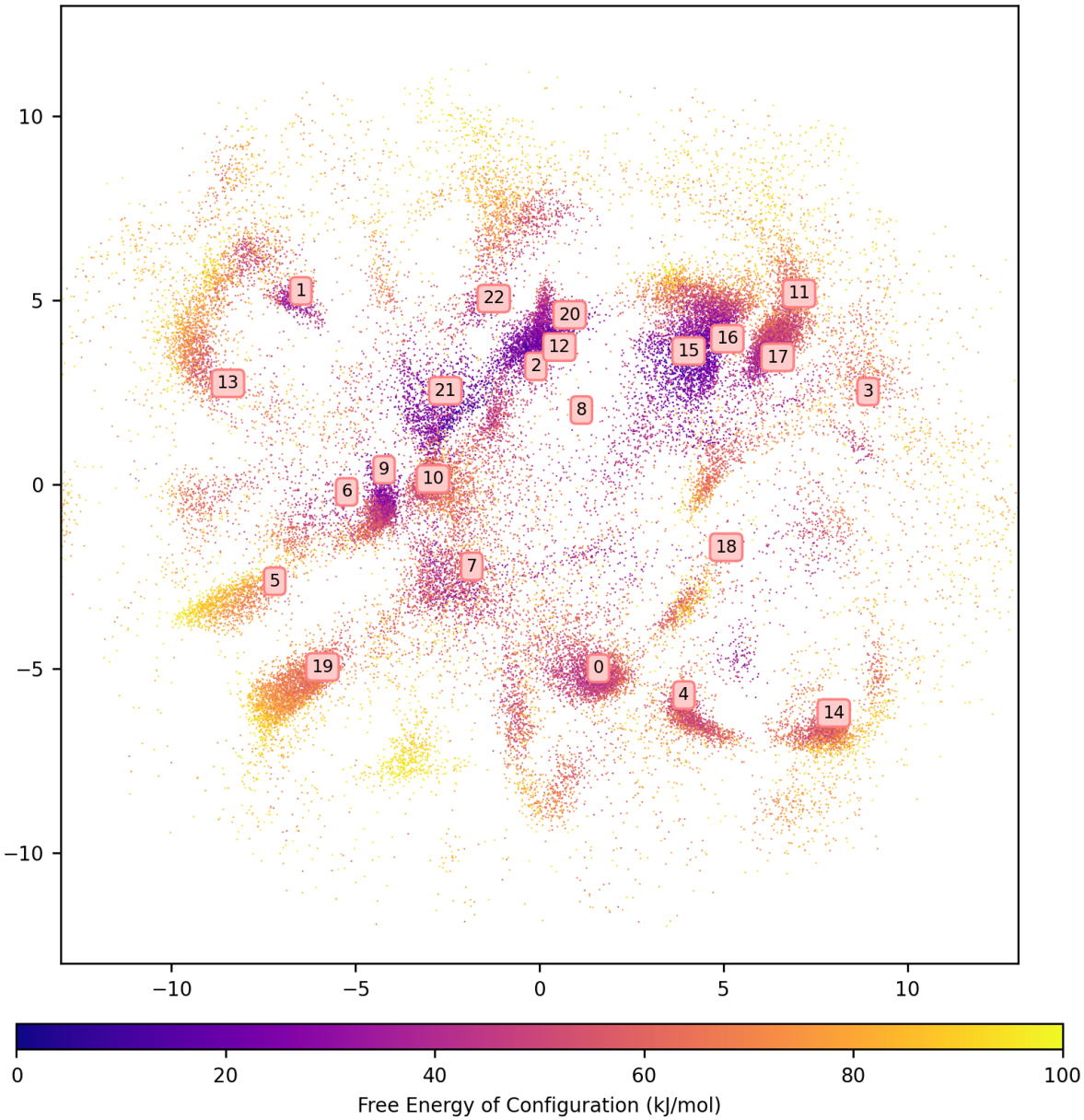


Figure B.7: 2D Sketch-map projection of taltirelin's 3D conformational free energy landscape in vacuum, with molecular structure of taltirelin inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

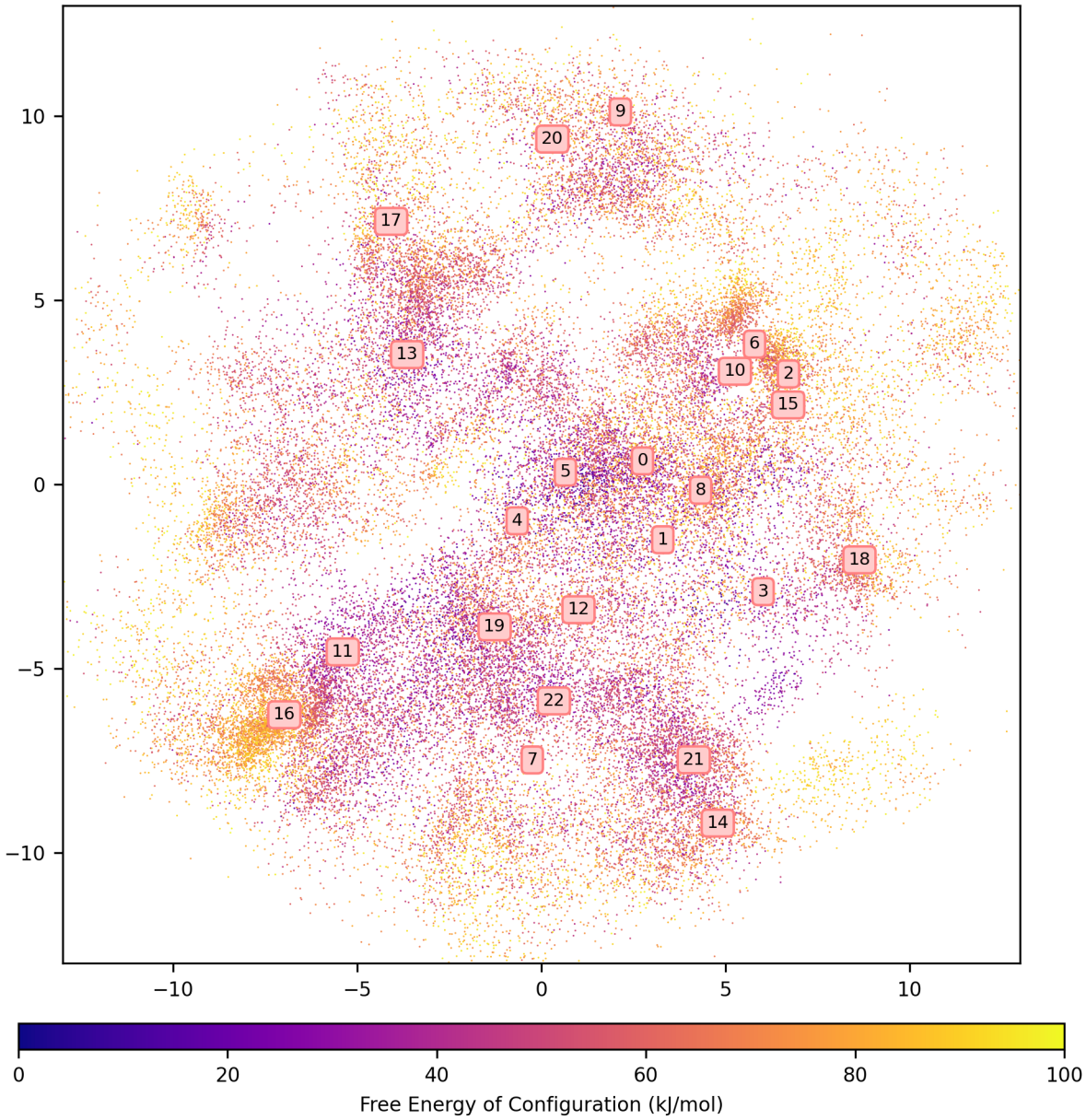


Figure B.8: 2D Sketch-map projection of taltirelin's 3D conformational free energy landscape in water. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

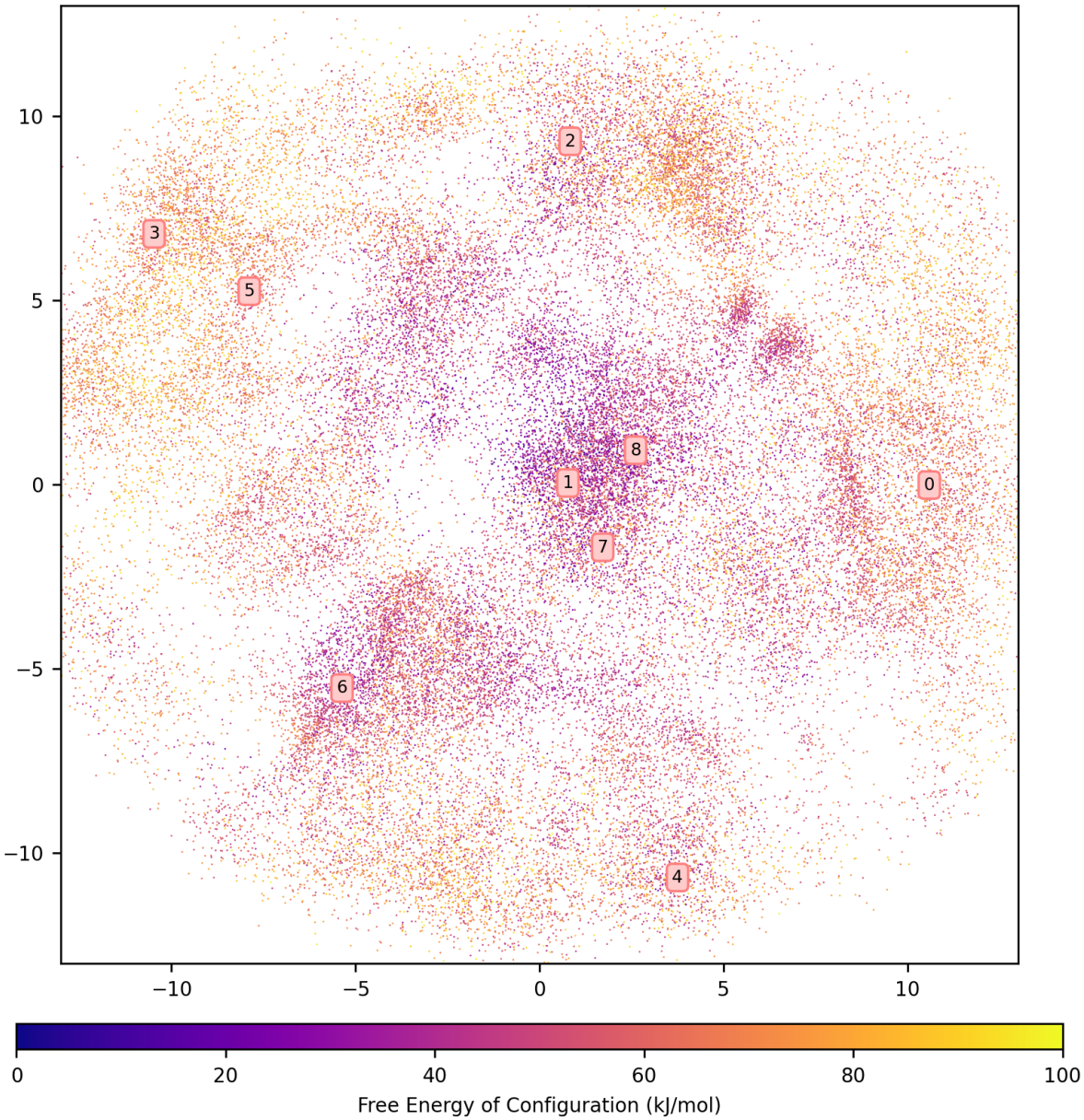


Figure B.9: 2D Sketch-map projection of taltirelin's 3D conformational free energy landscape in a water/methanol mixture. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

Table B.7: Labels, free energies and CV-space coordinates of taltirelin's conformers in vacuum. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
0	32.46	-1.21	2.29	1.84	2.91	-1.72	-2.39	-0.25	-0.09
1	21.95	1.49	0.60	1.71	2.76	-2.88	-1.71	2.75	-1.78
2	17.09	1.60	-2.48	1.29	0.13	1.30	1.83	1.85	-1.08
3	44.69	-1.39	2.09	1.60	-0.81	3.01	-2.97	-0.83	1.66
4	34.87	-0.95	2.38	1.57	2.79	-2.90	1.24	-0.67	0.36
5	55.04	2.84	1.35	-1.66	1.62	-0.86	-0.99	3.03	-2.03
6	32.91	-1.13	2.29	-3.09	2.70	0.93	-2.60	2.66	-1.38
7	19.97	1.61	3.04	3.01	2.61	0.82	-2.24	0.22	-0.44
8	23.30	1.53	-2.31	1.29	2.73	1.37	1.63	0.49	-0.50
9	12.05	1.68	3.01	-3.13	2.65	0.79	-2.52	2.92	-0.87
10	41.79	-0.04	2.99	3.04	-0.22	0.11	-1.64	2.35	-1.51
11	43.48	2.59	-2.25	-0.82	-0.16	-2.73	2.46	-1.48	2.09
12	6.75	2.03	-2.55	1.33	0.86	0.81	1.15	3.02	-1.21
13	45.46	-2.65	-0.53	2.97	2.79	1.09	-2.65	2.56	-1.56
14	43.02	-2.86	-0.47	3.09	2.59	-2.76	1.68	-0.43	1.42
15	6.94	2.19	-2.69	1.30	-0.41	0.65	1.63	-1.01	0.47
16	7.90	2.28	-2.55	0.89	-0.48	0.81	1.92	-1.11	1.39
17	19.45	2.16	-2.34	1.23	-0.80	3.14	-3.09	-0.84	1.74
18	26.16	2.47	-2.14	1.27	2.81	-1.69	-2.52	-0.52	-0.41
19	49.34	2.65	1.17	-1.85	0.99	-0.81	-1.07	0.28	-0.55
20	3.81	1.95	-2.21	1.32	0.46	1.07	1.21	3.01	-1.36
21	13.04	1.83	-2.50	1.68	1.46	-1.07	3.02	2.52	-1.44
22	20.98	1.79	-2.71	1.57	1.54	-2.86	1.71	2.73	-1.34

Table B.8: Labels, free energies and CV-space coordinates of taltirelin's conformers in water. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
0	15.46	2.92	-2.71	3.05	-0.65	-0.77	2.62	-0.66	-1.11
1	5.25	2.51	-2.44	1.53	2.89	-1.58	-3.12	-0.41	-1.51
2	38.44	1.97	-2.87	1.58	-0.72	3.03	-2.18	-0.82	2.03
3	8.78	2.70	-2.49	1.83	2.36	-1.27	-0.77	-0.18	1.64
4	4.88	2.50	-2.70	1.94	1.47	-0.94	-1.14	-0.10	-1.37
5	11.32	2.20	-2.26	2.93	-0.36	-0.95	-1.20	-0.64	-1.72
6	38.58	2.10	-2.36	1.48	-0.56	-2.95	1.69	-0.95	1.90
7	21.89	-1.26	2.75	1.89	2.38	-1.33	1.57	2.85	-1.16
8	5.44	2.26	-2.29	1.77	2.56	0.87	-2.82	0.04	1.78
9	19.55	2.46	-2.37	1.31	2.86	0.83	-2.62	2.53	1.81
10	23.42	2.20	-2.75	1.35	-0.41	1.00	-1.32	-0.98	1.81
11	12.91	2.55	-2.75	-1.23	1.11	-1.11	-1.18	0.10	-1.43
12	23.69	-1.08	2.34	1.93	2.75	1.29	2.29	-0.00	-1.28
13	14.54	2.79	-2.36	1.54	1.49	-1.33	-2.89	3.02	-1.51
14	20.20	-1.44	2.38	-1.34	1.18	-0.97	2.17	0.17	1.54
15	37.92	2.54	-2.76	1.08	-0.53	-1.35	-0.91	-0.87	1.98
16	50.20	3.04	1.17	-1.40	1.01	-0.85	-0.51	0.43	1.60
17	30.14	2.21	-2.75	1.64	2.29	-2.83	-1.63	2.90	2.03
18	18.38	2.93	-2.77	-1.35	1.10	-1.12	2.21	0.38	1.09
19	0.00	2.95	-2.80	1.92	2.63	-0.94	-1.30	-0.03	-1.32
20	14.93	2.47	-2.51	-1.38	1.00	-1.17	1.32	0.36	-1.54
21	23.16	-1.33	1.96	2.94	2.16	3.07	1.02	-0.03	-1.25
22	22.04	-1.60	1.98	3.10	2.36	-2.76	-1.98	-0.25	-1.48

Table B.9: Labels, free energies and CV-space coordinates of taltirelin's conformers in a water/methanol mixture. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
0	38.67	-1.35	-1.06	2.85	-0.37	-2.85	-3.07	-0.85	1.29
1	7.58	1.96	-1.80	3.06	-0.65	-0.81	-1.73	-0.40	-0.98
2	7.03	2.48	-2.33	-1.49	1.06	-0.87	1.82	0.16	-1.26
3	43.27	-3.05	0.05	1.45	2.88	0.97	1.72	2.80	2.25
4	14.73	-1.03	2.60	3.13	-0.77	2.90	1.23	-0.29	-1.30
5	45.28	-1.76	-0.38	1.48	1.60	-2.91	-2.38	2.79	-1.76
6	20.82	-1.32	2.18	-1.20	0.90	-0.92	-0.41	0.20	-1.57
7	17.82	1.57	-2.47	1.35	2.23	-2.95	-1.80	-0.13	-0.52
8	7.37	2.52	-2.34	3.02	-0.33	-1.03	2.59	-0.78	-1.02

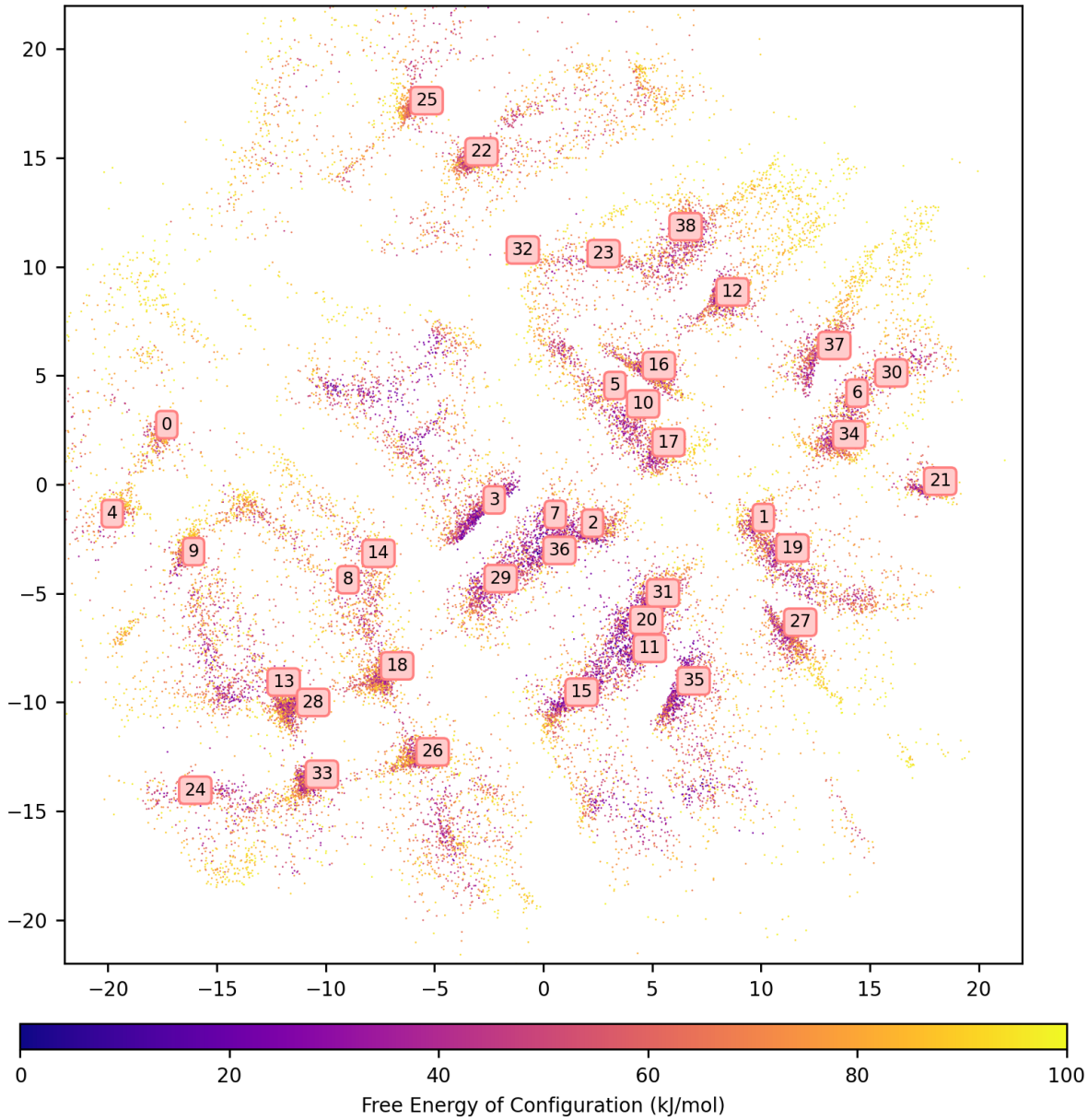


Figure B.10: 2D Sketch-map projection of m-nisoldipine's 3D conformational free energy landscape in vacuum, with molecular structure of m-nisoldipine inset. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

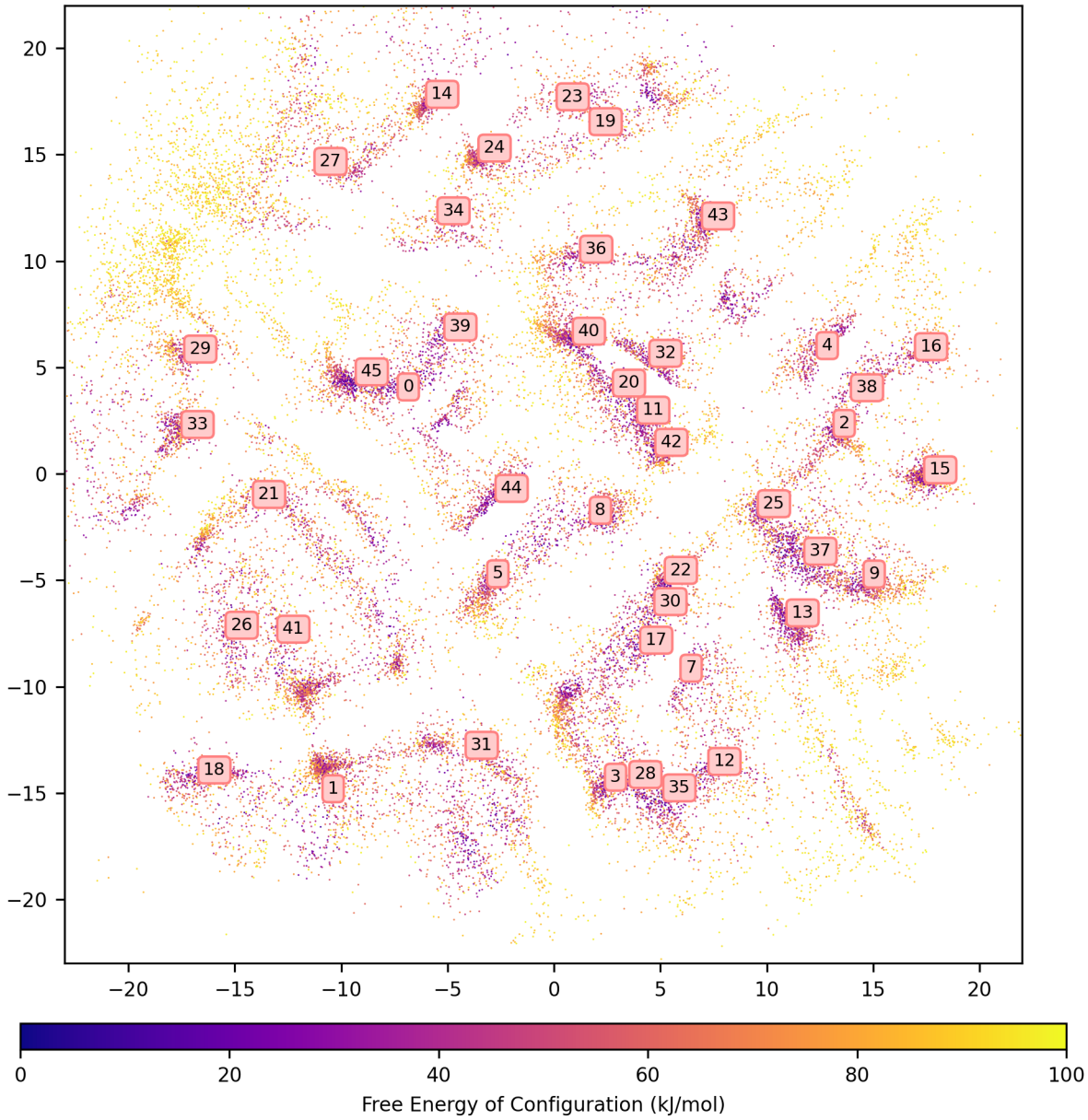


Figure B.11: 2D Sketch-map projection of m-nisoldipine's 3D conformational free energy landscape in an acetone/ethanol mixture. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

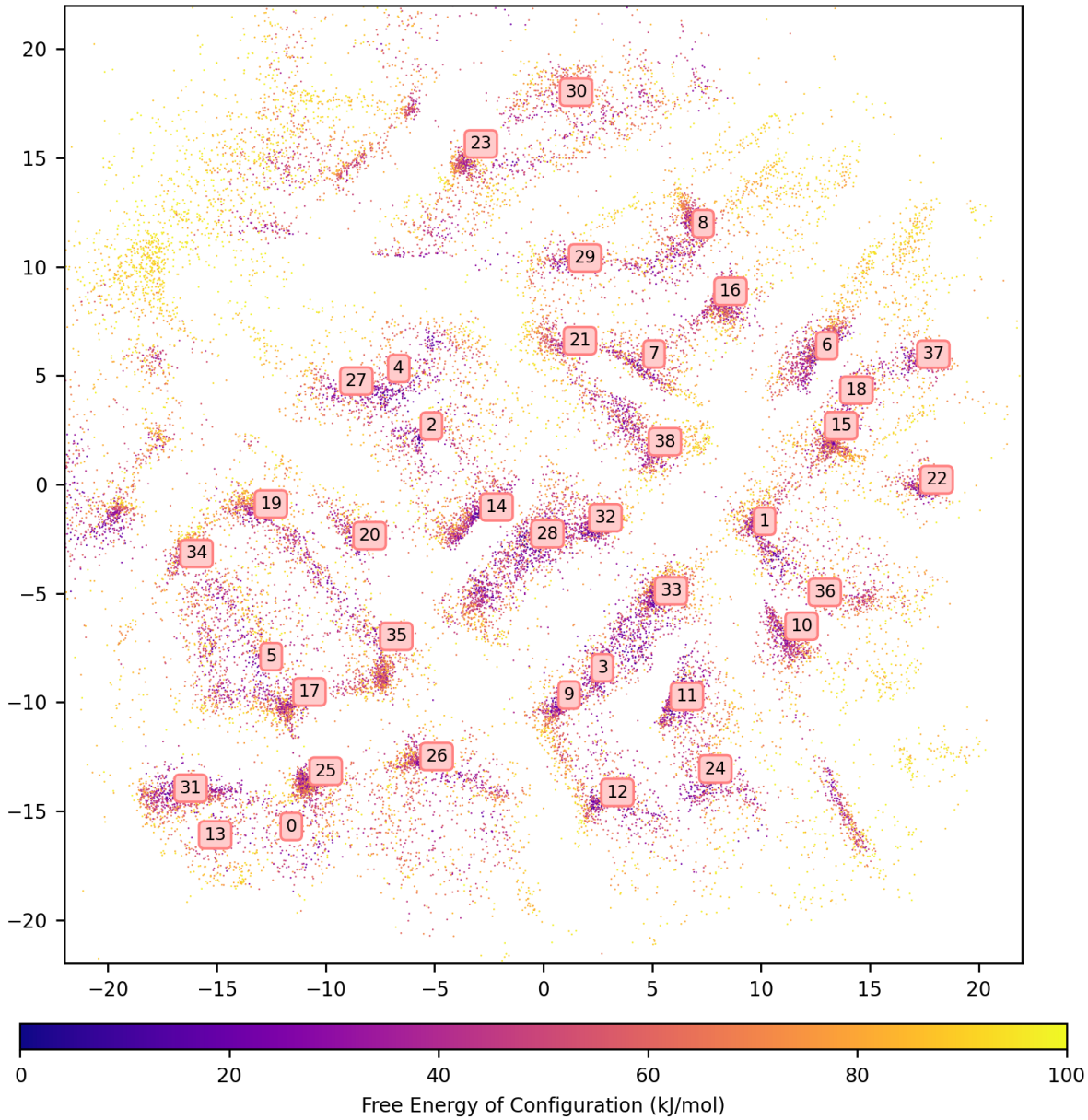


Figure B.12: 2D Sketch-map projection of m-nisoldipine's 3D conformational free energy landscape in an ethyl acetate/hexane mixture. Distances between configurations are preserved over small separations but the axes themselves have no physical meaning.

Table B.10: Labels, free energies and CV-space coordinates of m-nisoldipine's conformers in vacuum. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
0	21.47	0.07	0.78	-0.02	-2.96	-0.30	-3.00	-1.29	2.81
1	19.56	0.25	1.49	3.08	3.03	3.10	-3.07	-1.30	-3.04
2	4.93	-3.09	1.20	0.14	-3.07	3.01	3.05	-1.41	2.81
3	5.06	-3.11	1.31	0.17	-3.08	-3.12	-2.98	-2.82	-1.03
4	20.19	-0.45	-2.10	0.23	-3.13	-0.14	3.06	-1.44	-3.01
5	21.24	2.89	1.46	-2.58	3.06	-3.08	-3.14	-3.07	0.95
6	16.01	-0.33	-1.73	-3.10	-2.94	-3.08	2.84	3.12	2.86
7	9.00	2.96	1.38	0.23	-2.96	-2.85	-3.12	-2.37	1.24
8	27.17	-3.12	0.73	-0.06	2.99	-0.04	-2.90	3.06	3.00
9	20.71	-2.92	-1.84	0.03	3.06	0.21	3.07	-1.38	3.10
10	18.01	2.67	1.35	-3.01	2.89	3.13	-3.06	2.94	3.09
11	4.37	0.27	1.07	-0.03	-3.02	-2.89	-3.14	-3.08	2.93
12	18.27	-3.04	-1.61	-3.07	3.04	-2.98	-3.00	2.92	-1.17
13	17.31	-2.62	-2.27	-0.21	-3.02	-0.20	3.09	1.38	0.79
14	30.10	-2.96	0.64	-0.22	-3.07	0.11	-3.00	2.83	-1.13
15	10.26	-0.64	1.18	0.07	3.03	3.11	3.03	1.67	0.84
16	15.88	2.96	1.52	-2.79	-2.97	2.94	-2.99	-3.05	-1.17
17	20.79	-2.92	1.32	-2.85	-3.02	-3.07	3.09	-1.35	-2.88
18	18.53	2.91	0.78	-0.11	-3.06	-0.32	-3.01	1.19	1.06
19	19.03	0.26	1.40	-3.04	3.01	-3.08	2.83	-3.11	2.83
20	7.26	0.01	1.10	0.11	-3.14	2.90	3.03	-1.90	1.08
21	6.78	-0.14	-1.89	0.02	-2.84	-2.82	-3.11	-1.39	3.04
22	29.93	-2.86	-1.92	3.12	2.74	-0.28	-2.88	1.22	1.17
23	24.07	3.06	-2.01	-3.00	2.95	3.08	3.04	1.82	1.26
24	17.51	0.09	1.25	-2.82	2.94	3.07	-2.99	-2.99	1.38
25	24.60	0.27	-2.06	-2.93	-3.11	-0.32	-2.68	1.14	0.91
26	19.99	0.10	1.03	0.22	3.06	-0.40	-2.74	1.11	1.12
27	15.91	-0.06	1.32	-3.10	2.83	-3.09	-3.02	-3.04	-1.07
28	18.61	-3.06	-2.18	0.30	3.13	-0.26	-2.72	1.07	1.03
29	13.05	2.97	1.05	0.27	-3.12	-3.00	2.99	1.57	1.11
30	15.74	0.61	-1.92	3.12	3.13	-3.04	-2.96	3.06	1.15
31	3.50	-0.19	1.32	0.21	-2.93	-2.98	-3.11	-1.37	2.84
32	23.00	0.49	-1.99	3.02	3.11	2.93	-2.80	1.43	1.03
33	17.78	0.15	-2.06	0.04	3.05	-0.04	-2.81	1.16	0.95
34	14.67	0.35	-1.84	2.99	2.98	-3.05	2.85	-1.13	3.10
35	3.42	0.25	1.35	0.03	-3.12	3.00	-3.09	2.85	-1.13
36	0.00	2.91	1.08	-0.09	-3.07	3.00	3.08	-3.03	3.10
37	19.48	0.07	-1.69	2.72	-2.96	-3.09	3.06	3.11	-1.07
38	23.09	-3.09	-1.88	-2.99	-2.86	3.09	2.86	-1.61	1.21

Table B.11: Labels, free energies and CV-space coordinates of m-nisoldipine's conformers in an acetone/ethanol mixture. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
0	3.48	-2.86	-1.76	0.11	-3.04	-3.11	2.96	-3.11	-3.10
1	9.21	-0.28	-1.92	0.02	3.12	-0.06	-2.86	2.89	1.25
2	11.56	-0.53	-1.73	-3.11	3.00	3.01	-3.08	-1.26	2.89
3	5.54	0.34	-2.08	-0.06	2.95	3.13	3.13	1.33	1.20
4	12.67	-0.11	-1.66	-3.13	-2.97	-2.99	3.10	-2.77	-1.22
5	8.02	3.08	0.88	-0.24	2.95	-2.83	2.96	1.44	1.05
6	14.22	0.19	-1.89	3.10	-3.09	0.10	2.95	-2.88	-1.15
7	10.03	0.01	1.09	0.39	-3.10	3.01	-2.89	3.00	-0.99
8	8.86	-3.07	1.06	0.10	2.97	-3.11	3.14	-1.55	2.99
9	13.73	-0.72	0.88	-2.97	3.06	3.00	-2.93	1.18	1.02
10	16.02	-0.24	-2.10	3.12	3.13	0.13	2.81	-1.37	3.10
11	9.00	3.07	1.18	-2.91	2.88	3.11	-3.10	-2.92	2.78
12	6.95	-0.25	-1.99	0.00	2.86	-3.06	2.92	3.00	-0.75
13	7.83	-0.41	1.44	2.93	-3.02	-3.13	-3.00	-3.13	-0.99
14	22.94	0.04	-2.05	2.94	-3.04	-0.53	-3.00	1.31	0.75
15	5.87	-0.27	-1.83	0.14	-3.07	-3.05	2.89	-1.26	-3.10
16	14.87	-0.65	-1.99	3.03	-3.09	-3.02	-2.92	1.38	1.17
17	5.69	0.03	1.00	0.07	-2.96	3.06	-2.90	2.86	3.03
18	10.34	0.47	-1.85	0.15	-2.93	-0.03	2.93	3.07	-1.15
19	14.45	2.96	-1.82	-2.86	3.04	-0.15	3.03	-3.04	2.81
20	11.20	-2.49	0.83	-3.14	3.00	3.11	-3.07	-3.06	1.21
21	18.18	-2.76	1.39	0.18	-3.05	0.16	2.86	-1.23	3.11
22	5.31	0.04	1.29	0.03	3.07	3.07	2.94	-1.09	3.05
23	18.77	2.83	-2.23	3.03	2.83	-0.26	2.97	2.97	-1.01
24	23.03	-2.87	-2.22	-2.80	3.04	-0.36	-3.07	1.05	0.95
25	8.52	-0.55	1.15	-2.99	-2.98	3.11	3.05	-1.18	2.97
26	2.64	-2.89	-2.14	0.02	2.85	0.22	3.07	-2.99	-1.17
27	19.54	0.04	1.05	-2.88	2.93	-0.08	-2.90	3.11	0.94
28	11.56	-0.22	-1.72	0.02	-2.86	-2.78	3.05	-2.96	1.18
29	20.35	-0.53	1.28	-2.96	-3.02	-0.14	2.89	-1.22	3.02
30	14.55	0.15	-1.97	0.26	-3.10	-0.32	-2.63	1.20	1.06
31	15.67	-0.89	1.12	-0.10	-3.13	-0.12	-3.07	2.92	0.86
32	7.59	3.08	1.41	-3.01	3.07	-3.00	-3.06	3.07	-1.04
33	13.06	-0.38	0.85	-0.06	-3.09	0.13	3.12	-1.48	3.06
34	25.85	2.72	1.45	3.08	-2.92	0.41	3.01	1.11	0.94
35	1.88	-0.06	-2.04	-0.07	2.89	-3.09	-3.02	2.84	3.03
36	14.69	-2.84	-2.00	2.95	-2.94	2.94	3.13	1.45	0.83
37	8.41	0.22	1.23	-3.12	3.05	2.92	-2.99	-3.08	1.33
38	10.26	0.09	-1.96	3.05	3.02	2.93	2.96	-3.06	2.93
39	8.91	3.02	-2.09	-0.28	3.02	-2.96	-3.07	1.38	1.12
40	10.96	-2.92	1.29	-2.88	-3.01	2.99	-2.97	1.23	0.91
41	13.82	3.11	-2.22	0.25	3.07	0.01	2.97	-3.06	1.56
42	8.62	2.96	1.32	-3.11	-2.93	-3.11	2.96	-1.17	3.07
43	8.85	2.84	-1.83	-3.12	-3.13	3.08	-2.86	-1.41	3.08
44	5.06	3.07	1.53	0.07	-3.12	-3.09	-3.10	-2.95	-1.13
45	0.00	2.85	-2.13	-0.05	3.14	3.09	-3.14	-1.43	2.94

Table B.12: Labels, free energies and CV-space coordinates of m-nisoldipine's conformers in an ethyl acetate/hexane. The labeling convention is consistent with that of Figure 5.1a

Conformer	Free Energy [kJ/mol]	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
0	13.25	0.60	-2.20	-0.31	-3.13	-0.11	-3.09	-2.92	1.25
1	9.99	0.18	1.35	-3.11	-3.02	3.14	-3.06	-1.46	-3.05
2	0.00	2.65	-1.98	0.16	3.06	3.10	3.10	3.04	-0.90
3	5.47	0.44	1.28	0.04	3.13	-3.01	-2.99	2.80	1.25
4	4.66	2.62	-1.95	0.10	2.90	3.11	-3.04	3.13	1.62
5	6.66	-3.10	-2.16	0.11	2.95	-0.11	-2.99	-3.10	1.38
6	3.34	-0.00	-1.72	3.03	-3.02	-2.97	3.09	-3.05	-1.00
7	6.60	3.00	1.08	-3.09	3.09	2.97	-3.01	3.12	-0.89
8	9.79	-3.02	-1.97	-3.12	2.96	-3.06	-3.08	-1.69	2.94
9	9.30	0.14	1.08	0.12	-2.88	3.11	3.08	1.10	0.93
10	10.11	-0.17	0.97	-3.06	3.01	3.09	3.07	2.94	-1.02
11	6.23	0.06	0.99	0.13	-3.07	3.09	3.09	1.92	-1.28
12	5.50	0.08	-2.13	-0.12	3.10	3.10	-3.10	1.62	1.06
13	10.49	-0.35	-2.19	0.20	-2.87	-0.10	-3.11	-2.60	3.00
14	3.03	-3.02	1.24	0.33	-3.06	2.96	3.11	3.04	-0.74
15	10.52	-0.48	-1.94	-3.08	-3.03	3.12	-3.09	-1.40	-3.07
16	8.82	2.89	-1.69	-3.07	-3.02	2.90	3.14	-2.99	-1.01
17	14.72	-3.10	-1.95	0.09	-3.04	-0.05	-2.71	1.26	0.88
18	10.24	0.26	-2.23	-3.13	-2.90	3.02	3.04	-3.08	-3.09
19	15.02	-2.46	0.99	-0.12	-3.08	-0.02	3.06	-1.52	3.02
20	12.16	3.09	1.12	0.06	2.90	-0.25	-3.12	-2.82	-1.05
21	20.07	-2.97	1.33	3.11	-2.99	-2.99	-3.04	1.32	1.05
22	5.76	0.27	-1.92	0.16	3.03	-2.96	2.91	-1.29	3.06
23	20.82	-2.52	-2.27	2.81	-3.05	-0.05	-3.07	1.34	0.77
24	8.33	0.07	-1.56	0.23	3.13	-3.05	2.89	2.92	-1.09
25	15.16	0.22	-1.84	0.12	3.12	-0.41	-2.66	1.25	0.99
26	14.27	0.17	1.13	0.17	-3.10	-0.19	-3.09	1.56	0.77
27	13.09	-2.75	-1.94	0.37	-2.79	3.05	3.11	-1.39	2.99
28	1.74	-2.86	1.15	-0.17	3.04	-3.08	-3.03	-2.86	1.30
29	11.02	3.02	-2.19	-3.05	-3.09	3.07	-2.94	1.48	1.00
30	19.18	-2.82	-1.83	3.07	3.13	0.24	2.89	-2.92	-1.03
31	5.10	0.02	-1.87	0.13	2.92	0.08	2.96	-3.01	-1.06
32	8.97	2.92	1.24	-0.02	3.00	-3.12	2.88	-1.14	3.14
33	5.86	-0.47	1.01	0.15	3.07	2.94	-3.13	-1.24	-3.09
34	13.07	3.13	-2.05	0.28	2.99	0.10	3.05	-1.25	3.12
35	11.94	-2.68	1.01	0.05	-2.99	-0.38	-3.06	2.78	1.14
36	22.94	-0.24	1.65	2.99	2.94	-3.11	-3.10	2.88	0.81
37	13.34	-0.34	-2.20	2.95	3.05	3.10	-2.89	1.14	1.10
38	19.60	-2.82	1.46	-2.79	3.03	-2.97	-3.09	-1.62	3.01

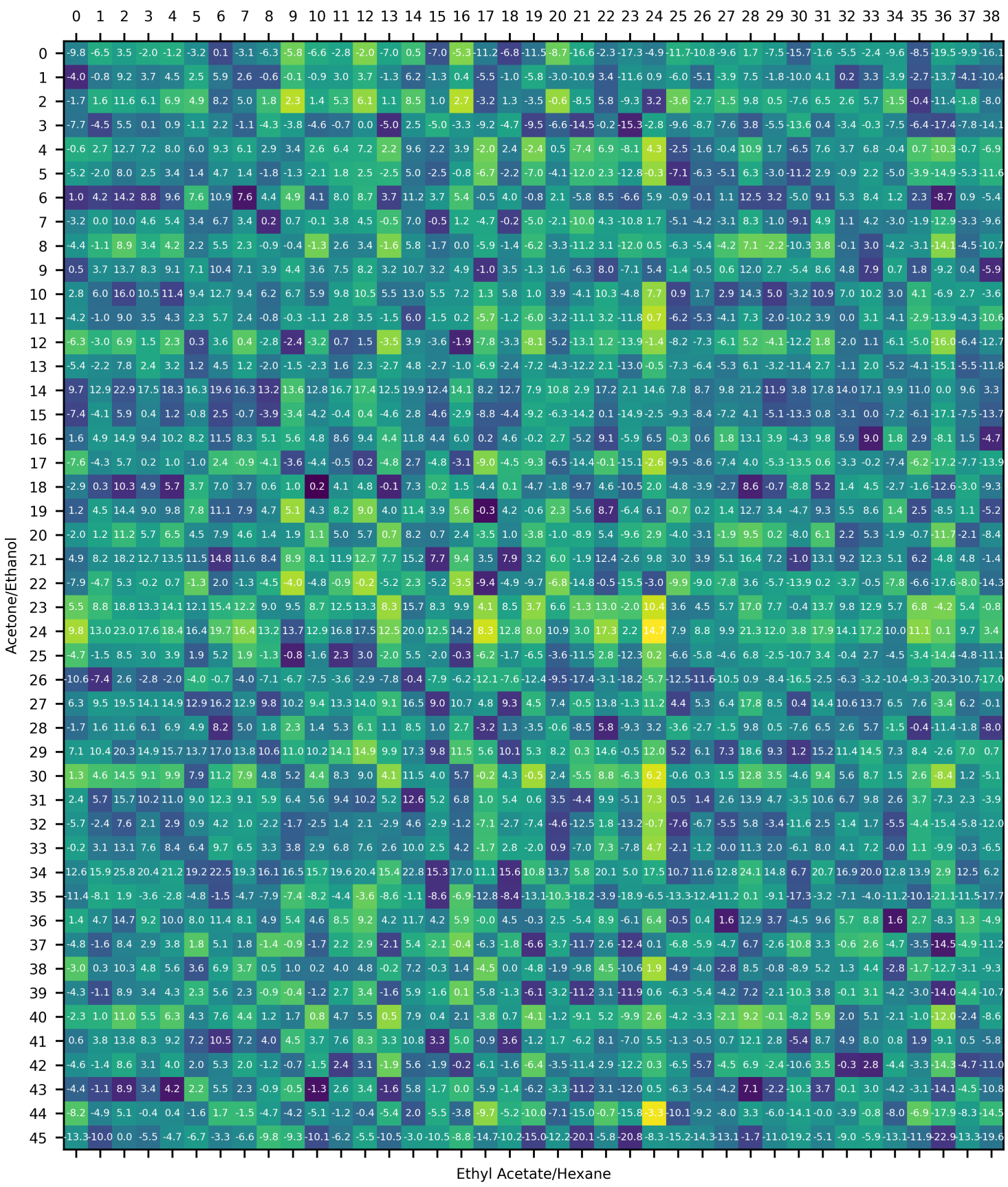


Figure B.13: Figure 5.22, enlarged

Appendix C

Supplementary Materials: Example of Code Tutorial

The following pages contain the static PDF of the jupyter notebook tutorial included with this project's code. This tutorial walks through the application of the analysis method on the example of alanine dipeptide. The interactive notebook is available at:

https://github.com/ucecvan/Twister.

In [1]:
```
from Twister import Twister
import matplotlib.pyplot as plt
```

This notebook is a walkthrough of the Twister module, which utilizes Density Peaks Advanced Clustering to study the conformational free energy surfaces of organic molecules. Twister takes as input a COLVAR file generated by the Plumed software package. This COLVAR file should contain within a record of the MD trajectory through the values of the torsions which describe the conformational space, as well as a record of the final bias in each torsion at each of these points in conformational space.

The conformational space we are studying here is the Ramanchandran plot of alanine dipeptide, or the conformational free energy surface of alanine dipeptide in terms of its phi and psi angles. The COLVAR file in this directory is the product of 1 microsecond concurrent WTmetaD simulation of alanine dipeptide in vacuo, where we are biasing both phi and psi independently. The COLVAR file records in it's first column the time at which the simulation trajectory was sampled, the second and third columns contain the values of phi and psi at the indicated simulation time. The fourth and fifth columns contain the amount of bias on phi and psi at the values in columns 2 and 3.

The COLVAR file should look something like this:

```
#! FIELDS time phi psi metad1.bias metad2.bias
#! SET min_phi -pi
#! SET max_phi pi
#! SET min_psi -pi
#! SET max_psi pi
 0.000000 -2.636189 2.965787 405.437016 401.372906
 0.200000 -2.826828 2.609132 405.375596 400.509111
 0.400000 -2.555259 3.072395 405.141060 400.972452
 0.600000 -2.886444 2.791932 405.068455 401.359558
 0.800000 -2.376418 2.984626 404.171513 401.329474
 1.000000 -2.491054 2.714065 404.815654 401.086567
 1.200000 -2.841954 2.599198 405.309145 400.442419
 1.400000 -2.873339 3.097467 405.146177 400.835999
```

Twister can interpret COLVAR files with different numbers of torsions, and not all torsions have to biased during metadynamics. All COLVAR files must be ordered as above, with time, then torsion, and then bias columns. Twister interprets the COLVAR file using a biid variable. biid is a list with elements representing the torsions in the molecule. Biased torsions are indicated with a 1 and unbiased torsions with 0. The biid for the COLVAR file above thus looks like:

In [2]:
```
biid = [1,1]
```

We instantiate our Twister object for Alanine Dipeptide with the biid variable:

In [3]:
```
AD = Twister(biid)
```

We read in the COLVAR file. The second argument specifies how many lines to skip at the start of the file.

In [4]:
```
AD.ColvarLoader('COLVAR2',5)
```

The ColvarLoader method automatically takes the last 2/3 of the COLVAR file, ignoring the start of the trajectory, where metadynamic bias is deposited at a high rate. In this demonstration notebook, we will be running our clustering on a further subset of just 5000 lines of the COLVAR file, so that the code can be demonstrated without excessive use of computational resources. The Downsample method returns a random, non-repeated sample of the current COLVAR attribute, which can be set using the ColvarSetter method. If it is desirable to use a larger dataset, change the argument in Downsample below.

In [5]:
```
Colvar5k = AD.DownSample(5000)
AD.ColvarSetter(Colvar5k)
```

Once the desired COLVAR variable is loaded in, it is time to run the main Clustering method. This method will run an initial DPA clustering on the biased distribution of datapoints in the conformation space defined by the torsions in COLVAR. The resulting densities will then be reweighed to account for the effect of the biases. DPA clustering will then be rerun on the reweighed densities. The cluster centers will correspond to density peaks, which we treat as maxima in the probability distribution. A Boltzmann inversion on all of the densities yields the free energies associated with each configuration, and the density peaks correspond to the conformational free energy minima. Running this method may take a few minutes, depending on your machine. The processor and memory demand increases with the square of the dataset size, so proceed with caution when increasing dataset size.

In [6]:
```
AD.Clustering()
```

The results of the Clustering attribute are stored in several attirbutes. The DPA object itself is stored in the DPAobj attribute. A summary of the properties of the cluster centers (free energy minima) are presented in the RefClusters attribute. The free energies generated are stored in energies, and the truncated free energies is stored in energies2 (see Note below). Similarly, the truncated Colvar, coordinates, and biases are stored in Colvar2, coords2 and bias2.

Note: As part of a heuristic to assist DPA in avoiding unneccesary noise, any data points which present an energy over 100kJ/mol are removed from the dataset before a final reclustering. This eliminates noise generated by data points in irelevant high-energy regions. equivalent data points in the Colvar, coords, and biases attributes are also removed, so a coherent truncated dataset is stored. In low dimensional cases, such alanine dipeptide, here, it is highly unlikely that any points will have such low density/high energy, so the energy cut-off datasets are the same as the original ones.

The RefClusters attribute contains the results for each of the cluster centers:

In [7]:
```
AD.RefClusters
```

Out[7]:
```
{0: [-1.363335,
  0.845914,
  411.329163,
  394.508638,
  1691,
  0.6597042006785401,
  82253.0],
 1: [1.208069,
  -0.83565,
  399.215922,
  399.895675,
  1372,
  8.162208056911219,
  39216.6],
 2: [-2.73749, 2.923128, 405.558802, 401.436067, 1922, 0.0, 5167
8.2]}
```

The format of RefClusters is a dictionary containing an entry for each cluster center. The keys are indexed from 0. Each entry follows the same format: First, the coordinates of center are presented, then the biases. These are followed by the population of the cluster (the number of other points assigned to the cluster by DPA). this is followed by the free energy of the cluster in kJ/mol, and lastly the time the configuration was sampled by the simulation.

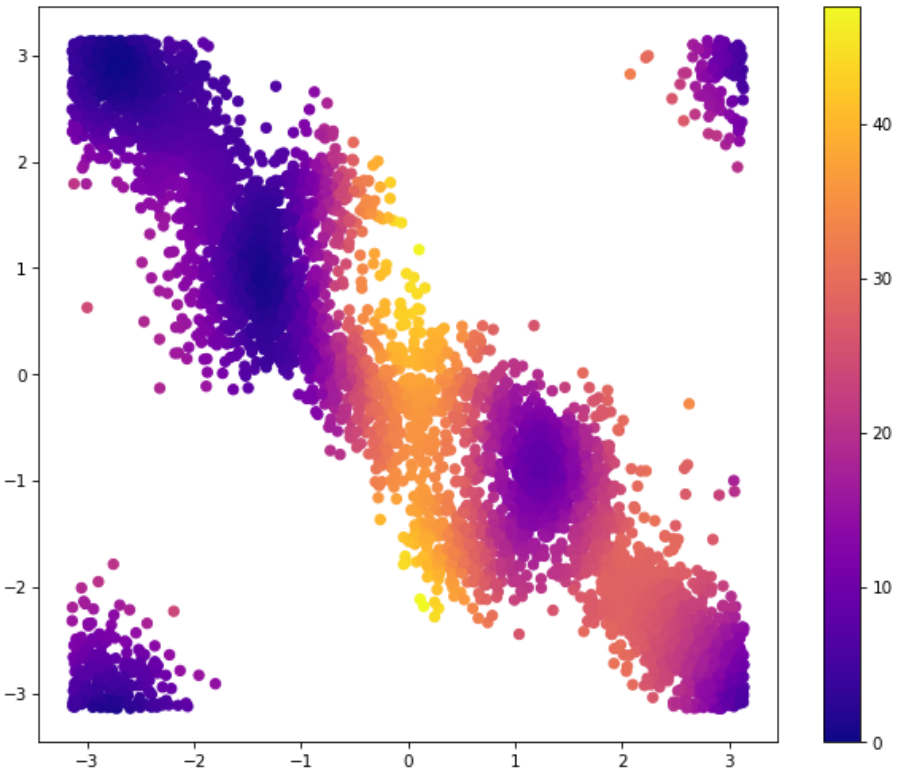
In [8]:
```
AD.RefClusters[0]
```

Out[8]:
```
[-1.363335,
 0.845914,
 411.329163,
 394.508638,
 1691,
 0.6597042006785401,
 82253.0]
```

For alanine dipeptide, we can plot a per point FES using the coords2 and energies2 attributes

In [9]:
```
plt.figure(num = 1,figsize = (10,8), dpi =75)
plt.scatter(AD.coords2[:,0],AD.coords2[:,1], c = AD.energies2, cmap =
plt.colorbar()
```

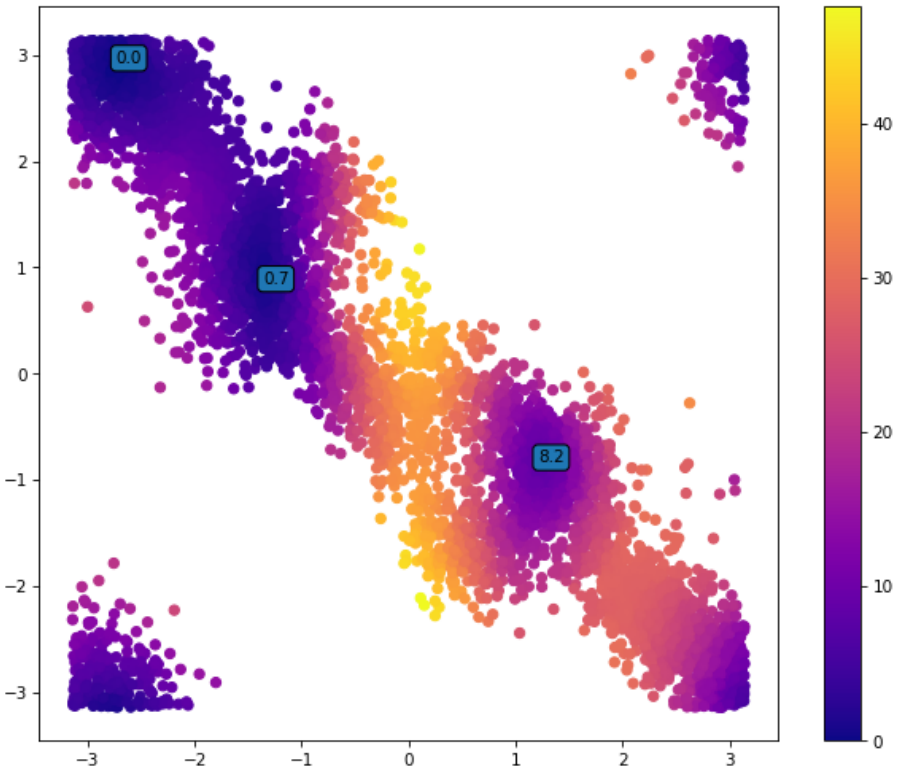
Out[9]:
```
<matplotlib.colorbar.Colorbar at 0x7f79b4801550>
```



We can annotate the above plot with information from RefClusters, for instance labelling the positions of the free energy minima with the free energy, rounded to 1 decimal place:

In [10]:
```
plt.figure(num = 1,figsize = (10,8), dpi =75)
plt.scatter(AD.coords2[:,0],AD.coords2[:,1], c = AD.energies2, cmap =
plt.colorbar()

for i in AD.RefClusters:
    plt.annotate(round(AD.RefClusters[i][5],1),(AD.RefClusters[i][0],
```



In order to determine that 5000 data points (or however many were used) is sufficient to capture the structure of the free energy surface, we repeat the clustering process on a smaller number of data points, and see if the cluster centers are in the same positions, and have the same relative free energies. We call this process Sampling Consistency Analysis and it is automated by the SCAnalysis method. It takes as input a list of dataset sizes to be tested. We will start at 500 data points, and add 500 at each iteration until we arrive at 4500. Be aware that this process repeats the clustering process from scratch for each dataset size, and can thus be quite time-consuming.

In [11]:
```
samplerange = range(500,5000,500)

AD.SCAnalysis(samplerange)
```

The results of the analysis are stored in the FAdata attribute.

In [12]:
```
dlist = AD.FAdata[0]
elist = AD.FAdata[1]
nlist = AD.FAdata[2]
clist = AD.FAdata[3]
cdlist = AD.FAdata[4]
```

All the results are lists where each member corresponds to one of datasets specified in samplerange.

dlist contains the mean separation of equivalent clusters between the smaller dataset and the complete dataset

elist contains the mean energy difference between between equivalent clusters in the smaller dataset and complete datasets

nlist contains the dataset sizes, essentially a duplicate of samplerange

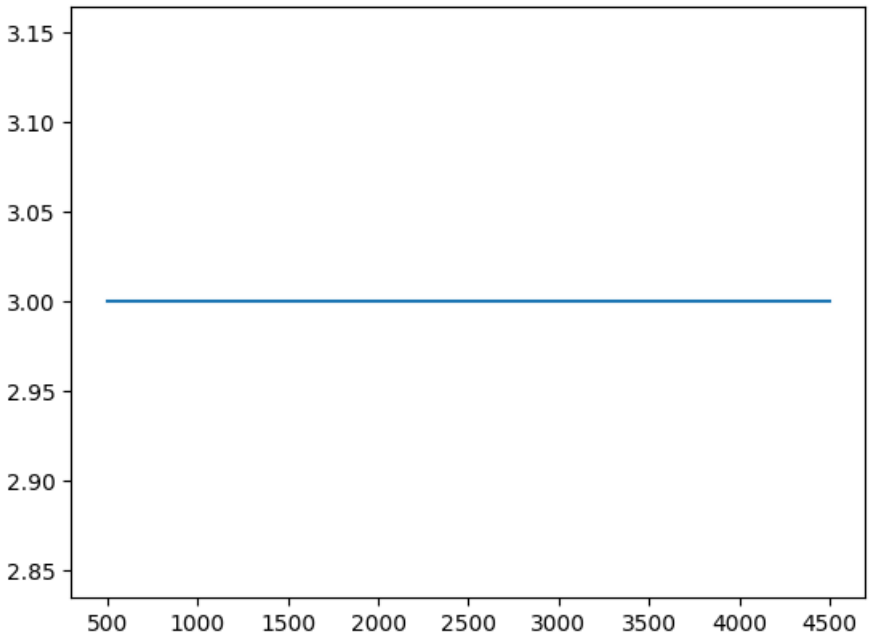
clist contains the number of clusters observed during each run on a smaller dataset

cdlist contains the equivalent to RefClusters for each clustering run on the smaller dataset

First, a check on the number of clusters identified per dataset size. For alanine dipeptide, this should be 3 clusters. Due to the very small dataset sizes at the lower end of our sample range, we may observe anomalies. As long as the second half of the plot remains constant, this is a good sign.

In [13]:
```
plt.plot(nlist,clist)
```

Out[13]:
```
[<matplotlib.lines.Line2D at 0x7f79b488beb0>]
```

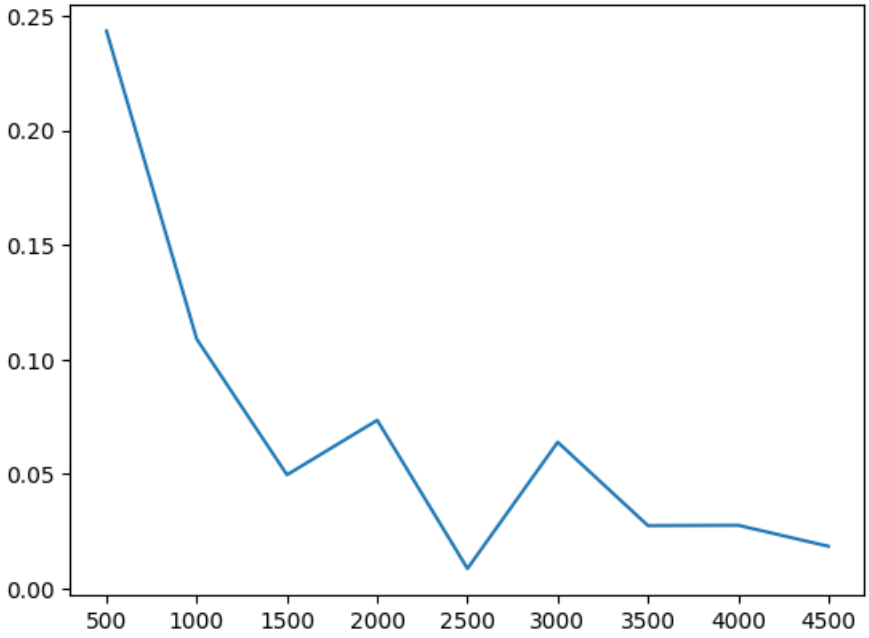


Next we observe how the mean separation and energy differences evolve as we get closer

to the final datset size. In the case of alanine dipeptide, the 5000-point FES we created above is already very close to the theoretical optimum, so we expect, for the datasets closer to 5000, to see positional differences of less than 0.25 rad, and energy differences of less than 0.5 kJ/mol. Anomalies in the number of clusters identified in the plot above will be reflected in the plots below.

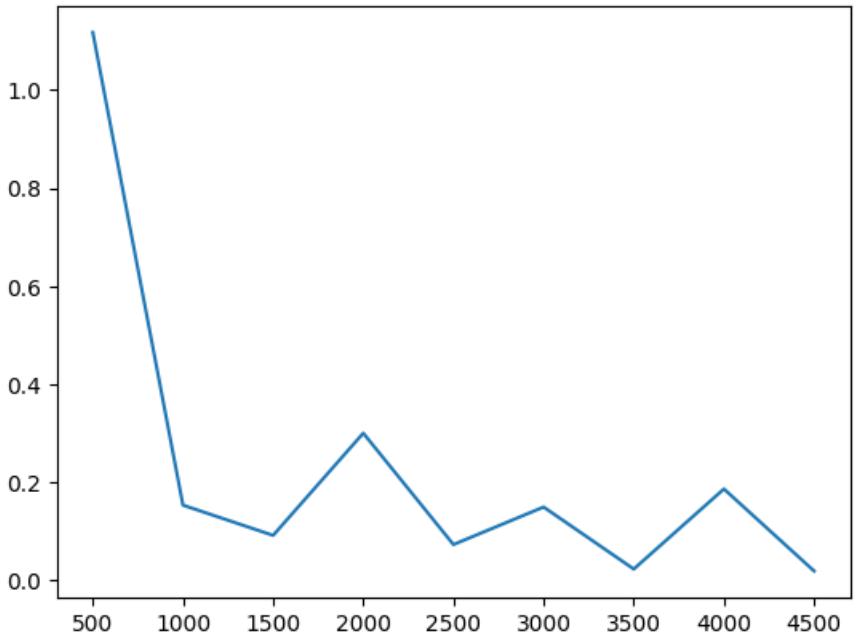
In [14]:
```
plt.plot(nlist,dlist)
```

Out[14]:
```
[<matplotlib.lines.Line2D at 0x7f79b5881190>]
```



In [15]:
```
plt.plot(nlist,elist)
```

Out[15]:
```
[<matplotlib.lines.Line2D at 0x7f79b4f533a0>]
```



This notebook worked through the trivial two-dimensional case of alanine dipeptide, using a small number of data points to illustrate the working principles of Twister in a cheap and visually accessible way. It should be noted that things get more difficult when considering cases where the conformational space is higher dimensional (a molecule with more torsions). Firstly, as the dimensionality grows, so does the number of data points required to map out the free energy surface. Secondly, when the dimensionality is higher than 2, a complete visualization of the free energy surface in terms of every torsion becomes impossible. In those cases, the results of the Sampling Consistency Analysis become even more important when evaluating the quality of the results.

In this case study, all of the default parameters were used. However Twister can accept custom distance metric in the attribute metric, for determining the separation between two points in the conformation space, should the default Euclidean metric not be appropriate. All energy calculations assume a simulation temperature of 300 K, unless a temperature is specified in the T attribute. For noise reduction, densities calculated by DPA are smoothed over a radius of 0.1 rad by default. This can be modified by changing attribute smoothrad. Note that increasing this radius comes with increased computational cost and reduced spatial resolution.

In [ ]:
```
```