SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area - PhD course in
Physics and Chemistry of Biological Systems

Nonparametric density estimation methods and applications to molecular simulations

Candidate:
Matteo Carli

Advisor:
Alessandro Laio

Academic Year 2020-21

Contents

List of Figures

List of Tables

Nomenclature

$\delta(\mathbf{x}-\mathbf{x}_0)$	Dirac delta distribution (simply known as delta function) centered on vector $\mathbf{x}_0$
$\mathbb{1}_d$	Identity matrix in d dimensions
$\mathrm{Box}_{[a,b]}(x)$	Boxcar function for the interval $[a,b]$: returns 1 if $x \in [a,b]$, 0 otherwise
ω_D	Volume of the unit-radius hypersphere in D-dimensions (see equation (D.3))
$\Theta(x-x_0)$	Heaviside theta distribution centred on point x_0
$B^D(r,\mathbf{x})$	D-ball of radius r centered at point $\mathbf{x}$
I_Ω	Indicator function of set Ω
$T_\Omega(\mathbf{x})$	Tangent hyperplane of manifold Ω at point $\mathbf{x}$

Acronyms

Notation	Description	Page List
kNN	k-Nearest Neighbours	19
BMTI	Binless Multidimensional Thermodynamic Integration	80
CDA	Constant Density Assumption	20
COD	Curse Of Dimensionality	2
CRB	Cramér–Rao Bound	24
CS	Core Set	42
CV	Collective Variable	2
DOFs	Degrees Of Fredom	9
DR	Dimensionality Reduction	1
FES	Free Energy Surface	2
ID	Intrinsic Dimension	1
IID	Independent and Identically Distributed	13
KD	Kernel Density	15
KDE	Kernel Density Estimator	15
MC	Monte Carlo	1
MD	Molecular Dynamics	1
MEC	Map-Existence Condition	57
MEP	Minimum Energy Path	86
MIAE	Mean Integrated Absolute Error	14
MISE	Mean Integrated Squared Error	14
MLE	Maximum Likelihood Estimator	6, 23
NEB	Nudged Elastic Band	112
NG	Neighbours Graph	19
NN	Nearest Neighbour	19
PDA	Pocket Doorway Area	44
PDF	Probability Density Function	2
SASA	Solvent-Accessible Surface Area	44

Chapter 1

Introduction

Molecular Dynamics (MD) and Monte Carlo (MC) simulations are popular techniques in the study of classical molecular systems. The microscopic detail they provide can elucidate biologically and chemically relevant mechanisms which cannot be directly studied experimentally. Due the rapid advances in computational hardware and software, recent years have seen a boost in the amount of data generated in these kind of simulations, making of molecular simulations a prototypical example of big data problem[35]. This calls for the development of approaches to treat, analyse and represent such data.

The data produced in molecular simulations are typically very long trajectories of highly dimensional vectors with the coordinates of hundred thousands atoms. One of the key task in order to be able to treat them is then the identification of a lower-dimensional, *but still fully informative*, representation: the so-called problem of Dimensionality Reduction (DR). Generally, a first step in this direction is made by choosing a space of descriptors; for example, in many applications in the analysis of molecular simulations it can be safe to neglect all the coordinates of the light atoms and of the solvent molecules. However, this step, called featurisation, typically brings to a description which is still relatively high-dimensional description. Fortunately, a fact comes to help: due to the specific form of the interactions between the atoms, configuration space is, pictorially speaking, almost empty. The data are typically localised, at least approximately, on a manifold, called intrinsic data manifold, of dimension much smaller than the number of descriptors, called Intrinsic Dimension (ID). Optimal DR should then allow to restrict the analysis to nothing more and nothing less than the intrinsic manifold; if more information is retained, this has consequence on the computational cost and on the efficacy of the analysis methods. If a further projection is done, some information is typically lost.

The intrinsic manifold can in principle parametrised by a set of generalised coordinates expressing all the relevant degrees of freedom of the system. In the field of molecular simulations these coordinates are referred to as a Collective Variable (CV). The Probability Density Function (PDF) over the full coordinate space can be marginalised over the CVs. If the CVs parametrise the intrinsic manifold, this marginalization does not imply any significant information loss. In general, even if the CVs describe the manifold only approximately, by taking the negative logarithm of this reduced PDF one obtains the Free Energy Surface (FES) of the collective variables (the reduced PDF and the FES are defined mathematically in the next chapter). On this hypersurface, if the CVs are appropriately chosen, the metastable states of the system appear as local minima separated by barriers. In order for the system to dynamically cross these barriers, a time which is exponentially increasing in the barrier height is required. Such transitions are in fact a typical example of rare event process[150]. The FES is typically highly rugged and complex[190], so that many of such barriers and wells are encountered in a typical reaction or biomolecular process. Thus, the FES is a key quantity to understand and characterise the properties of molecular systems[152]. It can provide information on the occupation of the states of the systems but also e.g. help measuring thermodynamic observables, elucidate reaction pathways and chemical mechanisms, serve as input for clustering and other pattern recognition algorithms. The concept of FES is useful even outside the realm of molecular simulations, wherever knowing the distribution of the data is required.

The characterisation of the data manifold in its full complexity is a very challenging theoretical and computational task The main difficulty is that the ID is typically high, of order 10 or more, and the so-called Curse Of Dimensionality (COD) arises[21, 73, 140]: data in high dimensions have a tendency to become all far from one another and most points happen to lie on the boundary of the data manifold[14, 203]. In these conditions, many analysis and learning methods fail to provide meaningful results. Yet, restricting to fewer dimension for manipulation or visualisation purposes can wash out relevant information and result in a misleading description. This poses to computational physicists a dilemma which seems to be difficult to solve in practical applications.

CV selection[70] is considered a key open problem in molecular simulations. When done manually, this task requires a detailed insight on the system. Therefore many machine learning techniques have emerged to tackle the problem of unsupervised dimensional reduction and CV identification. The simplest of them are linear projection methods[104, 105, 146, 188, 202] but these fail when the data manifold is not a hyperplane. However, the data manifolds are known to be highly nonlinear, twisted and topologically complex[17]. Many non-linear projection methods have been developed to tackle tbhis problem[43, 44, 169, 186], but also this methods may fail when the topology is not trivial: if a data manifold is, say, two-dimensional but topologically isomorphic to a thorus, it will be impossible mapping it to a two-dimensional description with a Cartesian metric. Some methods manage to cope with complex topologies[32, 111], but typically pay a price in terms of simplicity. Many other techniques for CV selection focus only on the description of the transition process, which is almost always well-described by a single coordinate, renouncing to provide a full description of the data manifold.

In order to get around this issue, in our group we developed an approach that is able to compute the free energy surface directly in the space of the descriptors without explicit dimensional reduction, namely without defining explicitly the CVs. This method, called PAk[157], is based on the k-Nearest Neighbour (kNN) density estimator, which in order to estimate local PDF on a point of the sample fixes a number of neighbours k and divides it by N and by the volume of the hypersphere centered on such point containing $k-1$ other neighbour. By estimating the ID, assuming local flatness of the data manifold and providing an adaptive procedure for the selection of k, PAk is able to greatly alleviate the COD without the need of defining any CV explicitly. Thanks to its good performance as a multidimensional free energy estimator it could be used in the analysis of a molecular system of ID up to 28. The free energy is then used as as input for a clustering analysis which allows to successfully identify the metastable states.

As we will see, the PAk approach, even if very powerful, is affected by important limitations. These limitations have been the object of our theoretical investigation.

In Chapter 2 we briefly introduce the concepts of dimensional reduction, the probability density function and the free energy. We then provide an overview of the most common nonparametric density estimation methods, namely the histograms, the kernel density estimators and the kNN estimator. We focus on nonparametric methods since they are very flexible, since they do not requiring any specific assumption on the functional form of the estimated quantity. The only parameter they all, in some form, require is the selection of a length scale, called also bandwidth, which determines how locally they operate. The main drawbacks of these methods are connected to the tuning of this bandwidth[177]. On one hand, their locality requires a high statistic for them to be accurate. To overcome the curse of dimensionality one is forced to select a larger bandwidth, to reduce noise in the estimates. On the other hand, the selection of a large bandwidth would cause to lose relevant detail and thus bias the results. In the literature on density estimation the search of a balance between these two needs is referred to the bias-variance tradeoff. One way to address this problem is to adapt the selected bandwidth point by point based on the local sample properties: make it smaller in regions where the statistic is high, larger where it is low. The kNN estimator is indeed a first step in the direction of adaptivity. In the following chapters we show how the adaptive neighbourhood selection proper of kNN can be improved to obtain estimators which perform even better in high-dimensional settings.

In Chapter 3 we present PAk, the multidimensional free energy estimator recently developed in our group[157]. We focus on the description and the discussion of the key ingredients which makes it more suitable than the standard kNN approach to estimate the free energy in high dimension. The first one is that the configuration space volumes which enter the definition of the estimator are measured in the low-dimensional intrinsic manifold rather than in the full embedding space. This trick prevents the positional information of the data from being diluted on irrelevant directions transverse to the data manifold. Assuming that the data manifold is Riemannian, namely locally flat, the data manifold is locally approximated by its tangent hyperplane and distances between neighbours, the only distances used in the estimator, can be measured in this low-dimensional Euclidean space. This allows to operate on the intrinsic manifold without parametrising it explicitly. The only prerequisite is estimating the local intrinsic dimension[68], as this is required to measure the volumes in the embedding manifold.

The second key ingredient of PAk is the definition of an optimal criterion for the neighbourhood selection for all points in the sample. Central to this purpose is the formulation of kNN in terms of a maximum-likelihood estimator. This point-adaptive neighbourhood selection makes the already-adaptive kNN doubly adaptive, thus much more robust to the curse of dimensionality.

The third key ingredient is introducing a variational slope parameter, which allows describing linear variations of the free energy within the neighborhood used to compute it. We will see that this is equivalent to taking the limit of a bandwidth going to zero. We call this property punctuality of PAk. This fact is exploited in Chapter 5 to define an efficient reweighting scheme for statically biased simulations.

In this chapter we also survey some of the main drawbacks which affect PAk. Firstly, despite proven robust to the course of dimensionality, its nonparametric nature dooms nonetheless its performance to drop in very high dimensions. Secondly, PAk's likelihood model introduces spurious correlations among estimates at neighbouring points. Thirdly, PAk provides free energy estimates only in correspondence of datapoints, but a generalisation of the methods for points lying outside this finite set would be required in many applications. In this chapter we also propose a scheme to efficiently compute interpolated free energies in generic points of configuration space which is coherent with the PAk approach: the PAk interpolator.

Chapter 4 presents an application of PAk, in its original formulation, to a problem of biochemical relevance, the study of the SARS-CoV-2 Main Protease. By analysing a very long molecular trajectory of the system we characterise its metastable states and propose potential druggable pockets. We analyse it using two different feature spaces: the space defined by all the ψ backbone dihedrals of the protease, and the space defined by the contacts between pairs of residues which break or form during the dynamics. The two metrics are both sensitive to local and global conformational changes in the peptide, but capture different details: the ψ coordinates keep track of the changes in the protein backbone; the mobile contacts metrics, instead, also keep track of the side-chains rearrangements, while neglecting fluctuations around the completely formed or completely unformed contacts. Both feature spaces contain $\mathcal{O}(10^2)$ coordinates, demonstrating the capability of PAk to provide free energy estimates in high dimensional spaces.

We are able to identify 18 metastable states of the system, which we characterise by considering the accessibility of the active site. Based on this analysis we propose some relevant contact patterns and three possible binding sites which could be targeted to achieve allosteric inhibition.

We show that all three proposed target sites are comprised in pockets with high druggability score according to the software PockDrug. By looking at sequences of proteins in the same Pfam family we find that all the residues involved in the proposed target sites are conserved. We consider it a hint that our proposed targets and the consequent allosteric mechanisms might be weakly exposed to mutations. The insight on this molecule's conformational changes provided by our analysis as well as the transferability of the same approach to other systems might prove useful for the design of farmaceutical inhibitors.

Chapter 5 introduces the first theoretical development presented in this Thesis, a procedure to estimate the free energy in high-dimensional spaces starting from a sample of points generated in a biased simulation. This protocol consists of computing the biased free energy at all points in the dataset using PAk and then reweighting this quantity point by point simply subtracting the numerical value of the applied bias in each point.

The simple additive form of this reweighting procedure is a nontrivial result. First of all, it crucially relies on the punctuality of PAk. One one hand this estimator optimally selects for every point in the dataset the size of the neighbourhood considered in the free energy estimate; on the other hand, the likelihood maximisation peculiar of PAk extrapolates the value of the free energy in the limit of neighbourhood size (or bandwidth) going to zero, which makes the estimate more punctual than in other kernel-based methods.

Secondly, since any free energy estimate involves integration over degrees of freedom which are not necessarily those which are biased, we describe the condition under which it is possible to reweight in an Umbrella-Sampling fashion the biased free energy over some coordinates $\sigma(\mathbf{x})$ when the applied bias potential is a function of some possibly different CVs $\mathbf{s}(\mathbf{x})$. In short, this is possible if all the information necessary to define the biasing CVs $\mathbf{s}$ is encoded in the coordinates σ over which the free energy is computed. In other words, if the intrinsic manifold the $\{\mathbf{s}_i\}_i$ lie on can be mapped to a submanifold of the manifold the $\{\sigma_i\}_i$ lie on.

We test our unbiasing approach on several model free energy surfaces and on realistic systems for which the ground truth free energies are estimated from an unbiased simulation. The results show that in all tested cases bPAk is an unbiased estimator of the ground truth values. We also discuss the applicability of this punctual form of the reweighting to other finite-size kernel methods in statically biased simulations.

In Chapter 6 we discuss what we consider the most important theoretical development presented in this thesis, the *Binless Multidimensional Thermodynamic Integration* (BMTI) free energy estimator. Its development was motivated by the fact that PAk, despite unbiased and robust, produces noisy estimates even in conditions of low dimensionality and high sample density, where other methods' fluctuations typically reduce. Thus, BMTI was conceived with the purpose of providing similar free energy values at neighbouring points, as it should be for continuous and smooth functions.

In the first part of the chapter we discuss the main ingredient of the approach: the accurate and robust estimates $\hat{\delta F}$ of the free energy differences between neighbouring points and of their error. These estimates are based on a nonparametric estimator of the average free energy gradient on optimally-sized neighbourhoods. The size of these neighbourhoods is selected using the same procedure as the one employed in PAk, inheriting, therefore its double adaptive formulation. Next, we show that the gradients estimated in this manner can be used to estimate the free energy differences $\hat{\delta F}$s between a data points and all its k neighbours. We prove that the $\hat{\delta F}$s are normally distributed around the true values and spread with the estimated standard deviation. Therefore we considered them as marginal random variables of a multivariate normal distribution whose diagonal covariance matrix has the estimated variances of the $\hat{\delta F}$s as entries. We interpret such distribution as a likelihood for the error-affected observations $\hat{\delta F}$ as a function of the free energies F seen as parameters. The corresponding Maximum Likelihood Estimator (MLE) produces the BMTI estimates $\hat{F}$.

The motivation of the name defining the acronym BMTI shall now be clearer: BMTI estimator does not specify any requirements for the position of its inputs, there is no grid and no binning to populate, so it is *binless*. Thanks to the point-adaptiveness, mutuated from PAk, of its constituent blocks $\hat{\delta F}$ it allows mitigating the COD, making it suitable for high-dimensional applications, hence *multidimensional*. Finally, in order to reconstruct the FES it proceeds in a TI-like fashion by considering all the possible paths connecting points in the neighbours graph and computes all the relative free energy differences along them simultaneously, by solving a linear system, using the estimates $\hat{\delta F}$ as increment and their estimated errors as weights.

BMTI estimates have proven to outperform other nonparametric methods when tested on many model systems. We also define an observable that measures the estimators' roughness. BMTIis has the smallest roughness of all other nonparametric estimators we tested while PAk, as expected, is the most rough. There was only one case in which BMTI failed: when applied to a sample in which the saddle points connecting the thee wells of the FES had not been sufficiently sampled. Since the likelihood model is defined only in terms of free energy differences, the connectivity of its neighbours graph is a strict requirement in order to produce meaningful results. The second problem which affects BMTI is the inefficiency of its error estimator, on which much research effort is being devoted.

Chapter 2

Theoretical Background

The typical output of molecular simulations are very long trajectories, namely sets N of time-ordered configurations of the system. In its rough form each configuration consists of the collection of all the Cartesian coordinates of the atoms and is, therefore, generally very high-dimensional even for relatively simple systems. For example, a configuration of a peptide includes not only the coordinates of all the atoms of the peptide itself, but also of all the atoms of the solvent in which it is immersed, for a total of $\mathcal{O}(10^4)$ real numbers even in the simplest cases. We refer to the space of these vectors as the raw coordinate space, and we denote by D_r its dimension. Thus, our raw dataset is the collection $\mathbf{X}_r := \{\mathbf{x}_{r,i}\}_i$ of the N configurations $\mathbf{x}_{r,i}$.

In typical applications, not all of these vector components are interesting and their great number would make a direct analysis impossible. As a first step, a set of $D \ll D_r$ features is therefore typically selected. These features are typically combinations or simple projection of the raw coordinates. This procedure, called *featurisation*[80], requires of course some insight on the system (although recently much effort has been put in automatic feature selection[135, 195]), and is typically done by keeping enough features so that no relevant information is lost. For example, in order to capture conformational changes of large proteins, one can keep track of the positions of all the internal dihedral angles of the protein, assuming that the solvent degrees of freedom do not play a significant role, and that the internal bonds and angles are approximately fixed during the dynamics. These selected features live in what we will call the coordinate space or the configuration space. The representation of the dataset in this space is $\mathbf{X} := \{\mathbf{x}_i\}_i$.

In all the following we will only assume that an appropriate metric can be defined on this space (not necessarily Euclidian). Moreover, we will assume that our data are a sample obtained by a MD or MC simulation in the canonical ensemble[190]. We make this choice since we are interested in the application of our methods in the field of molecular simulations. Again, however, this is done without loss of generality, since the methods hereby discussed can straightforwardly be applied, and indeed are[61, 81, 204] to many other fields where understanding and representing the distribution of high dimensional data is required.

For exposition simplicity, we assume that the configurations of our system are distributed with the canonical probability density:

$$\rho(\mathbf{x}) := \frac{e^{-\beta U(\mathbf{x})}}{\mathcal{Z}} \ , \tag{2.1}$$

where $U(\mathbf{x})$ is the potential energy, $\beta := (k_B T)^{-1}$, with k_B Boltzmann constant and T the temperature of the system, is the so-called inverse thermodynamic temperature or coldness[127, 130] and $\mathcal{Z}$ is the configurational partition function $\mathcal{Z} := \int e^{-\beta U(\mathbf{x})}\, \mathrm{d}\mathbf{x}$. The key thermodynamic potential in this case is the Helmholtz free energy of the system:

$$A := -\beta^{-1} \ln \mathcal{Z} \ . \tag{2.2}$$

2.1 Dimensional reduction

The chemical and physical interaction between the degrees of freedom of the system restrain the dynamics only to a part of configuration space. Indeed, configurations are typically concentrated in regions of configuration space having a dimension d much smaller than that of the full configuration space. Outside these regions, the probability density decays rapidly[151]. These regions of space are referred to as the *intrinsic data manifold* and their local dimensionality is commonly called the *intrinsic dimension*[74]. To distinguish them, we will call the dimension D of the full space the *embedding dimension.* Importantly, the the localisation of data in manifolds of relatively small dimension is not a peculiarity of data sampled in molecular simulations, but are a general property observed in a huge variety of datasets[17]. For example, it has been observed that the computational search speed of nearest-neighbours in generic dataset scales with the ID rather than with the embedding D[20, 110].

The ID is a key player in all the approaches we used and developed. It can be estimated following various algorithms[27, 28], which, importantly, need to be able to cope with manifolds which are, in general, curved, twisted and topologically complex[17].

The ID quantifies the number of Degrees Of Fredom (DOFs) that are necessary to provide a complete description of the system without loosing any relevant information. It is therefore a crucial quantity, since the manipulation, analysis and representation of very high dimensional data is practically unfeasible and, beyond featurisation, an additional step of explicit *dimensional reduction* is generally required in practical applications. In principle, of course, a proper dimensional reduction would restrict the full space to a set of d variables. However, these d coordinates are typically nonlinear functions of the coordinates $\mathbf{s}(\mathbf{x})$. In the field of molecular simulations these coordinates are referred to as *collective variables* (CVs). We remark that if a manifold of intrinsic dimension d is topologically complex it is very difficult (if not impossible) finding an explicit expression for the d CVs that would describe it.

2.1.1 The free energy surface

When properly defined, a set of CVs is capable to capture all the relevant detail of the process under study without significant information loss. The reduced probability density is formally defined as the marginal of $\rho(\mathbf{x})$ with respect to all the CVs:

$$\rho(\tilde{\mathbf{s}}) = \int \rho(\mathbf{x})\, \delta(\tilde{\mathbf{s}} - \mathbf{s}(\mathbf{x}))\, \mathrm{d}\mathbf{x} \tag{2.3}$$

We can consider the restriction of configuration space $\Omega_{\tilde{\mathbf{s}}}$ where the $\mathbf{s}(\mathbf{x}) = \tilde{\mathbf{s}}$ as something in between a microstate and a macrostate of the canonical ensemble, also called a *mesostate*[89, 147, 171]: if the choice of CVs performs a dimensional reduction then the set of points identified by the CVs taking a specific vector of values $\tilde{\mathbf{s}}$ is extended rather than infinitesimal and cannot be regarded as a microstate; on the other hand, it is not guaranteed that a small multidimensional interval of the CVs $[\tilde{\mathbf{s}} + \delta\mathbf{s}]$ identifies a set of points with the dignity of macrostate, namely having well-defined macroscopic properties and a large thermalisation times outside the interval w.r.t. to the thermalisation time within it. As such, it can make formally sense to consider the partial free energy of a mesostate identified by a CV:

$$A_{\tilde{\mathbf{s}}} = -\,\beta^{-1} \ln \mathcal{Z}_{\tilde{\mathbf{s}}} := -\beta^{-1} \ln \int_{\Omega_{\tilde{\mathbf{s}}}} e^{-\beta U(\mathbf{x})}\, \mathrm{d}\mathbf{x} = -\beta^{-1} \ln \int_{\Omega_{\tilde{\mathbf{s}}}} \mathcal{Z}\, \rho(\mathbf{x})\, \mathrm{d}\mathbf{x} = -\beta^{-1} \ln \int_{\Omega_{\tilde{\mathbf{s}}}} \rho(\mathbf{x}) \mathrm{d}\mathbf{x} + f_c \tag{2.4}$$

where $\mathcal{Z}_{\tilde{\mathbf{s}}}$ is the configurational partition function restricted to the microstate $\Omega_{\tilde{\mathbf{s}}}$ and $f_c = -\beta^{-1} \ln \mathcal{Z}$ is an additive constant equal for all values of the CVs. Notice that, by definition of the CVs $\mathbf{s}(\mathbf{x})$, all the $\Omega_{\tilde{\mathbf{s}}}$ are disjoint sets and so the partial free energies $A_{\tilde{\mathbf{s}}}$ add up to the total Helmholtz free energy: $A = \sum_{\tilde{\mathbf{s}}} A_{\tilde{\mathbf{s}}}$. If the CVs identify slow DOFs of the system, a small interval of CV values can even happen to identify some macrostate of the system, in which case it can even make sense to distinguish internal energy and entropy contributions to the partial free energy[78, 126]. Anyhow, no matter this nomenclature, in the molecular simulations community the following quantity is generally considered:

$$F(\tilde{\mathbf{s}}) \;:=\; -\beta^{-1}\ln\rho(\tilde{\mathbf{s}}) \;=\; -\beta^{-1}\ln\int\rho(\mathbf{x})\,\delta(\tilde{\mathbf{s}}-\mathbf{s}(\mathbf{x}))\,\mathrm{d}\mathbf{x} \;=\; -\beta^{-1}\ln\int_{\Omega_{\tilde{\mathbf{s}}}}\rho(\mathbf{x})\,\mathrm{d}\mathbf{x}\ , \tag{2.5}$$

which differs from the actual free energy $A_{\tilde{\mathbf{s}}}$ only by the additive constant f_c. The collection of $F(\tilde{\mathbf{s}})$ for all the values of $\mathbf{s}$ the CVs describe a hypersurface, which is referred to as the *free energy surface* (FES) of the CVs.

The reduced probability density in equation (2.3) and the free energy surface, or simply free energy, in equation (2.4) will be the focus of our research throughout this thesis work. We will talk about estimating $\rho(\mathbf{x})$ and $F(\mathbf{x})$ where $\mathbf{x}$ represents any choice of coordinates as long as a metric is defined on them. In fact, all considered estimators are formulated in a completely general, distance-based way. Unless differently specified, in this work we will consider free energies in units of k_BT.

2.2 Multidimensional density estimation: nonparametric methods

As pointed out by Silverman[178], the need for nonparametric methods was first proposed in the famous pioneering work by Fix and Hodges[71] in the context of discriminatory analysis and has since then become very popular in many fields of machine learning[17, 74, 80, 170] and, notably, of density estimation[101, 170, 177]. Parametric methods assume a functional form for the probability distribution and focus on the optimisation of few parameters in order to best fit the observations. They are employed for a wide range of applications, ranging from classification to density estimation[17, 21, 58, 90, 114, 167]. However, while making it arguably less costly to return a smooth and well-behaved function, especially with few data, parametric methods require a much better knowledge of the underlying distribution. A wrongly specified model, i.e. one which does not capture some crucial features of the distribution, introduces in fact a bias that cannot be healed even in the case of large samples[21, 170]. There are many cases in which a simple parametric model which generalises the system's density is very difficult, if not impossible, to provide. Molecular simulations make no exception, since even the smallest and simplest molecules display metastability (multimodality in terms of p.d.f.) and often complex and rugged free energy landscapes.

In order to study all such systems, nonparametric methods are better suited. Nonparametric methods, in fact, make fewer and less rigid assumptions on the distribution and are rather fully driven by the data sample[177]. Such assumptions are typically some local properties[74]. Unlike what their name suggests, these methods do have indeed at least one parameter, which typically controls the level of smoothness[144]: we will refer to this as *smoothing parameter* or *bandwidth* or *scale parameter*, since it controls the scale at which the influence of a single datapoint decays. In his book[170], Scott discusses various possible ways to distinguish parametric from nonparametric density estimators. An estimator is commonly called nonparametric when it is consistent for a wide range of true density functions, a definition which does not mark a sharply clear boundary between the two classes. However, this criterion can be rephrased quantitatively in terms of locality[187]: an estimator $\hat{f}$ is called nonparametric if, given two different datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$, the influence of $\mathbf{x}_j$ on $\hat{f}(\mathbf{x}_i)$ vanishes asymptotically, i.e. in the limit of infinite statistics.

Nonparametric approaches, thanks to their characteristics, are ideal for the explorative study of some dataset distribution features, such the intrinsic dimensionality, topology and shape of the data manifold (the support of the distribution), monomodality, multimodality or glassiness, the presence of entalpic or entropic barriers[179], data structures like clusters and so on and so forth. However, aside from preliminary investigation of features, nonparametric methods are well-established also as self-standing learning methods. Thanks to their flexibility and the possibility to use them even with little or no a priori assumptions, nonparametric methods are are broadly adopted in the field of density estimation[80, 101]. Such flexibility comes at a cost of course: nonparametric methods are typically affected by a higher noise (variance) and potentially also bias, a drawback which can be exacerbated in high dimensions[74]. Since their prediction is based on local properties of the sample, all the interesting regions of the density domain must be populated in order to produce meaningful results. As we know, however, the task of sampling low-density regions is exponentially harder with increased dimensionality. A possible approach would be looking at data at a coarser scale, i.e. increasing the smoothing parameter. This however is a risky procedure, since it makes the model less informative and more parametric, although not supported by a model, which can introduce a severe bias. In order to exploit the advantages of nonparametric methods without incurring misleading results, a balance between sensitivity and statistics and resolution must be found: it is the so-called *bias-variance trade-off*, which is common to all nonparametric methods[80]. In the rest of the chapter we briefly review some of the main nonparametric density estimation methods. In particular we focus on the so-called counting methods[126], excluding from our treatment orthogonal series estimators[31], since they are often unstable and non-consistent and their application involve a high degree of technicalities[101]. In the following chapters we will show how many of the most common drawbacks affecting nonparametric estimators can be tackled efficiently in order to obtain good performance even in high dimensionality.

2.2.1 Histogram methods

The most basic, most common and oldest density estimation method is the histogram[177]. This approach consists of dividing the domain into non-overlapping regions called *bins* and counting the number of datapoints falling within each of these bins. If all the bins have the same size, then the shape of a suitably parametrised histogram is appoximately the same as that of the PDF. Thus, other than estimating the PDF, the histogram is a simple and powerful tool for representing and visualising the distribution and features of the data.

We shall present this method in its simplest one-dimensional version, in which the bins are segments and the smoothing parameter is a scalar h called binwidth. All definitions and nomenclature are easily generalisable to the multidimensional case (in the case of D dimensions, bins are not segments but D-dimensional volumes, and the binwidth can be expressed a D-dimensional vector if the grid spacing is different in different directions).

Let us assume N data points in one dimension $\{x_i\}_i$ lie on a segment $[a, b]$, which will be the support of our histogram. This segment is divided into M bins $(T_1, \ldots, T_M)$ of width h, where $T_1 = [a, a+h)$, $T_2 = [a+h, a+2h)$ and so on. The number of points falling in each bin constitutes the counts histogram $\{\hat{n}_m\}_{m=1}^{M}$. Defining I_S the indicator function of a segment S, such that $I_S(x) = 1$ if $x \in S$ and 0 otherwise, the number of counts in bin T_m is $\hat{n}_m = \sum_{i=1}^{N} I_{T_m}(x_i)$. By dividing the counts by the number of total observed points, one obtains the so-called frequency histogram $\{\hat{p}_m\}_m$, with:

$$\hat{p}_m := \frac{\hat{n}_m}{N} \ . \tag{2.6}$$

If the N datapoints are Independent and Identically Distributed (IID), then the bin counts are binomial random variables: $\hat{n}_m \sim B(N, p_m)$ such that

$$\langle \hat{p}_m \rangle = p_m := \int_{T_m} \rho(x) \, \mathrm{d}x \qquad \text{and} \qquad \mathrm{Var}[\hat{n}_m] = N \, p_m \, (1 - p_m) \ , \tag{2.7}$$

i.e. $\hat{p}_m$ is an estimator of the true probability to sample a datapoint in that bin (whose consistency can be shown)[170]. Looking at equation 2.7, one can approximate the PDF at the centre c_m of bin T_m assuming $\rho(c_m) \approx p_m/h$. In fact, what is generally done is to extend this approximations to all the points in the bin (making a slightly greater systematic error than in the case of bin centres), so that the density of a datapoint x_m falling in bin T_m is approximated as:

$$\rho(x_m) \approx p_m \, h \quad , \quad \forall \, x_m \in T_m \tag{2.8}$$

and thus, following from equations (2.6) and (2.7), it is estimated as:

$$\hat{\rho}(x_m) := \frac{\hat{p}_m}{h} = \frac{\hat{n}_m}{N \, h} \ , \tag{2.9}$$

while the uncertainty on such estimator can be expressed as the standard deviation:

$$\sigma[\hat{\rho}(x_m)] = \frac{\sqrt{\mathrm{Var}[\hat{n}_m]}}{N\,h} \approx \frac{\hat{\rho}(x_m)}{\sqrt{\hat{n}_m}} \ . \tag{2.10}$$

This setting corresponds to approximating the whole PDF as a step-wise function[101], called the frequency density histogram or area-normalised histogram, which can be shown to be a consistent estimator of the PDF[170]. Expression (2.9) can be rewritten using the indicator function of the bins:

$$\hat{\rho}(x) := \sum_m \frac{\hat{n}_m}{N\,h}\, I_{T_m}(x) \ . \tag{2.11}$$

Despite being consistent and at a the same time a very simple and easily implemented tool, the histogram is far from being the ideal choice in terms of both variance and bias. In fact, histograms are very sensitive to the choice of the smoothing parameter but also of the position of the bin centres. One can see from equation (2.10) that the variance can be reduced by increasing the binwidth, which has the effect of increasing the bin counts. However, increasing the smoothing parameter stretches the validity of approximation (2.8), introducing a bias which is greatest on the borders of the bins and obviously increases with the bin size. Imposing a step-wise behaviour to the estimator of a typically smooth function introduces artificial discontinuities in the PDF, makes it difficult to estimate derivatives and disfavours the intuitive contour plot visualisation. Moreover, for the same reason, the shape of a histogram can vary significantly with small adjustments in the domain origin[177]; the problem is intensified in the multidimensional case, where the degrees of freedom to adjust the grid orientation make the representation even more parametrisation-sensitive. Finally, another source of bias is the fact that histogram have finite support, namely they are non-zero only in regions where data have been sampled; while this has convergence advantages when the PDF is integrated, this tails behaviour might be pathological.

Besides the last one, all these problems could be addressed by choosing a denser discretisation of the domain, which however would reduce the number of data per bin and increase noise; this effect is more and more severe with increasing dimensionality, where populating the histogram requires exponentially larger samples the choice a small binwidth could produce even empty bins. A good balance must be found and, indeed, we are facing an emblematic example of bias-variance trade-off problem. By choosing a probability convergence criterion (e.g. minimisation of Mean Integrated Absolute Error (MIAE) or Mean Integrated Squared Error (MISE) or many possible others), it is possible to find the optimal value for the histogram binwidth[101, 170]. We will not get into the details of such discussion. Another possible way to get around some of these issues is to use variable partition histograms[101], in an attempt to make the smoothing parameter adaptive: instead of using a fixed bandwidth (or D-dimensional banwdidth vector in the multidimensional case) for all bins, this is varied locally. This requires choosing a sensible partitioning criterion, which can turn out to be challenging especially in high dimensions.

2.2.2 Kernel Methods

Due to its high versatility and simplicity, Kernel Density (KD) estimation is probably the most common approach to nonparametric density estimation both in the univariate and in the multivariate case and it can indeed be considered a prototype for all of them. Except for very simple tasks, it is preferred to histograms since it can outrival it both in terms of reducing parametrisation rigidities and in terms of smoothness.

One of the main drawbacks of histograms is the fixed nature of the domain partition: even when defining the bins adaptively, which is not a very common practice, the grid structure of the domain tassellation does not always suit the geometry of the problem: rare-data regions might result penalised to promote higher resolution in high-density regions or vice-versa. To overcome this rigidity, KD estimation considers only finite-size regions around the datapoints. The intuitive idea behind a Kernel Density Estimator (KDE), instead, is to assign a fixed amount of probability $p_N = 1/N$ to each point and to build the PDF by adding at the position of each sample point a function that normalises to p_N. The features of the class of functions chosen determine the quality and the properties of the estimator, such as smoothness, convergence and its capability to resolve or generalise from the data.

KDEs were first introduced by Fix and Hodges[71], who proposed what is now known as naive estimator[177]: in correspondence of each datapoint a box function of fixed size is added. In practice, if in one dimension we set a bandwidth h, the density at point x is computed as the number of sample points which fall in an interval $(x - \frac{h}{2}, x + \frac{h}{2})$ divided by h and N, much like in the histogram case (2.9):

$$\hat{\rho}_{\text{naive}}(x) = \frac{1}{N\,h}\left\{ \text{ number of sample points falling in } \left(x - \frac{h}{2}, x + \frac{h}{2}\right)\right\} \tag{2.12}$$

In the limit of infinite statistics $N \to \infty$ and vanishing smoothing parameter $h \to 0$, the probability $P_{x,h}$ that the random variable $y \sim \rho(y)$ takes values in $\left(x - \frac{h}{2}, x + \frac{h}{2}\right)$ is correctly estimated with this method:

$$\lim_{h\to 0} \hat{\rho}_{\text{naive}}(x) \;=\; \lim_{h\to 0} \frac{1}{h}\, P_{x,h} \;=\; \frac{\mathrm{d}}{\mathrm{d}x}\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} \rho(y)\,\mathrm{d}y \tag{2.13}$$

When all points lie on a regular grid and the size of the box is that of grid spacing, the naive estimator is equivalent to a histogram having such grid nodes as bin centres.

However, while eliminating the histogram problem of bin centring, the naive estimator is still discontinuous, since it is a step-wise function. Nonetheless, the problem can be easily tackled by allowing smoother functional forms than the box function, so the concept was soon formalised by Rosenblatt[161] (first appearence of KDE[178]) and Parzen[145], who generalised the method to a wider class of functions and studied convergence and asymptotic properties. The first extension to the multidimensional case was by Cacoullos[26]. Again, for simplicity and without loss of generality, we will present mostly the implicit reminder that all can be extended to the multidimensional case.

The intuitive view of KD estimation as a sum of N bumps of a given shape at the position of the sample points can be viewed in a somehow more formal way by thinking about the Dirac delta function as a convolutional kernel. If we knew the value of a PDF $\rho(x)$ for all values of x, from the definition of the Dirac delta we could write the value of the PDF at point x_0 as:

$$\rho(x_0) = \int \rho(x)\delta(x - x_0)\,\mathrm{d}x \; . \tag{2.14}$$

However, we can only observe a discrete a finite distribution of points, which we can represent as the sample density[74]:

$$\rho_s(x) = \frac{1}{N}\sum_{j=1}^{N} \delta(x - x_j) \tag{2.15}$$

Plugging $\rho_s(x)$ into equation (2.14) would give $1/N$ if $x_0 \equiv x_j$ for some sample point x_j and 0 otherwise. In order to be able to generalise from observed data, the delta function can be relaxed to a smoother kernel κ_h with width regulated by a finite scale parameter $h > 0$, but which still preserves the delta function normalisation property:

$$\int_{-\infty}^{\infty} \kappa_h(x - x_0)\,\mathrm{d}x = 1 \; . \tag{2.16}$$

The corresponding kernel density estimator, in one dimension, is defined as:

$$\hat{\rho}\left(x\right) := \frac{1}{N}\sum_{j}^{N} \kappa_h\left(x - x_j\right) \; . \tag{2.17}$$

Typically, these kernels are chosen to be radially symmetric probability densities, with $\kappa_h(x) \geq 0$ and $\kappa_h(x) = \kappa_h(-x)$; in this case expression (2.17) defines a bona fide PDF[101] and one recovers the Dirac delta in the limit of small bandwidth:

$$\lim_{h \to 0} \kappa_h(x - x_0) = \delta(x - x_0) \; . \tag{2.18}$$

If the chosen and kernel is differentiable, the estimated PDF inherits such property, making it smooth and giving KDEs competitive advantage over histogram.

Very commonly in literature the dependence on the smoothing parameter is not incorporated in the kernel function, but rather a kernel $K \equiv \kappa_{h=1}$ is used, so that the consequence of rescaling the smoothing parameter must be explicitly treated. We write an expression of a KDE in this form, this time directly in the multivariate case, so that the dependence on the dimensionality D is also evident:

$$\hat{\rho}\left(\mathbf{x}\right) := \frac{1}{N\,h^D}\sum_{j}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_j}{h}\right) \tag{2.19}$$

Expression (2.19) is still not the most general form for a multivariate KDE, since the chosen bandwidth h is treated as a scalar. However, the amount of smoothness can also be chosen not to be isotropic and can in general be represented by a $D \times D$ symmetric matrix H[170]; expression (2.19) is then recovered for $H = h\;\mathbb{1}_D$.

If the kernel κ_h is well behaved, the estimators in equation (2.17) and (2.19) are unbiased and consistent if $h \to 0$ as $n \to \infty$ when $n\,h^D \to \infty$. Asymptotic convergence is obtained regardless of the functional form of the kernel, as long as regularity conditions apply[101], although the shape and properties of κ_h do influence the speed of convergence. Nonetheless, while dimensionality effects are important, the shape of the kernel does not seem to affect convergence heavily[101]. This offers great flexibility, and allows to choose among a wealth of possibilities, some of which illustrated in the table:

Kernel shape	$K(x)$
Box	$\frac{1}{2}\,I_{[-1,1]}\,(x)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\,exp(-\frac{1}{2}x^2)$
Epanechnikov	$I_{[-1,1]}\,(x)\;\frac{3}{4}(1-x^2)$
Cosine	$I_{[-1,1]}\,(x)\;\frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right)$
Triangle	$I_{[-1,1]}\,(x)\;(1-\mid x\mid)$
Biweight	$I_{[-1,1]}\,(x)\;\frac{15}{16}\,(1-x^2)^2$

Table 2.1: Some of the most common kernels used in KDE[80, 177].

The simplest of all kernels above is the previously discussed box-shape kernel, i.e. the naive estimator. In the multidimensional case two are the most common options for the naive estimator: either taking the D-squared box, thus use the indicator function of the hypersquare of unit side centered in $\mathbf{x} \equiv \mathbf{0}$; or considering the D-ball of unit radius and zero origin $B^D(1, \mathbf{0})$ and thus using its indicator function $I_{B^D(1,\mathbf{0})}$ divided by the volume of the D-dimensional unit hypersphere ω_D (see equation (D.3)). The Gaussian KDE is by far the most used, both in the univariate and in the multivariate case, in which the degree of smoothing is controlled by the diagonal - or even dense - inverse covariance matrix. Except for the Gaussian, all other kernel in table 2.1 have compact support and are all subcases of a more general expression for polynomial kernels[101]. Worth a mention is the so-called Epanechnikov kernel, which has been shown to be optimal in terms of asymptotic MISE[65]. Notice that in the formulas presented so far we assumed data to lie in a Euclidean metric space, in which the concept of vector difference, distance, norm are those familiar to us. All the kernels in table 2.1 are symmetric w.r.t x, i.e. they only depend on the norm of $\|x\|$. Extending this concept, it is possible to use KDEs and estimate the PDF at given points even in non-Euclidean metrics simply knowing a distance between data point (e.g. on curved Riemannian manifolds[148]).

KDEs are thus a very simple and powerful tool, but they are also very general, since under weak assumptions both parametric and nonparametric methods can be proven, at least asymptotically, to be some sort of generalised kernel density estimator[187]. In the case of histogram the equivalence between equations (2.9) and (2.19) can be easily seen by applying the definition of $\hat{n}_m$. In what follows, we will point out similar analogies also while presenting the other methods.

As for all nonparametric methods, the crucial problem for KDEs is the selection of the bandwidth. In regions of high data density, a large value of h may lead to over-smoothing and a washing out of detail that might otherwise be extracted from the data. However, reducing the level of smoothing may lead to noisy estimates elsewhere in data space where the density is smaller, making the estimated PDF artificially rough and spiky. Once again, this problem is accentuated in high dimensions: the MISE can be decomposed into a variance component scaling as h^{-D} and a bias, going as $h^2\sigma_K^2 \sim h^{D+2}$, where σ_K^2 is the variance of the kernel function an typically goes as $\sim h^D$[37].

As mentioned by Silverman[178], the problem of bandwidth selection in KD estimation was first discussed even by Fix and Hodges in the original paper. Due to the popularity of KDEs, a great amount of research effort has been devoted to tackling this issue [92, 180, 191]. According to different possible optimality criteria, various optimal values for the choice of a unique bandwidth have been proposed[101]. A further improvement can be obtained by allowing the smoothing parameter to be coordinate-dependent, i.e. by making it adaptive. This, however, can require some prior knowledge of the system studied or some accurate modelling. A famous metod is the so-called variable kernel estimator[24], which relates the smoothing parameter at a point to the distance of its nearest neighbor of some order k. Other methods propose iterative procedures to refine the local choice of h[4, 5, 87, 177]; in this case, however, the performance of the estimator still depends on the smoothing parameter chosen in the first (blind) iteration.

2.2.3 k-Nearest Neighbours methods

A brilliant but very simple way to adaptively select the bandwidth of a KDE is adopted in the k-Nearest Neighbours (kNN) estimator. Instead of fixing the scale parameter h, the hyperparameter of kNN sets the number of neighbors to be considered for each point, which in turn determines the local level of smoothing. As for the KDE methods, also this other important class of nonparametric density estimation methods was firstly proposed in the same work by Fix and Hodges[71, 178]. It was then formalised and generalised to the multivariate case few years later[118], while in reference [120] bias and variance, tails behaviour and other theoretical properties of the estimator are discussed.

The first key ingredient to implement kNN is to compute the distances $\{ d(\mathbf{x}_i, \mathbf{x}_j) \}_{i,j}$ among all couples of datapoints. For every point $\mathbf{x}_i$ it is possible to rank all the N points in neighbouring order: the index i_0 will be exactly i, i_1 the index of the Nearest Neighbour (NN) of $\mathbf{x}_i$, i_2 the index of the second-nearest and so on until i_{N-1}, which is the furthest point from $\mathbf{x}_i$ in the database. Then one has to decide the number k of neighbours to be considered for each point. The distance of the kth neighbour, $r_{i,k} := d_{i,i_k}$, implicitly sets a typical scale around point $\mathbf{x}_i$: if data are dense around in that region, then $r_{i,k}$ will be small; if data are rare $r_{i,k}$ will be large. The selection of k defines a neighbourhood for each point and induces a directed weighted graph structure on the dataset, in which nodes are represented by sample points, an edge directed from point i to point j is established if j is in the neighbourhood of i (among i's first $k-1$ NN) and the weight of the edge equals the distance between the two connected nodes. We name this graph the Neighbours Graph (NG) of the dataset: it will be a crucial ingredient of all our work.

Hence, the kNN density estimator at point $\mathbf{x}_i$ is defined as:

$$\hat{\rho}_i := \frac{k}{N} \frac{1}{\omega_D \, r_{i,k}^D} =: \frac{k}{N \, V_{i,k}} \, , \tag{2.20}$$

where ω_D is the volume of the unit-radius hypersphere in D-dimensions, which enters the definition of the volume of the D-hypersphere of radius $r_{i,k}^D$ centered on $\mathbf{x}_i$: $V_{i,k} := \omega_D \, r_{i,k}^D$. Basically, kNN instead of counting the number of points within a ball of fixed radius h, fixes the number of points k to look at, measures the radius of the ball in D dimensions centered at a given point $\mathbf{x}_i$ containing no less and no more of k datapoints. Evidently, a convention must be chosen to define such ball: one could choose the largest possible, having as radius $r_{i,k}$, the distance of the closest excluded point $\mathbf{x}_{i_k}$; the smallest possible, of radius $r_{i,k-1}$ just as big as needed to include the $(k-1)$th neighbour of $\mathbf{x}_i$; or anywhere in between these two distances. We have adopted the first convention, which tends to slightly underestimate the PDF. However, like for all other conventions, the estimator is unbiased if $k \to \infty$ and $k/N \to 0$ while $N \to \infty$[101]. Concerning the hyperparameter k selection, as for all nonparametric methods bias-variance trade-off considerations apply; expression (2.20) implicitly considers the PDF constant over a region of radius $r_{i,k}$, so this condition – we call it Constant Density Assumption (CDA) – should not be strongly violated. Various authors suggest that the choice of k should scale with sample size and dimensionality as $k \propto N^{\,4/D+4}$[6].

Thanks to its adaptivity to the local sample density, the kNN density estimator performs well even in high dimensions and has has a smaller variance compared to a standard KDE[177]. However, kNN is not immune to the curse of dimensionality, since in very high dimensions the meaningfulness of a distance ranking fades away[19]: the difference between the distance of the kth NN and of the $(k+1)$th NN goes to 0 when $D \to \infty$, which can make the definition (2.20) unstable. Various attempts have been made to address this problem[132, 139, 157]. This will be discussed more in detail in the following chapter.

So far, we have defined the estimator in equation (2.20) only on points of the dataset. Given a generic point of coordinates $\mathbf{x}$, one can still measure the distances w.r.t all datapoints $\{\mathbf{x}_i\}_i$ and define the distance of the lth closest point to $\mathbf{x}$, of index $l(\mathbf{x})$ as, $r_l(\mathbf{x}) := d(\mathbf{x}, \mathbf{x}_{l(x)})$. Fixing a value for the number of neighbours k one obtains the function $r_k(\mathbf{x})$ which expresses the local value of the smoothing parameter. Thus, what sometimes is done in literature is to give a KDE expression that extends the kNN estimator to all the domain:

$$\hat{\rho}(\mathbf{x}) := \frac{1}{N\,(r_k(\mathbf{x}))^D} \sum_{i=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{r_k(\mathbf{x})}\right) \,, \tag{2.21}$$

where to recover expression (2.20) one should sit on a sample point $\mathbf{x}_i$ and take K to be a ball kernel; with any choice of $K(\mathbf{x})$, expression (2.21) is a subcase of the variable kernel estimator[24]. This continuum version of kNN, that we will call continuated kNN, has however several drawbacks. Firstly, $r_k(\mathbf{x})$ is defined everywhere, but is highly discontinuous, which is reflected in a discontinuity and non-differentiability of the estimator itself. Secondly, the tails of the estimated PDF are artificially fat, decaying as $\|\mathbf{x}\|^{-D}$, which makes the estimator ill-behaved for peripheral points of the dataset; in one dimension this estimator is not even normalisable[162]. Finally, notice that the continuated kNN estimator is not defined for $k < 2$, since it would not find any point within radius $r_1(\mathbf{x})$; the case of too-small k however is not interesting and should be avoided due to low statistical significance of the resulting estimates.

kNN is an excellent strategy to implement a simple adaptive method to estimate the density. It has the advantage of selecting a higher smoothing parameter in low density region, reducing variance, and a smaller one, retaining more detail, in data-rich regions. While standard KDEs tend to undersmooth the tails and oversmooth the peaks of the PDF, by selecting a global bandwidth for the whole dataset, kNN is known to oversmooth the tails, introducing a great bias[177]. Overall, we can conclude that, if uninterested in a smooth PDF, kNN is preferable to KDEs for the estimation of complex and/or multidimensional PDFs and performs better with a poorer statistics, although the curse of dimensionality affects it nontheless. Extrapolated or border estimates (at coordinates falling outside the core of the datasets) must be taken cautiously.

Chapter 3

Improving kNN-based density estimation

3.1 Overcoming some limitations of kNN

3.1.1 The importance of restricting to the relevant submanifold

In the first chapter we covered the topic of PDF estimation with particular focus of Euclidian spaces $\mathbb{R}^D$ of coordinates $\mathbf{x}$. As anticipated, these procedures do not differ formally from the case in which they are applied to a set of collective variables $\mathbf{s}(\mathbf{x})$ living in a possibly-non-Euclidean metric, as long as a metric is defined in this space. Besides being crucial in data manipulation and visualisation, in the field of density estimation restricting to a reduced coordinate space can be advantageous to cope with the curse of dimensionality and becomes in most cases imperative in order to obtain sensible results; moreover, in many enhanced sampling techniques data are generated via the exploration of a CVs space, which makes that space the natural one to compute quantity such as the reduced density (2.3) or the FES (2.5). However, the quality of these estimates is critically bound to the quality of dimensional reduction (DR). In the introduction we briefly mentioned on how the problem of DR is tackled in the literature, by finding an explicit mapping of the full coordinate space into a much smaller space of descriptors through linear or non-linear transformations $\mathbf{s}(\mathbf{x})$. This mapping can be specified manually, by exploiting chemical/physical/mechanistic insight into the system under consideration or can be pursued automatically via the available unsupervised DR techniques. Both manual and automatic procedures, or a combination of the two, have potentially serious drawbacks and introduce a remarkable layer of complexity to the task of density estimation and are therefore a possible source of bias.

However, an alternative is to never perform explicitly this transformation, with the only assumption that data lie on a Riemannian manifold Ω of known dimensionality d, which is embedded in the Euclidean space $\mathbb{R}^D$. If this hypothesis holds, since we are interested in nonparametric density estimation, local by definition, we can consider distances between points on the data manifold only at a scale at which it looks locally flat. In this regime, Ω at a given point $\mathbf{x}$ is well approximated by its tangent hyperplane $T_\Omega(\mathbf{x})$. Therefore, we are well justified to take our nonparametric estimators with support only on this $\mathbb{R}^d$ restriction, i.e. to be normalised considering only d dimensions. Nonetheless, since data are restrained on the manifold, in a neighbourhood of each datapoint $\mathbf{x}_i$ one can measure distances equivalently in the Euclidean $\mathbb{R}^D$ metric or in the $\mathbb{R}^d$ metric of $T_\Omega(\mathbf{x}_i)$: opting for the first one, we would never need to define an explicit chart of Ω.

This idea, originally proposed in reference[139] applied to KDEs and then further developed in reference[157] using kNN, is a key ingredient to help mitigating the curse of dimensionality. In fact, a data point $\mathbf{x}_i$ can be seen as adding a lump of probability density only on the relevant d dimensions, instead of diluting this information over a much larger number of irrelevant dimensions $D \gg d$. This approach relies on a good estimate of the local intrinsic dimensionality $d(\mathbf{x})$ of the manifold, which can be achieved with various possible techniques[28, 66, 68, 84]. Here and in what follows we assume for simplicity that d is constant over the whole dataset and estimate it with the TWO-NN approach described in reference [68]. This nonparametric and unsupevised ID estimator also allows to cope with complex topologies, since it only looks at the manifold locally rather than globally.

3.1.2 A likelihood model for kNN

In section 2.2.3 we introduced the kNN density estimator starting from a KDE perspective in which no explicit assumption was made on the distribution of the data. There was however an implicit assumption: the request that the bias of the estimator is not too high is equivalent to assuming that the PDF over a region of size selected by the chosen k is sufficiently flat or slowly-varying (as in the case of the histogram). This requirement can be formalised in a way that allows to obtain kNN as MLE[157], which is our first step towards further improvements to kNN.

Given a point $\mathbf{x}_i$, we indicate by $r_{i,j}$ the Euclidean distance between point i and its jth nearest neighbour. Importantly, we assume the existence of a data manifold of intrinsic dimension d to which data are softly confined, as discussed in section 3.1.1. Thus, Euclidean distances measured in $\mathbb{R}^D$ are the same as those measured in $\mathbb{R}^d$, which in turn, for neighbouring points of i, is approximately equivalent to the distance in the manifold metric. We also measure volumes in $\mathbb{R}^d$, so that the volume of the unit hypersphere is ω_d. Therefore, the volume of the hyperspherical shell between neighbours $j-1$ and j of a point i is given by $\nu_{i,j} = \omega_d(r_{i,j}^d - r_{i,j-1}^d)$, where $r_{i,0}$ is conventionally set to 0 (hence, the volume of the d-sphere enclosed between a point i and its k-nearest neighbour is $V_{i,k} := \sum_{j=1}^k \nu_{i,j} = \omega_d \, r_{i,k}^d$).

If we fix a number of neighbours k and consider the actual PDF ρ constant within a radius $r_{i,k}$

from point i, then it can be proven[68] that the hyperspherical shell volumes $\{\nu_{i,j}\}_j$ are IID random variables exponentially distributed with rate corresponding to the number density $N\rho$:

$$P\left(\nu_{i,j} \in [\nu, \nu + \mathrm{d}\nu]\right) \;=\; (N\rho)\, e^{-(N\rho)\,\nu}\, \mathrm{d}\nu \;. \tag{3.1}$$

We can therefore write a joint probability distribution for the k independent random variables:

$$f_\rho\left(\{\nu_{i,j}\}_j\right) \;=\; \prod_{j=1}^{k} f_\rho\left(\nu_{i,j}\right) \;:=\; \prod_{j=1}^{k} (N\rho)\, e^{-(N\rho)\,\nu_{i,j}} \;. \tag{3.2}$$

As we can see from equation (3.2), the random variable $\nu_{i,j}$ has expected value $(N\rho)^{-1}$; in fact, on average, in the volume of one shell we find one point: $\langle \nu_{i,j}\, N\rho \rangle = 1$.

We can encode this PDF into a log-likelihood for the observed shell volumes $\{\nu_{i,j}\}_j$ as a function of the parameter ρ by taking the logarithm of the right-hand side of equation (3.2):

$$\mathcal{L}_{i,k}(\rho) = \mathcal{L}(\,\rho \mid N, \{\nu_{i,j}\}_j) \;:=\; \sum_{j=1}^{k} \left(\ln(N\rho) - (N\rho)\nu_{i,j}\right) \;=\; k\log(N\rho) - (N\rho)V_{i,k} \;. \tag{3.3}$$

By maximising this log-likelihood w.r.t. ρ one obtains the solution $\hat{\rho}_i = k/(N\, V_{i,k})$, which is exactly the kNN estimator defined in (2.20). It is worth asking if such an estimator, obtained from a likelihood maximisation, can still be regarded as nonparametric; according to the definition by Scott[170] discussed in section2.2 it is indeed, since it is only locally parametric and the asymptotic limit $N \to \infty$ makes this local dependence more and more punctual: $r_k(\mathbf{x}) \to 0$.

Applying the asymptotic limit of the Cramér–Rao Bound (CRB)[51, 155] for the variance of the estimator $\hat{\rho}_i$ it is possible to estimate its statistical error in terms of standard deviation:

$$\varepsilon_{\hat{\rho}_i} \;:=\; \left\langle -\frac{\partial^2}{\partial\rho^2}\,\mathcal{L}\left(\,\rho \mid N, \{\nu_{i,j}\}_j\right) \right\rangle^{-\frac{1}{2}} \;=\; \frac{\hat{\rho}_i}{\sqrt{k}} \;, \tag{3.4}$$

which has the same expression of the histogram uncertainty (2.10). Since we know that $\hat{\rho}_i$ is an unbiased estimator and thus $\langle \varepsilon_{\hat{\rho}_i} \rangle = \rho/\sqrt{k}$, we can once again see an instance, even for kNN, of the bias-variance trade-off: the statistical error of the estimator decreases by increasing the smoothing parameter, but this makes the approximation on which equation (3.1) relies shakier, since the density in a ball of radius $r_{i,k}$ might become non-constant.

3.1.3 Adaptive neighbourhood size selection

As discussed in section 2.2.3, the kNN estimator is an adaptive version of a kernel density estimator, since a fixed k would select a small smoothing parameter in regions of high density end a large where statistics is scarce. However, as we have seen, it is not immune from the bias-variance trade-off problem, despite being less exposed to it. Luckily, there is room for further improvement, because the effect of k on variance can be partly decoupled from that on variance: while a low variance can still only be granted by a high statistics, requiring a large k, systematic error in kNN is introduced only when k selects a region over which the PDF is not slowly-varying; the selected bandwidth should be small where the PDF changes rapidly, but can be larger when the PDF varies at a slower pace, regardless of its absolute value. In Figure 3.1b one can see an example of this: in green a slowly varying low-density region, where the PDF remains constant on a large region: a decent statistic can be included in the estimate; in blue a constant density region is perturbed by a high density peak rising quite close to the central point: only few points can be included in the estimate; in red a high-density region, where very many points have been collected and despite r_k being quite small a big k can be selected.

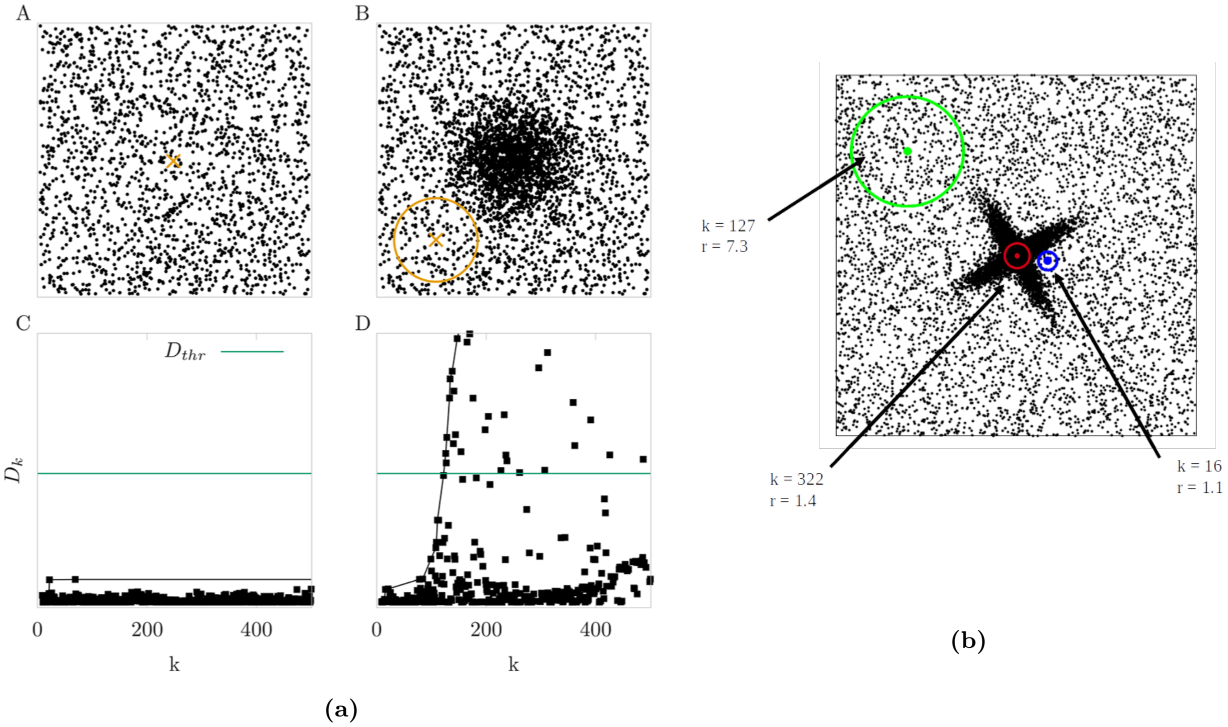


Figure 3.1: Adaptive selection of the optimal neighbourhood size $\hat{k}$ for the kNN estimator. **(a)** Schematic representation of the log-likelihood test in the case of two different bivariate PDFs. (A and B) Sample of 2000 points extracted from a uniform distribution and the same sample with 2000 additional points extracted from a Gaussian distribution. (C and D) D_k given in equation (3.6) as a function of k estimated for the two points highlighted in orange in panels A and B. The green line corresponds to the threshold D_{thr}. Image from reference [157]. **(b)** Optimal selected k and corresponding scale parameter r_k in different conditions in the case of a bivariate PDF. Sample of 2000 points extracted from a uniform distribution plus additional 2000 points extracted from a peaked distribution.

Based on this observation, Rodriguez et al.[157] proposed a quantitative procedure to optimise the largest possible value of k for which the PDF can be considered approximately constant within a certain level of confidence. To apply this procedure, given point i, we start with a low value for k and proceed iteratively by comparing two likelihood models: the first one ($M1$) in which the PDF estimated with kNN at point i and at its $(k+1)$th nearest neighbour, denoted by index l, have two different values ρ and ρ'; the second one ($M2$) in which i and l have the same kNN density ρ. In $M1$ the total log-likelihood is maximised with respect to two DOFs, ρ and ρ', while in $M2$ only one DOF, ρ, is allowed. These two maximised log-likelihoods, $\mathcal{L}_{M1}$ and $\mathcal{L}_{M2}$ are then compared: as long as they are similar within a certain confidence interval the putative optimal value for k is increased by one and the test is iterated; when the likelihood of model $M1$ utilises its extra DOF in the maximisation to become significantly larger than $\mathcal{L}_{M2}$ it means that the constant-density approximation does not hold for the k under test and so $k-1$ is adopted as optimal neighbourhood size $\hat{k}_i$ for the neighbourhood of point i.

Phrasing everything in mathematical terms, the optimal log-likelihoods of $M1$ and $M2$ are obtained (cf. (3.3)) as:

$$\mathcal{L}_{M1} := \max_{\rho,\rho'} \mathcal{L}_{i,k}(\rho) + \mathcal{L}_{l,k}(\rho') = k \log\left(\frac{k^2}{V_{i,k}\, V_{l,k}}\right) - 2k(1+\log N) \tag{3.5a}$$

$$\mathcal{L}_{M2} := \max_{\rho} \mathcal{L}_{i,k}(\rho) + \mathcal{L}_{l,k}(\rho) = 2k \log\left(\frac{2k}{V_{i,k} + V_{l,k}}\right) - 2k(1+\log N) \tag{3.5b}$$

These two likelihoods are then compared by a likelihood ratio test[134], which becomes a difference test in the case of log-likelihoods:

$$D_k = 2(\mathcal{L}_{M1} - \mathcal{L}_{M2}) = 2k\left(\log V_{i,k} + \log V_{l,k} - 2\log(V_{i,k}+V_{l,k}) + \log 4\right) . \tag{3.6}$$

The difference D_k between the two maximised log-likelihoods for a given k is distributed as a χ^2 RV with one DOF, since $M1$ has two parameters and $M2$ has one. We consider a p-value of 10^{-6} as significance threshold for the difference between the two models, corresponding to $D_{\text{thr}} = 23.928$. As long as $D_k < D_{\text{thr}}$ we consider the radius selected by the present k compatible with the constant density assumption around point i and we proceed to test the case of $k+1$ neighbours. When the inequality is violated the test fails and the previous value of k is retained. In brief, the condition for the optimal neighbourhood choice for point i is:

$$\hat{k}_i \quad \text{s.t.} \quad \forall\, k \le \hat{k}_i\ ,\ D_k < D_{\text{thr}} \quad \text{and} \quad D_{k+1} \ge D_{\text{thr}}\ . \tag{3.7}$$

The set of adaptively selected neighbourhood sizes for every point $\{\hat{k}_i\}_i$ can be used in all previous formulas of the kNN estimator to obtain a better-performing estimator, that we will name $\hat{k}$NN. This estimator is indeed doubly adaptive, since it builds adaptively on the already-adaptive kNN method. Although this may sound like a computationally costly procedure, by fixing a maximum number max$\hat{k}$ of explored values in the optimisation of k the estimation of the optimal $\hat{k}$ only scales linearly with the number of sample points N[79]. All kNN-based estimators discussed hereafter only involve operations on this sparse NG defined by fixing the set of $\{\hat{k}_i\}_i$.

In Figure 3.1a one can see the statistical test to select $\hat{k}_i$ at work. In the case of a uniform density D_k does not even get close to the threshold (in green); in the case of the Gaussian distribution superimposed on a uniform background the test is easily passed for all values until approximately 10^2, but then some values of D_k start shooting high and rapidly D_{thr} is reached. To visualise the rapidity of this triggering process in panels C and D the highest values of D_k recorded up to that specific k are connected by black solid line, which is horizontal in the first case and almost vertical in the second one. Tested in various conditions of statistics, dimensionality and underlying PDF, this neighbourhood selection procedure has proven quite consistent also relaxing the chosen significance threshold defining D_{thr} to larger values, even by two or three orders of magnitude corresponding to p-values of $10^{-4} \div 10^{-3}$.

As a side note, we point out that this whole procedure of neighbourhood size optimisation can be employed to address the problem of the bandwidth selection discussed in section 2.2 not only in the case of kNN-based methods, but also in the case of KDE. An application to Gaussian kernels is discussed in Appendix B.

3.2 PAk: a point-adaptive kNN-based density estimator

3.2.1 A free-energy formulation of $\hat{k}$NN estimator

Equation (3.3), its solution and the corresponding estimated error (3.4) can be rewritten in form of free energy. The free energy formulation of the estimator is more convenient to describe landscapes affected by metastability, where the PDF varies by orders of magnitude. To proceed, we apply the simple identity:

$$F(\mathbf{x}) := -\log[N\rho(\mathbf{x})] \ , \tag{3.8}$$

in the spirit of the definition of FES in equation (2.5), where β has been set to unity, so units of k_BT are assumed. Notice that, upon identification $\tilde{\mathbf{s}} \equiv \mathbf{x}$, the only difference between these two definitions is a constant factor depending only on N, which can be neglected.

By simply plugging this definition of F into equation (3.3) one obtains the log-likelihood of the volume shells around a given point i as a function of the free energy:

$$\mathcal{L}_i^{\hat{k}\mathrm{NN}}(F) = \mathcal{L}_i^{\hat{k}\mathrm{NN}}(F \mid \{\nu_{i,j}\}_j) := \sum_{j=1}^{\hat{k}_i} \left(-F - \nu_{i,j}\, e^{-F}\right) \tag{3.9}$$

which is maximised by:

$$\hat{F}_i := \operatorname*{argmax}_F \mathcal{L}_i^{\hat{k}\mathrm{NN}}(F) = -\ln \hat{k}_i + \ln \sum_l \nu_{i,j} = -\ln \frac{\hat{k}_i}{V_{i,\hat{k}_i}} \tag{3.10}$$

and has an error estimate $\varepsilon_{\hat{F}_i} = \hat{k}_i^{-\frac{1}{2}}$. So far we have just rewritten in free energy terms the upgrades to kNN introduced in section 3.1. The likelihood formulation provides a correct error estimate for the model and, as we will see, lays the foundation for further improvements to the model.

Before moving on to introduce the improvements, it is worth discussing which is the quantity that $\hat{F}$ actually estimates. So far, we have mentioned three levels of dimensional reduction (DR) (c.f.r. section 2.1): from the raw coordinate space to what we called the full coordinate space, via a choice of descriptors or featurisation $\mathbb{R}^{D_r} \ni \mathbf{x}_r \mapsto \mathbf{x} \in \mathbb{R}^D$; from this latter space to the CVs space Σ via an explicit DR; an implicit DR simply operated via the computation of the ID, from the space of the descriptors (or even of the CVs) to the intrinsic d-dimensional data manifold $\Omega \subset \mathbb{R}^D$ (or $\Omega \subset \Sigma$), approximated at each point $\mathbf{x}$ by its tangent hyperspace $T_\Omega(\mathbf{x})$ (or $T_\Omega(\mathbf{s})$), affine space of $\mathbb{R}^d$. Estimator (3.10) computes the negative log-PDF in the space it is applied. If such space is Σ, the estimated quantity is exactly the FES in equation (2.5). If however it is applied in the full coordinate space, $\hat{\rho}_i$ estimates unbiasedly the actual PDF $\rho(\mathbf{x}_i)$ for each point $\mathbf{x}_i$; therefore, in this case $\hat{F}$ estimates the potential energy of the configuration $\mathbf{x}_i$ entering the canonical Boltzmann factor in equation (2.4). Nonetheless, it can still make sense to refer to it as a "free energy", since for each point $\mathbf{x}_i$ the estimator considers an extended region Ω_i of volume $V_{i,\hat{k}_i}$ in configuration space, which cointains an infinite number of configurations. If the collection of all the $\{\Omega\}_i$ were a tassellation the concepts of internal energy $U_i = \langle U(\mathbf{x})\rangle_{\Omega_i}$ and Gibbs entropy $S_i = -k_B \langle \log \rho(\mathbf{x})\rangle_{\Omega_i}$ would be straightforward; unfortunately, kNN and related estimators do not consider a tiling of disjoint regions, but a set of regions with many overlaps, which would make it hard to give an explicit definition. Finally, a different reason why F can be regarded as a free energy is the fact that for every d-dimensional vector in the tangent space of the data manifold T_Ω, $(D-d)$ transverse directions are also explicitly included in $\mathbf{x}$: although their PDF is strongly suppressed outside Ω, those DOFs are formally averaged on, justifying the term.

3.2.2 Linear corrections to $\hat{k}$NN: introducing PAk

In reference [157] the author propose to modify the $\hat{k}$NN likelihood in order to account for possible corrections to the assumption that the PDF (and likewise the free energy) around point i is constant. They suggest that, moving away form i, F might be allowed to vary linearly in the neighbour rank j with a slope a, tweaking equation (3.9) into:

$$\begin{aligned}\mathcal{L}_i^{\mathrm{PA}k}(F,a) \;=\; \mathcal{L}_i^{\mathrm{PA}k}\left(F,a \mid \{\nu_{i,j}\}_j\right) \;:=& \sum_{j=1}^{\hat{k}_i}\left(-F+aj-\nu_{i,j}\, e^{-F+aj}\right)\\ =& \;-F\hat{k}_i+a\frac{\hat{k}_i(\hat{k}_i+1)}{2}-\sum_{j=1}^{\hat{k}_i}\nu_{i,j}\, e^{-F+aj}\end{aligned} \tag{3.11}$$

where a is treated as a variational parameter and thus it is maximised over both F and a:

$$\hat{F}_i \;:=\; \underset{F}{\operatorname{argmax}}\;\max_{a}\;\mathcal{L}_i^{\mathrm{PA}k}\left(F,a \mid \{\nu_{i,j}\}_j\right)\;. \tag{3.12}$$

Notice that expression (3.12) has no closed-form solution and must be solved iteratively, e.g. using the Newton-Raphson method[79].

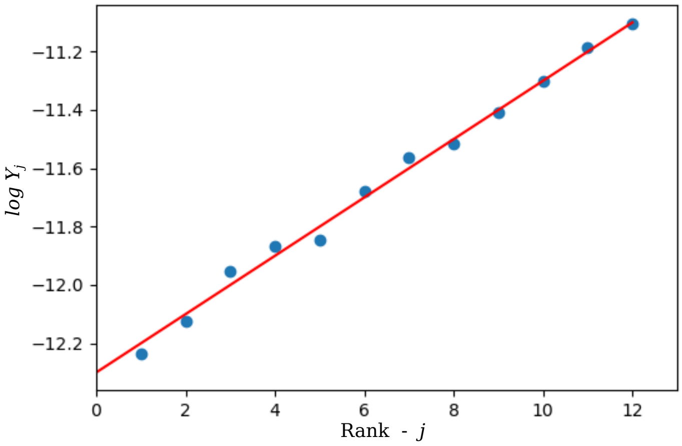


Figure 3.2: Pictorial illustration of PAk's log-linear fitting procedure. AAAAA spiegare meglio e modificare immagine, in ordinata ci va il logaritmo

As pointed out in reference [29], for each point i this maximisation is equivalent to the solution of a log-linear regression model[133] in which the j observed responses are the random variables $\mathbf{Y}^i = \left\{Y_j^i\right\}_j := \left\{\frac{1}{\nu_{i,j}}\right\}_j$ distributed exponentially with expected value $\langle Y_j^i\rangle = e^{-F_i+aj}$. Therefore, F_i is the intercept of such model and its value is equivalent to taking the limit $j \to 0$ or, analogously, sending the smoothing parameter of the model $r_{i,\hat{k}_i} \to 0$. In other words, this fitting procedure makes PAK less influenced by local fluctuations of the neighbours, thus making its estimates at neighbouring points more uncorrelated. We call this feature the *punctuality* of PAk.

The maximum likelihood approach also allows estimating the uncertainty of the MLE $\hat{F}_i$, again in the CRB setting[51, 155], which gives:

$$\varepsilon_i^{\mathrm{PA}k} = \sqrt{\frac{4\hat{k}_i + 2}{(\hat{k}_i - 1)\hat{k}_i}} \ , \tag{3.13}$$

where the differences w.r.t. the $\hat{k}$NN case are given by the presence of variational parameter a.

3.2.2.1 Using PAk to analyse biased trajectories

The PAk estimator allows estimating the free energy of a set of data points harvested from a multidimensional probability density. In molecular dynamics simulations of systems affected by metastability, one often biases the dynamics by an external potential, which is built in order to enhance the probability to observe transitions between the metastable states in a short simulation time. For all kernel methods, the standard reweighting schemes require subtracting the local exponential average of the biasing potential over the neighbourhood, which can be very noisy. In reference [29], whose results are presented in Chapter 5, we show how this problem can be overcome, at least if the bias is not time-dependent, thanks to the PAk's punctuality, which nontrivially allows the definition of a punctual reweighting scheme.

3.2.3 Estimator performance

3.2.3.1 Validation of estimators

In this thesis work we will face multiple times the necessity of assessing the efficiency of estimators and of their error predictions. In order to do so, we will mainly look at the observables described in what follows. The name of the quantities considered is F since we are mainly interested in free energy estimators, but these performance quantifiers are quite generally applicable to any estimator.

3.2.3.1.1 L_1 error or absolute error. Quantifies the local deviation of the estimator at a point $\hat{F}_i$ from the true value F_i. For example, in the case of free energies, plotted as a function of the true values of F will provide insight of the estimator performance in the various regimes of sample point densities. Some estimators might for example perform better than others where the statistic is high, but underprerform when it is low. The L_1 is defined:

$$\epsilon_i^{L_1} = |F_i - \hat{F}_i| \ . \tag{3.14}$$

3.2.3.1.2 Average L1 error or absolute error. Quantifies the global accuracy of an estimator and is defined as the sample average of the observable in equation (3.14):

$$\mathcal{E}^{L_1} = \frac{1}{N} \sum_{i=1}^{N} \epsilon_i^{L_1} \tag{3.15}$$

3.2.3.1.3 Pull variable[57]. It is used to compare two different error-affected estimators $\hat{F}^a$ and $\hat{F}^b$ of the same quantity F. It is defined as:

$$\chi_i := \frac{(\hat{F}_i^a - \hat{F}_i^b)}{\sqrt{\varepsilon^2_{\hat{F}_i^a} + \varepsilon^2_{\hat{F}_i^b}}} \ , \tag{3.16}$$

where ε_F indicates the uncertainty on the quantity F. If $\hat{F}^a$ and $\hat{F}^b$ are compatible and independent from one another, the distribution of the pull over a full statistical sample $\{\mathbf{x}_i\}_i$ is expected to be a Gaussian with zero average and unitary variance: $\chi_i \sim \mathcal{N}(0,1)$. One of the two estimators, say $\hat{F}^a$ can as well be substituted by the true values F_i, in which case the corresponding errors $\varepsilon_{\hat{F}_i^a}$ are zero. This is used to test the performance of the remaining estimator and of its error. In this case, the shift of the pull distribution from zero accounts for biasedeness of $\hat{F}^b$; variance of the distribution instead quantifies error accuracy: if the distribution is too spread it means the error is underestimated (or that sample points are correlated), if it is too narrow it means it is overestimated.

3.2.3.2 PA*k* performance

Reporting the results about PA*k* validation from reference [157], which introduced it, we evaluate PA*k* with three tools: the correlation plots between ground truth and predicted free energies for every point, $\hat{F}_i$ vs F_i; the pull distribution of the estimated free energies, according to equation (3.16), giving $\chi_i := \frac{F_i - \hat{F}_i}{\varepsilon_i}$, which should be distributed like a standard normal $\mathcal{N}(0,1)$; the mean absolute error of the estimator in equation (3.15), which reads $\mathcal{E}^{L_1} = \frac{1}{N}\sum_{i=1}^{N} |F_i - \hat{F}_i|$.

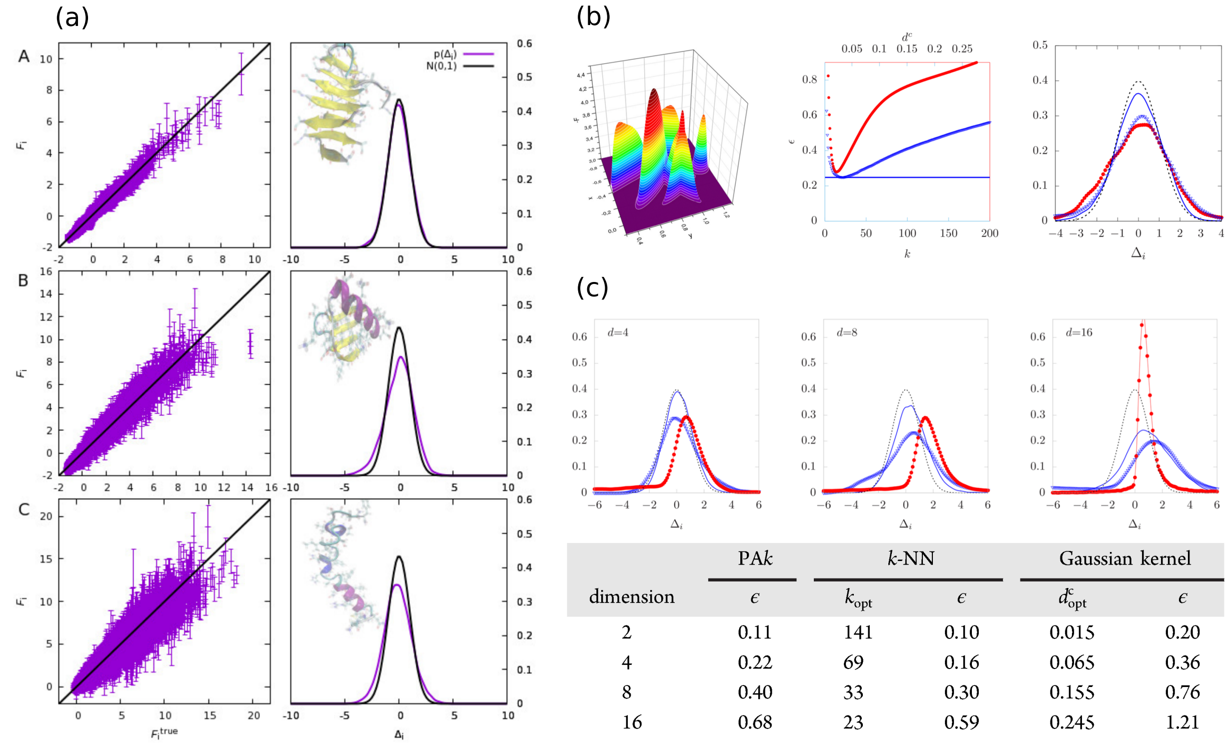
The tests assessing the performance of PA*k* estimator and comparing it to standard *k*NN and to Gaussian KDE are presented in Figure 3.3. By looking at correlation plots, pull distributions and absolute errors we see that PA*k* and its error estimator are unbiased and accurate in a range of dimensionalities d ranging from 2 to 8. Looking at the table in panel (c), reporting the mean absolute errors of the three estimators on synthetic datasets with d from 2 to 16, we see that PA*k* slightly underperforms *k*NN with k_{opt} selected in order to minimise $\mathcal{E}^{L_1}$ with respect to the analytical ground truth free energy; however, PA*k* always gives very similar results, despite the fact that its optimisation in fully unsupervised; the choice of k_{opt} is instead supervised: this optimisation could not be carried out in realistic settings where the true free energy is unknown; it is very likely that fixing k arbitrarily kNN would underperform PAk. Concerning the Gaussian KDE, PAk always beats it also in terms of absolute error, despite the fact that d^c_{opt} is optimised supervisedly.

In panels (b) and (c) we see that the pulls obtained with PAk are better than with the other estimators. In the leftmost pull distribution, in $d = 16$, we see that all free energy estimators are biased. This is an evident manifestation of the COD, which causes all points to be far away from each other and thus forces the estimators to ignore a lot of detail present in the analytical potentials. However, this might also be a problem of this specific dataset; furthermore, not in all applications a rigorous detail is required in the reconstruction of the FES; indeed, PAk has been succesfully used in practical applications on real datasets with intrinsic dimensionalities greater than 8: it has routinely been used on datasets whith ID between 8 and 15[29, 157, 179], but it has also returned biochemically meaningful results when applied on a huge realistic systems of ID between 26 and 28[30], as we will discuss in Chapter 4. We will also show in Chapter 6 that PAk gives stikingly good results on a synthetic dataset of ID 9. Importantly, PAk always displays the best pull distribution among the three estimators, a sign of its error estimator robustness.

For a comprehensive discussion on PAk performance we refer the reader to reference [157]. Thereby, PAk free energies estimated on the full coordinate space without explicit dimensional reductions, are shown to fairly agree with those computed in a space of collective variables specifically designed for an optimal description of the system. From here to the end of this chapter we will give an overview on PAk's stregths and weaknesses. Wherever our claims might appear not sufficiently supported by evidence so far provided, they will emerge in all the next chapters, when PAk will be shown at work, both when used in applications and when compared to other estimators.

3.2.4 PAk in a nutshell

PAk is a non-parametric unsupervised estimator which generalises the kNN approach. The first improvement introduced is conceptual and can be applied in principle to the whole class of generalised kernel methods[139] using only distances and not coordinates, including kNN-based algorithms[157]: the assumption that the estimates should be restricted to the tangent hyperplane of the intrinsic data manifold. By considering only such space of dimensionality d much lower than the embedding space dimension D an implicit DR is performed, which greatly alleviates the curse of dimensionality and grants sensible results even if D is very high. Formally, the intrinsic manifold's tangent hyperplane $T_\Omega(\mathbf{x})$ is coordinate-dependent; however, its parametrisation hould not be defined explicitly: as long as the ID of the dataset is computed accurately and the selected bandwidth $r_{i,\hat{k}_i}$ is such that within such distance from point i the PDF varies slowly enough, distances can be measured in $\mathbb{R}^D$ and volumes in $\mathbb{R}^d$; this allows to perform a significant DR without ever defining any CV explicitly.



	PA*k*	*k*-NN		Gaussian kernel	
dimension	ϵ	k_{opt}	ϵ	d^c_{opt}	ϵ
2	0.11	141	0.10	0.015	0.20
4	0.22	69	0.16	0.065	0.36
8	0.40	33	0.30	0.155	0.76
16	0.68	23	0.59	0.245	1.21

Figure 3.3: Performance of PA*k* estimator. All images are taken from reference [157] **(a)** Correlation plots and pull distributions of the free energy estimator on three test systems, A, B and C, having ID $d = 2, 4, 7$ and described in Appendices A.8.1, A.8.2 and A.8.3 respectively. In purple observed data, in black theoretical behaviour **(b)** Test on a the bidimensional artificial dataset whose negative free energy is represented on the left; three estimation methods are employed: fixed Gaussian kernel (red dots), standard kNN (blue triangles), and PAk (blue solid line). Center: average L_1 error $\mathcal{E}^{L_1}$ varying the chosen k for the kNN and of the Gaussian KDE smoothing parameter h, whose values are indicated on the bottom and top x axis respectively. Right: pull distributions for the three estimators; for kNN and the Gaussian KDE the values of k and h chosen are those who minimise $\mathcal{E}^{L_1}$ on the dataset and are labelled k_{opt} and d^c_{opt} respectively; the blue dashed line is the standard Gaussian; the error on the Gaussian KDE estimates are estimated by bootstrapping[63]. **(c)** Test on four artificial data sets of different dimensioality $D = d$; each dataset is composed by four d-variate Gaussians of different heights and variances and are described in the Supporting Information of reference [157]; top row: pull distributions computed for three different density estimators (same methods and colour code as panel (b)) on three systems ($d = 4, 8, 16$); bottom row: table reporting the average absolute error $\mathcal{E}^{L_1}$ (labelled ϵ) for the three methods and the optimal values for k_{opt} and d^c_{opt}.

Secondly, PAk original paper[157] introduces a likelihood formulation for the kNN class of methods, from which estimators of the PDF and of the free energy can be derived as MLEs. This formulation provides a natural error estimate. In its probability density phrasing, the kNN likelihood allows to define a procedure for the determination of the local bandwidth, as we will recap in the next paragraph. Yet, the – completely equivalent – free energy formulation paves the way for further developments to kNN, starting with PAk and continuing through the method that will be introduced in Chapter 6. Despite their formulation in terms of free energy, all estimator that will be discussed can in principle be used in any context where a PDF should be computed, due to the simple relation (3.8); since the estimators of the free energy neglect a constant additive term, in order to guarantee a normalisation to unity a known reference PDF value $\rho(\mathbf{x}_{\mathrm{ref}})$ should be adopted; alternatively, as mentioned in section 3.2.1, a tassellation of the domain $\{\Omega_i\}_i$ should be provided, so that the summation $\sum_i \hat{\rho}_i V_{\Omega_i}$ is computable and can be set to one. Despite the likelihood maximisation procedure might apparently characterise all these approaches as parametric, their local nature, as already discussed, make them fully-fledged nonprametric methods. Moreover, in reference [29] (and in Chapter 5) we argue that, via the neighbour-order-dependent regression achieved by (3.12), PAk takes the locality aspect one step further, defining a *punctual estimator*.

Thirdly, reference [157] introduces an unsupervised procedure to optimise independently for each datapoint the maximum number of nearest neighbours that can be included in the estimate without introducing systematic errors. This step, which can be applied to all kNN-based methods and not only to PAk, improves greatly the adaptivity of the estimators, making them more accurate and robust. Even if some small biases were induced at this stage of the procedure, we expect them to be healed by the log-likelihood maximisation step; what distinguishes PAk is in fact its two-step approach: the adaptive neighbourhood selection followed by the maximisation of PAk likelihood using, besides the free energy F, the additional variational parameter a accounting for deviations from the constant density assumption.

Tested on various systems, PAk has been proven an unbiased punctual free energy estimator, outperforming other non-parametric and unsupervised methods. The competitive advantage is especially visible in high dimensionality and with modest-size samples. PAk builds on the well-known kNN density estimation procedure, thus it retains the same benefits as its parent method. In terms of accuracy it outperforms it while paying a very reasonable price in terms of simplicity and computational scalability. Arguably, all improvements introduced with PAk have to do with enhancing the point-adaptiveness of pre-existing methods. Although our focus will be shifted occasionally from one PAk feature to another, we believe that all of the ingredients of PAk play a crucial role in determining its performance.

3.2.5 Survey on PAk's drawbacks

In this section we briefly outline some of the drawbacks and problems of the PAk estimators. These drawbacks motivated our research work, and this thesis is an attempt to address some of them.

3.2.5.1 Pak is not defined other than for sample points

The PAk estimator provides free energy estimates at all points in a given sample. However, some applications might require estimating the free energy even in other points of configuration space, which do not belong to the sample. This calls for the definition of an interpolation scheme, which

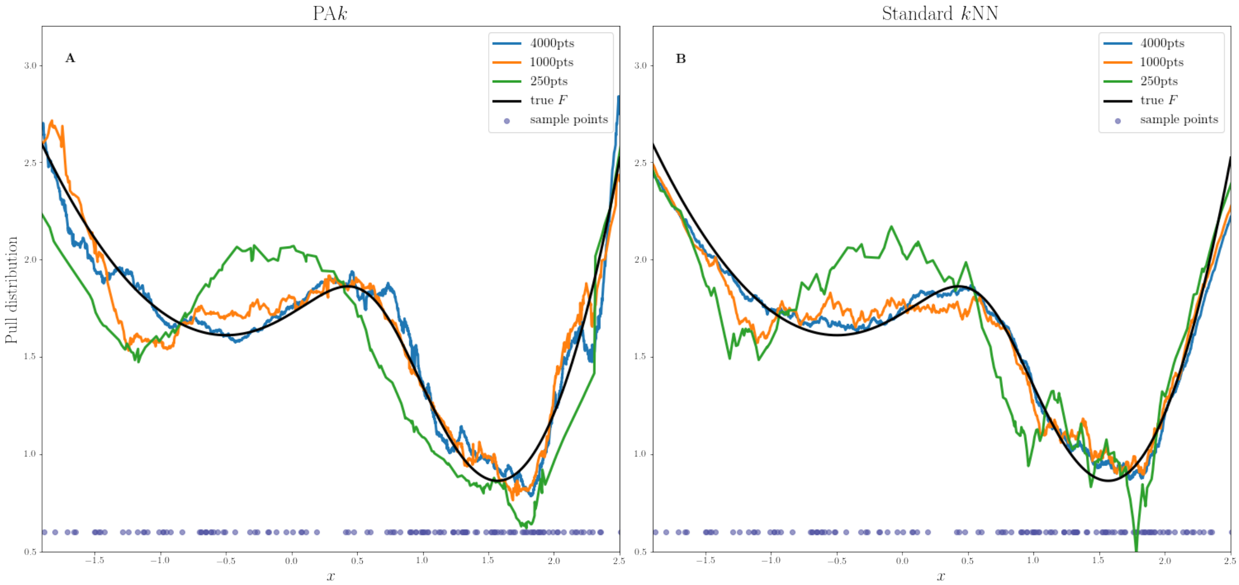


Figure 3.4: Pictorial illustration of PA*k*'s correlation-induced roughness as compared to *k*NN. In both panels performance of PA*k* (A) and standard *k*NN (B) on a 1-dimensional potential (cfr. Appendix A.1 for the functional form) for various sample sizes. In blue samples of 4000 points, in yellow 1000, in green 250. The *k* of *k*NN is chosen as 1/10 of the sample size. In violet the configurations of a decimated sample (125 points).

allows estimating the free energy in a point of generic coordinates, and not only on the data points used to estimate it. We propose a possible solution to this problem in section 3.2.6.

3.2.5.2 Curse of dimensionality

The PA*k* approach makes a progress with respect to other nonparametric methods in alleviating the curse of dimensionality. As mentioned above, it improves its accuracy by restricting, without explicitly defining it, to the low-dimensional intrinsic data manifold and by featuring a doubly-adaptive bandwidth selection. However, since PA*k* is a nonparametric method, and does not provide a model for the free energy surface it is still strongly affected by the COD. As we will see in Chapter 6, e.g. in Figure 6.7 and in the discussion in section 6.3.2.2.1, the linear corrections in the log-likelihood in equation (3.11) which define PA*k* starting from the *k*NN log-likelihood in equation (3.9) plays a key role in guaranteeing accurate results even at very high free energy values, where datapoints are rare. However, if the statistic is too poor w.r.t. what the high dimensionality would require, the performance of PA*k* are not great, as we will see in some specific examples. The development of improved, possibly partly-parametric, models, more robust to the COD than PA*k*, has been a first focus of our research.

3.2.5.3 Spurious correlations and roughness

The kNN-based density estimation methods consider every point on average $\langle\hat{k}\rangle$ times. This causes local fluctuations of the estimators around the ground truth distribution to be amplified and propagated at a range corresponding to the local selected k. To put it differently, if the free energy estimates $\hat{F}$ at different points were all independent, we would observe a white noise around the ground truth value; instead, estimates at a given point i consider in their computation the configurations of all the neighbours of i and vice-versa, so that the free energy of neighbouring points result correlated (even if they are considered independent by the model in equation (3.9) and we observe a correlated noise. In 1 dimension, for example, this emerges as the mid-frequency undulating behaviour visible in Figure 3.2.5.3 in both panels (representing PAk in panel A and standard kNN in panel B). This problem, phrased in slightly different terms, has been known since long ago[177]. It is of course a problem of all kernel methods, but it affects more severely approaches whose kernels do not have fast-decaying tails, as e.g. Gaussian KDEs do. All kernel methods are consistent[101], so simply kernel with fattest tails will have the related estimator converge more slowly. In practice, for all methods this noise is healed by statistics. In the case of kNN, moreover, the limit $N \to \infty$ automatically takes the limit $h \to 0$, since k is fixed. By looking at panel B of Figure 3.2.5.3 we can see the estimates of kNN estimator on a 1-dimensional potential for various sample sizes n (with k chosen as $N/10$). In green we observe the smallest sample, of 250 points, which has big oscillations. Noise is rapidly absorbed when the sample size goes to 1000 (in yellow) and then 4000 (in blue).

PAk, which is also based on kNN, has an additional factor that influences the noise in its estimates. The free energies F of the neighbours of a point i entering kNN likelihood in equation (3.9) as parameters are corrected linearly in PAk likelihood in equation (3.11) as $F - a\,j$, where j is the neighbourhood rank and a is a scalar which is the same for all ranks j. These terms, evidently, do not take into account any spatial information on where the jth neighbour of i is located w.r.t. i: they are linear in rank space rather than in configuration space. For example, in 1 dimension the model does not know if a neighbour j is on the left or on the right of i. This causes the MLE to fit locally, by looking at it in configuration space rather than in neighbourhood rank space, assuming a sort of cusp-like (or hypercone-like) free energy profiles in order to extrapolate the value for $\hat{F}_i$. This fact introduces spurious noise and mid-range ripples which remain visible even when the statistic improves, while other nonparametric methods, such as standard kNN, converge more rapidly. We can see it by looking at Figure 3.2.5.3: while with 4000 points (in blue) kNN (in panel B) has managed to smoothen and damp the noise present at lower statistic, PAk, in blue, still maintains a spiky behaviour. For future reference, we refer to this feature as PAk's intrinsic roughness. This undesired feature, however, as we have seen and will see also in the next chapters, does not affect

PAk unbiasedness and statistical robustness. In Chaapter 6 we will describe a possible stratedy to mitigate this problem.

3.2.6 Computing the free energy in a generic point: the PAk interpolator

A great amount of approaches for interpolating are available in the literature[8, 77], from the basic nearest neighbour interpolation, to inverse distance weighting, to Gaussian KDE to the use of radial basis functions. These methods are only effective when the dimensionality is low, but are rapidly affected by the COD.

We propose an approach that applies the PAk scheme to define an interpolator which works also in high dimensional spaces. The procedure is straightforward: for every point ξ in which we want to compute the interpolated free energy we compute the distances to all the points in the sample $\{r_{\xi,i}\}_i$ and we rank the neighbours i of ξ from 1 to N. Then we estimate the optimal neighbourhood size $\hat{k}_\xi$ as described in section 3.1.3, but with an important difference: the kNN number-density estimates which enter the log-likelihoods in equations (3.5) for the likelihood ratio test should be $\rho_\xi := k/\omega_d \; r^d_{\xi,k+1} =: k/V_{\xi,k}$ rather than the usual definition of $V_{\xi,k}$ as the distance to the kth neighbour; this accounts for the fact that ξ is not a point of the sample: in other words, the biggest volume shell centred on $\mathbf{x}_\xi$ that contains only one point is that of radius $r_{\xi,2}$. Once the optimal number of neighbours for point ξ is found, PAk likelihood is defined as in the standard PAk by equation (3.11), but considering that this time $\nu_{\xi,1} := \omega_d \; r^d_{\xi,2}$ and that for $j > 1$ $\nu_{\xi,j} := \omega_d \; (r^d_{\xi,j+1} - r^d_{\xi,j})$. This likelihood can be maximised, returning the maximum-likelihood estimate of the interpolated free energy at point ξ: $\hat{F}_\xi$. This procedure, introduced by us in reference[29] has been proven to define an unbiased interpolator for the free energy at configurations not belonging to the data sample in various applications[204]. We call it the PAk interpolator. PAk interpolator has the same features as its parent method PAk, discussed in section 3.2.4. The computational cost of computing the interpolated free energy with this method is the same as a simple PAk estimate.

Chapter 4

An application to the SARS-CoV-2 Main Protease

In this chapter we present a project in which, making use of the PAk estimator presented in Chapter 3, we characterise the metastable states of SARS-CoV-2 Main Protease to propose potential druggable pockets. This work is published with the title *Candidate Binding Sites for Allosteric Inhibition of the SARS-CoV-2 Main Protease from the Analysis of Large-Scale Molecular Dynamics Simulations*[30]. This application allows illustrating the advantages of the PAk estimator in a real-worls application, but also its drawbacks, posing the basis for future developments.

We analyse a $100\mu s$ Molecular Dynamics (MD) trajectory of the SARS-CoV-2 Main Protease with the purpose of explicitly characterising and describing these metastable states. In some of these configurations, the catalytic dyad is less accessible. Stabilising them by a suitable binder could lead to an inhibition of the enzymatic activity. The idea that motivates this analysis is exploring the viability of allosteric inhibition. Based on a characterisation of global and local properties of the states of the molecule, we are able to propose a few possible targets which could serve as binding sites for drug-like compounds with the purpose of allosteric inhibition.

The core of all this analysis procedure is a pipeline of unsupervised methods which combines: an accurate ID estimate[68]; the PAk estimator (introduced in section 3.2.2 of Chapter 3), which allows characterising a free energy landscape as a simultaneous function of hundreds of variables; a Density Peak algorithm to find configuration clusters based on their FES[61, 158]. These tools allow us to identify several conformations that, when visited by the dynamics, are stable for several hundred nanoseconds.

4.1 SARS-CoV-2 Main Protease and its inhibition

The severe acute respiratory syndrome, which has broken out in December 2019 (COVID-19), is caused by coronavirus 2 (SARS-CoV-2)[199, 208]. Its main protease ($\mathrm{M^{pro}}$ or $\mathrm{3CL^{pro}}$) was the first protein of SARS-CoV-2 to be crystallised, in complex with a covalent inhibitor, in January 2020[103]. It is essential in the viral life cycle since it operates at least eleven cleavage sites on large viral polyproteins that are required for replication and transcription[103, 207], so it is an attractive target for the design of antiviral drugs[153]. Since there is no known human protease having a cleavage specificity similar to the one of $\mathrm{M^{pro}}$, it may be possible to design molecules that do not interact with human enzymes[103, 207].

$\mathrm{M^{pro}}$ is a homodimer. Each monomer has 306 residues and is composed of three domains. Domains I and II (residues 10-99 and 100-182, respectively) have an antiparallel β-barrel structure. The binding site of the substrate is enclosed between these β-sheets[207]. Domain III (residues 198-303) contains five α-helices and has a role in the regulation of the protein dimerization[207]. The two residues His^{41} and Cys^{145} form the catalytic dyad. The structure and way of functioning of the SARS-CoV-2 $\mathrm{M^{pro}}$ are similar to the ones of the SARS-CoV $\mathrm{M^{pro}}$[10, 201]. This is expected, due to a 96% sequence identity between them.

The most direct strategy to block the action of the $\mathrm{M^{pro}}$ is through small molecules that directly interact with the catalytic site. The first *in silico* trials were made with covalent inhibitors known to be interacting with the catalytic site of SARS-CoV $\mathrm{M^{pro}}$ such as N3[103] or 11r[207]. Many efforts followed in the field of virtual screening. In this kind of studies, computational docking of millions of molecules is performed, the behaviour of the best candidates is usually then tested through MD simulation [38, 76, 102, 106, 128, 141].

Another possible route that can be followed to stop the action of the $\mathrm{M^{pro}}$, is allosteric inhibition[49, 119]. The functional definition of allosteric regulation implies the energetic coupling between two binding events[69, 129]. The binding of the allosteric ligands affect orthosteric pockets by altering protein dynamics, either through large-scale structural changes or through more subtle changes in correlated residue motions [136, 197]. Following the idea of conformational selection[198], allosteric effectors will act as inhibitors by stabilising configurations in which the access to the active pocket is at least partially closed. In short, the idea is to block the protease in one of its metastable conformations, in which the catalytic dyad cannot regularly operate, inhibiting in this way the whole protein functionality. This approach, at least in principle, has several advantages. First of all, it offers the possibility to drug sites far from the catalytic pocket, thus enlarging the chance to discover active compounds and to obtain non-competitive inhibition. If an allosteric site is identified and targeted, using this strategy one can develop drugs which are highly specific since they do not bind in active sites, which are typically conserved in protein families [137]. Owing to these advantages, allostery has been established as a mechanism for drug discovery, for example to target G-protein-coupled receptors(GPCRs)[48, 91] or protein kinases[55, 143, 200].

4.2 Search for metastable states

Our strategy to identify candidate binding sites for allosteric inhibition is fully based on the analysis of a long MD trajectory. This can be seen as a first important drawback of PAk in its original formulation: it cannot be used to analyse trajectories generated under the action of an external bias, as it happens in many enhanced sampling methods. We will see in the nexh Chapter how this limityation can be overcome.

We analyse a $100 \mu s$ MD trajectory of the M$^{\text{pro}}$ generated in the D. E. Shaw Lab[52]. Our scope is to search for possible metastable states of the protease, namely configurations which do not change significantly on the scale of several tens of ns. These configurations are important for developing drugs for allosteric inhibition, since they are already (marginally) stable, and by designing a ligand which increase their stability they can become kinetic traps[137]. The local minima of the free energy, if deep enough, correspond to the metastable states, approximately the same that would be found by performing a costly Markov State Modeling analysis[154].

We look for metastable states estimating the free energy landscape with PAk, whose competitive advantage is that it allows performing the analysis in very high-dimensional spaces, taking into account at the same time several hundreds different variables without explicit dimensional reduction[157, 179]. Then the estimated FES undergoes a completely unsupervised and nonparametric density-peak clustering algorithm. This procedure allows finding the free energy minima, and thus the metastable states, with no prejudice on their structure.

4.2.1 Choice of descriptors

We extract from the $100 \mu s$ MD trajectory of the M$^{\text{pro}}$ enzyme 10.000 equally spaced frames, one every $10ns$. Since the molecular complex is a homodimer, we consider the 20.000 total frames of the two monomer trajectories as a sample of the conformational space of a single monomer. However, the trajectories of the two monomers are considered and analysed separately, in order to verify *a posteriori* whether the configurations they explore are similar or not.

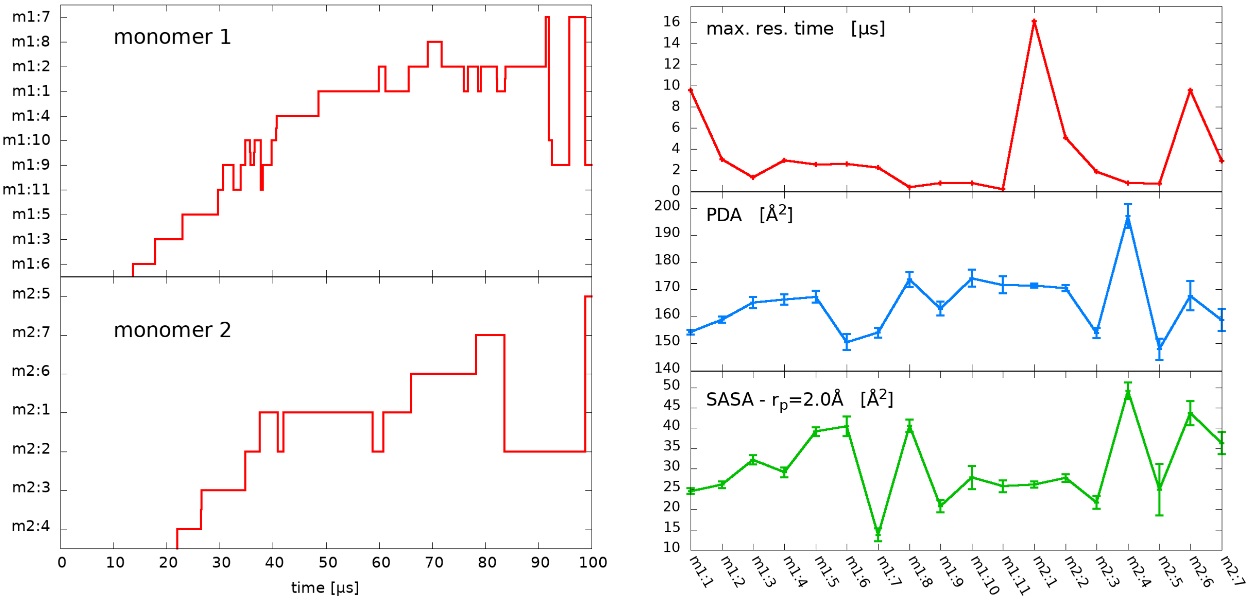
We carry out our analysis in two different descriptor spaces: the space defined by all the ψ backbone dihedrals of the protease, and the space defined by the contacts between pairs of residues which break or form during the dynamics. Both spaces consider the enzyme globally, not limiting the analysis to the catalytic dyad or to the binding pocket, which is essential to unveil possible allosteric states. Our two metrics are both sensitive to local and global conformational changes in the peptide, but capture different details: the ψ coordinates keep track of the changes in the protein backbone; the mobile contacts metrics, instead, also keep track of the side-chains rearrangements, while neglecting fluctuations around the completely formed or completely unformed contacts. In both metric spaces in which we perform our analysis, we neglect the 10 residues at the C-terminus of the peptide, since they are highly mobile in both monomers and might introduce noise in the analysis. The distance functions among points in the two metric spaces are:

- the *ψ-backbone-dihedral distance*[50]; such distance between configurations t and t' is defined as $\theta_{t,t'} = \sum_i ((\psi_{i,t} - \psi_{i,t'}))^2$, where $\psi_{i,t}$ is the value at time t of the ψ dihedral angle that involves the α-carbon of residue i of the monomer, index i runs between 1 and 296 and the notation $((\bullet))$ stands for 2π-periodicity within the brackets;

- the *contact-map distance*[50], restricted only to contacts which vary significantly during the simulation. To define these mobile contacts, we first compute the contact-map matrix C for each frame, restricted to residues 1-296. For each couple of residues ij we first evaluate the distances between all the couples of heavy atoms, with one atom belonging to i and the second one belonging to j. C_{ij} is then equal to $\sigma(d_{min})$ where d_{min} is the smallest distance between the couples of atoms, and σ is the sigmoidal function: $\sigma = (1 - (d/r_0)^{10})/((1 - (d/r_0)^{20}))$, with $r_0 = 4.5$Å. We consider as *mobile* the contacts which are completely formed ($C_{ij} > 0.8$) in at least 5% of the frames and completely broken ($C_{ij} < 0.2$) in at least 5% of the frames. Moreover, we neglect those contacts which have a value between 0.2 and 0.8 (i.e. close to r_0) in more than 50% of the frames. This procedure selects 155 relevant mobile contacts for the first monomer (m1) and 184 for the second (m2). Most of these contacts are in common, as reasonable since the two monomers are chemically identical; the union of the two sets has 235 elements. Denoting by $\mathcal{M}$ the set of mobile contacts of a monomer, the contact-map distance between configuration t and t' is $d_{t,t'} = \sum_{(i,j)\in\mathcal{M}} \sqrt{(C_{ij}(t) - C_{ij}(t'))^2}$, where $C(t)$ is the contact matrix of configuration t.

4.2.2 Our nonparametric unsupervised pipeline at work

The free energy landscape of each dataset is estimated following the procedure described in the previous chpapter. First of all, the intrinsic dimension (ID) of the manifold containing the configurations is calculated [68]. In the spaces of the ψ dihedrals we get an ID of 28 for m1 and of 26 for m2. In the spaces of the mobile contacts, we get an ID of 17 for both monomers. The free energy F of each configuration is then calculated using the PAk estimator in equation (3.12). Finally, using



(a) (b)

Figure 4.1: Global observables for the 18 identified states. **(a)** Trajectories for the two monomers in the space of the states. The frames that do not belong to a core set are relabeled by the state identifier of last visited core state; notice there is no label assigned to the first 10 to 20 μs indicating that no statistically meaningful metastable state is visited in the first part of the trajectory. **(b)** Global observables of the states. Top: the maximum residence time for each state, taken as the longest time interval over which the state label does not change. Middle: average PDA of the frames belonging to the core of a state. Bottom: average SASA of the catalytic dyad of the frames belonging to the core of a state; the SASA is computed choosing a probe radius $r_p = 2.0$Å.

Density Peak (DP) clustering[158] in its unsupervised variant[61], we build a topography of the free energy landscape.

When we find the free energy minima we assign all the frames to one of these minima according to the DP procedure. The set of configurations assigned to a single free energy minimum defines a free energy basin. Then, following ref [61], we find the saddle point between each pair of basins. The cluster Core Set (CS) of a basin is the set of configurations whose free energy is lower than the free energy of the lowest saddle point of the basin.

The described approach requires choosing the metric and a single metaparameter, the statistical confidence Z at which a basin is considered meaningful. A basin a is considered meaningful if $(F_{ab} - F_a) > Z(\varepsilon_{F_a} + \varepsilon_{F_{ab}})$ for all the basins b which share a border with a. Here, F_a is the free energy minimum of basin a, ε_{F_a} is its uncertainty, F_{ab} is the free energy of the saddle point between basin a and b, and $\varepsilon_{F_{ab}}$ is its uncertainty. In our analysis Z is set to the value Z=1.4, which corresponds to a confidence level of approximately 85 %. This means that we expect to have nearly a 15% of artificially split free energy basins. We have verified that, by varying Z around this value, the description does not change significantly: the most populated free energy basins remain approximately unchanged.

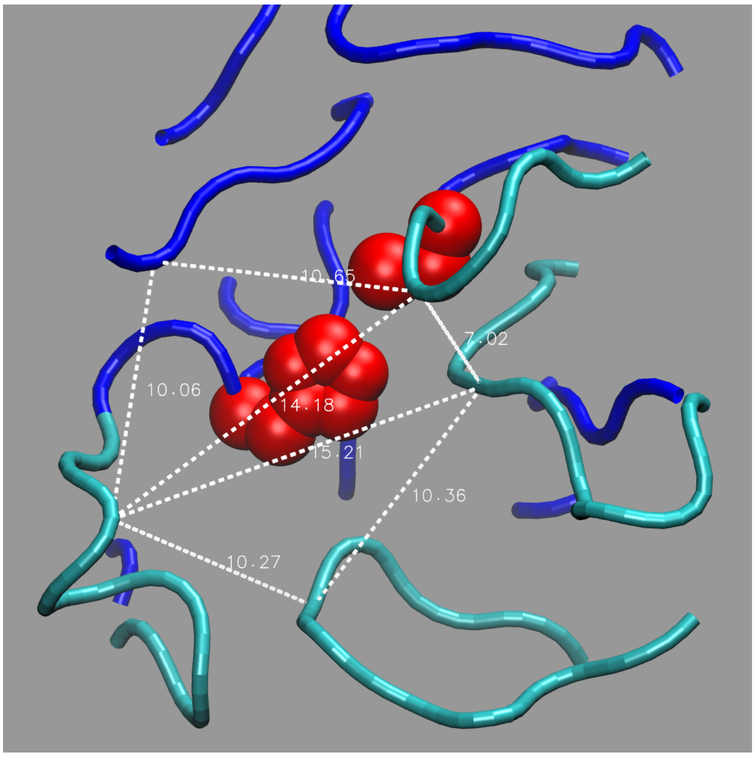


Figure 4.2: Pictorial illustration of the PDA observable. Visualisation of the backbone (in dark blue) of the residues surrounding the catalytic dyad (in red) and thus shaping the enzyme's binding pocket. In light blue the most flexible loop surrounding the cavity are represented: the left and upper flap, the linker and right loop. In white dashed lines, the segments connecting the five Cα atoms which delineate the three triangles whose total area we call PDA. Such triangles are: Ser^{46}-Gly^{143}-Met^{165} and Thr^{25}-Ser^{46}-Gly^{143}, Gly^{143}-Met^{165}-Arg^{188}. The segment labels report distances in Å.

In the following analysis we call *state* a set of configurations which belong to the core set of the same free energy basins according to both metrics. If, for example, a given basin number found using the dihedral metric is split in two different basins according to the contact metric, in our analysis we will consider two states. As a consequence, our states are structurally uniform according to both metrics. We consider in our analysis only states with a population of at least 8 core state configurations. With this criterion, we identify 11 relevant states in the trajectory of m1 and 7 in the trajectory of m2, for a total of 18 metastable states.

4.3 Characterisation of metastable states

4.3.1 Global observables

Firstly, we want to make sure that the metastable states detected analysing the m1 and m2 trajectories separately are the same as if we run the algorithm on the merged 20.000 configurations. We check it in the case of the mobile contacts metric. We find that all the clusters involve either only frames from the first monomer or from the second. There is no relevant cluster that shares structures from both monomers, meaning that in terms of the contact map the configurations of m1 are different from the configurations of m2. Due to their chemical identity, in an ergodic simulation the configurations explored by the two monomers should be nearly identical. Therefore, the first important result of our analysis is that $100\mu s$ of MD simulation are not sufficient to explore ergodically all the configuration space, as recently claimed also by Cocina et al.[42]. This is also visible by looking at Figure 4.1a: most states are visited only 2-3 times. Consequently, the mean residence time cannot be meaningfully estimated. We instead compute, the maximum residence time, considering it a proxy of the metastability of each state. These times are shown in the upper panel of Figure 4.1b and range from $0.20\mu s$ to $16.07\mu s$.

To quantify the accessibility to the catalytic site we estimate two observables: the first one is the well-konwn average Solvent-Accessible Surface Area (SASA) of the dyad[11, 115, 176]. The second one was defined by us, designed specifically for the active site of the M^{pro}; we call it Pocket Doorway Area (PDA) and it quantifies the opening of the catalytic pocket from the position of four selected $C\alpha$ carbons. PDA is defined as the sum of the area of the three triangles formed by the $C\alpha$ carbons Thr^{25}-Ser^{46}-Gly^{143}, Ser^{46}-Gly^{143}-Met^{165} and Gly^{143}-Met^{165}-Arg^{188}, which form the tips of 5 loops delimiting the cavity. For a visual representation of the PDA see Figure 4.2. The two quantities, presented in the middle and lower panels of Figure 4.1b, are in general quite correlated, although not in all the states. Indeed, contrary to PDA, SASA is sensitive to what happens in the direct proximity of the catalytic residues, while neglecting more macroscopic rearrangements of the catalytic pocket.

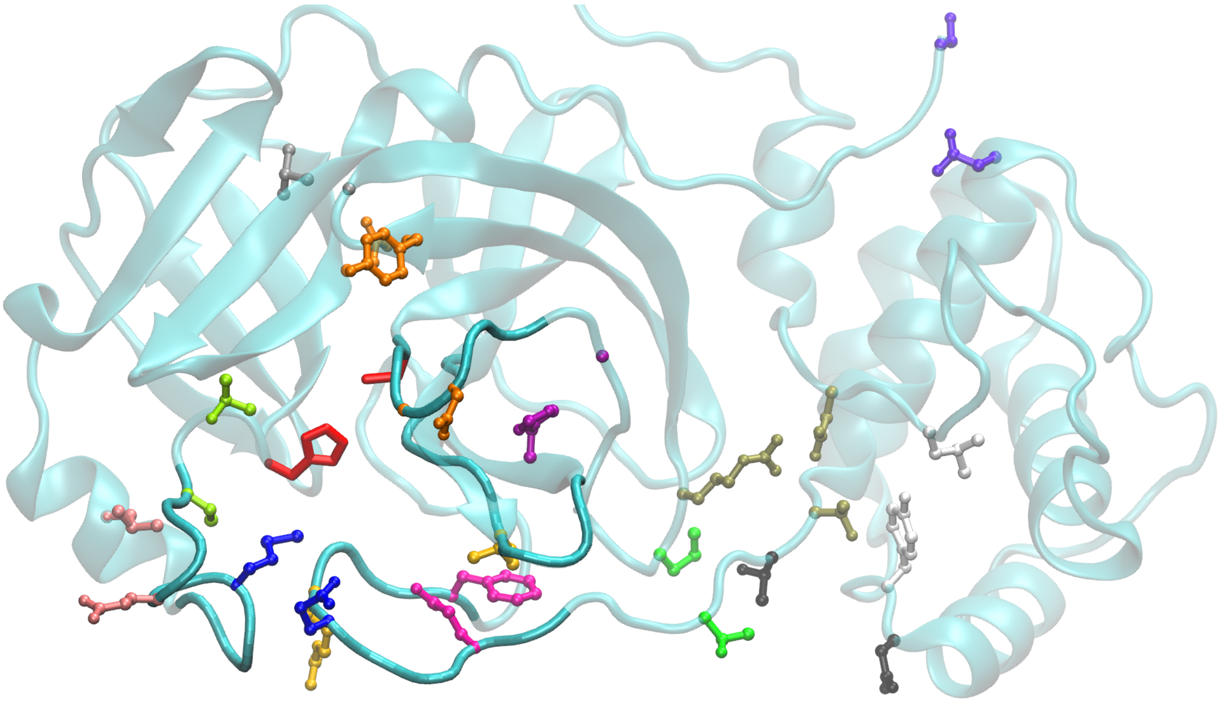
4.3.2 Structural characterisation of the states

We characterise the states by analyzing in detail their contact structure and their backbone arrangement. In the case of the mobile contacts, we analyse the intra-monomer contacts which change significantly between at least two of the 18 states; furthermore, we also track the behaviour of few inter-monomer contacts that might reflect some changes in the metastable states' catalytic cativity[10, 12]. When the average over the configurations belonging to a state of the contact matrix entry C_{ij} for a given contact (i, j) is < 0.3, we consider that the contact is not formed in that state; when the average is > 0.7 we deem it to be formed; if neither of the case applies we label the contact as undetermined for that state and indicate it with letter n. The contact structure of the selected states is summarised by the table in Figure 4.3a. The contacts displayed are all the mobile contacts that change relevantly among states; contacts not displayed in Table 4.3a either have an undetermined label n for most of the stares or they involve residues directly contiguous to some contact displayed.

Turning to the backbone, we analyse the ψ dihedral angles in the loops closing the cavity and

state ID	Thr^{25} - Cys^{44}	Asn^{28} - Tyr^{118}	Asn^{28} - Gly^{143} & Ser^{144}	Glu^{47} - Leu^{57}	Met^{49} - Gln^{189}	Tyr^{118} - Asn^{142}	Leu^{167} - Arg^{188}	Phe^{185} & Val^{186} - Gln^{192}	Leu^{141} - Gly^{2}*	Gly^{2} - Ser^{214}	Val^{18} - Gly^{120}	Arg^{131} - Thr^{199}	Arg^{131} - Asp^{289}	Pro^{132} - Thr^{196}	Gly^{138} - His^{172}	Asp^{197} & Thr^{198} - Asn^{238}	Tyr^{239} - Leu^{287}	Ala^{285} - Ala^{285}*
m1:1	1	1	0	0	1	0	0	1	0	n	1	1	1	0	1	0	1	1
m1:2	n	1	0	0	1	0	0	0	0	n	1	0	1	1	1	1	1	1
m1:3	0	1	1	0	1	1	0	1	1	1	1	1	1	0	0	0	1	1
m1:4	0	1	0	1	0	0	0	1	0	1	1	1	1	0	1	0	1	n
m1:5	0	1	1	1	0	1	1	1	1	1	1	n	1	1	0	1	1	1
m1:6	n	1	1	0	0	1	0	1	1	n	1	0	n	0	1	1	1	1
m1:7	0	1	0	1	0	0	0	1	0	1	1	0	1	1	1	1	1	1
m1:8	0	1	0	0	0	0	0	1	0	0	1	1	1	0	1	0	1	1
m1:9	0	1	0	1	0	0	0	1	0	1	1	0	0	0	1	1	1	n
m1:10	0	1	0	1	0	0	0	1	0	n	1	0	n	0	1	1	1	1
m1:11	0	1	0	1	0	0	0	1	0	1	1	1	1	0	1	0	1	1
m2:1	1	1	1	0	1	1	0	1	n	1	1	0	n	1	0	1	0	1
m2:2	1	1	1	0	1	1	0	1	n	1	1	0	n	1	0	1	0	1
m2:3	0	0	0	0	0	1	0	1	n	1	0	0	n	1	0	1	1	1
m2:4	n	0	0	0	0	n	0	1	1	1	n	0	1	0	0	1	1	1
m2:5	0	0	0	0	n	1	1	0	n	1	n	n	1	0	1	1	0	n
m2:6	1	1	1	0	n	1	0	1	n	1	1	0	0	1	0	1	0	1
m2:7	n	1	1	0	1	1	0	1	n	1	1	0	0	n	0	1	0	1

(a)



(b)

Figure 4.3: **(a) Table presenting the status of selected intra-monomer contacts and inter-monomer contacts**. In the case of inter-monomer contacts, the residue of the monomer which is excluded by the metric that defines a state is marked with a star(*). For each contact (columns) the average over the configurations of a given state is reported in the corresponding row. Such contacts are divided into two subgroups by a double vertical line: on the left those between residues belonging to the flexible loops which control the access to the binding pocket and on the right other contacts. For readability, the entries take only three possible labels: 0 when the contact is not formed, 1 when it is formed and *n* in all other case. Contacts whose label does not vary in any of the states of a given monomer are reported in light gray colour. **(b) Visualisation of selected inter-monomer contacts**. A VMD[99] representation of monomeric M$^{\text{pro}}$ in state m1:1; on the left hand side the enzyme binding pocket, which encloses the catalytic dyad (in red); all other highlighted residue couples refer to the contact with the corresponding colour in the table.

state ID	Ile^{43}-Pro^{52} loop	Phe^{140}-Cys^{145} loop	Phe^{185}-Tyr^{201} loop	2	61	153	154	168
m1:1	$\alpha\beta\beta\alpha\alpha\alpha\alpha\alpha c\beta$	$\beta\beta\beta\alpha\beta\beta$	$\beta\beta c\beta\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	c	α	α
m1:2	$\alpha\beta\beta\alpha\alpha\alpha\alpha\alpha c\beta$	$\beta\beta c\alpha\beta\beta$	$\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta c\beta\beta\beta\alpha$	β	β	c	α	α
m1:3	$\alpha\alpha\alpha\alpha\alpha c\alpha\alpha c\beta$	$c\alpha\beta\alpha\alpha\beta$	$\beta\beta c\beta c\beta\beta\beta\beta\beta\beta\beta c\beta\beta\beta\alpha$	β	β	β	c	α
m1:4	$\alpha\alpha\alpha\alpha\beta c\beta\alpha\beta$c	c$\beta c\alpha\beta\beta$	$\beta\beta c\beta c\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	β	c	α
m1:5	$\alpha\alpha\alpha\alpha\beta c\beta\alpha c\beta$	$c\alpha\beta\alpha\alpha\beta$	$\beta\beta c\beta c\beta c\beta\beta\beta\beta\beta c\beta\beta\beta\alpha$	β	β	β	α	α
m1:6	$\alpha\beta\beta\alpha\alpha\alpha\beta\alpha c\beta$	$c\alpha\beta\alpha\alpha\beta$	$\beta\beta c\beta\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta c\alpha$	c	β	β	c	α
m1:7	$\alpha\alpha\alpha\alpha\beta c\beta\beta c\beta$	$\beta\beta\beta\alpha\beta\beta$	$\beta\beta c\beta c\beta c\beta\beta\beta\beta\beta c\beta\beta c\alpha$	β	β	β	α	α
m1:8	$\alpha\beta\alpha c\beta c\alpha\beta\beta\beta$	$\beta\beta\beta\alpha\beta\beta$	$\beta\beta c\beta\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	c	α	α
m1:9	$\alpha\alpha\alpha\alpha\beta c\beta\beta c\beta$	$\beta\beta c\alpha\beta\beta$	$\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	β	α	α
m1:10	$\alpha\alpha\beta\alpha\beta\alpha\beta\beta c\beta$	$\beta\beta c\alpha\beta\beta$	$\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	β	α	α
m1:11	$\alpha\alpha\alpha\alpha\beta c\beta\beta c\beta$	$c\beta c\alpha\beta\beta$	$\beta\beta c\beta\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	β	α	α
m2:1	$\alpha\beta\beta\alpha\alpha\alpha\alpha\alpha c\beta$	$\beta\beta\beta\alpha\alpha\beta$	$\beta\beta c\beta c\beta\beta\beta\beta\beta\beta\beta c\beta\beta\beta\alpha$	β	β	β	c	α
m2:2	$\alpha\beta\beta\alpha\alpha\alpha\alpha\alpha c\beta$	$\beta\beta\beta\alpha\alpha\beta$	$\beta\beta c\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	c	α	α
m2:3	$\alpha\beta\beta\beta\beta\alpha\alpha\beta\beta\beta$	$\beta\beta\beta c\alpha\beta$	$\beta\beta c\beta c\beta\beta\beta\beta\beta\beta\beta c\beta\beta c\alpha$	α	α	c	α	α
m2:4	$\beta\beta\beta\beta\beta\alpha\beta\alpha\beta\alpha$	$c\beta\beta\alpha\alpha\beta$	$\beta\beta c\beta\beta\beta c\beta\beta\beta\beta\beta\beta\beta\beta\beta\alpha$	α	β	c	α	α
m2:5	$\alpha\beta\beta\beta\beta\alpha\beta\beta\beta\alpha$	$\beta\beta\beta\beta\beta\beta$	$c\alpha\beta\alpha\alpha\alpha\alpha\alpha\alpha\beta\beta\beta\beta\beta\beta\beta\alpha$	β	β	c	α	β
m2:6	$\alpha\beta\beta\alpha\beta\alpha c\alpha c\beta$	$\beta\beta\beta c\alpha\beta$	$\beta\beta c\beta c\beta\beta\beta\beta\beta\beta\beta c\beta\beta\beta\alpha$	c	β	c	α	α
m2:7	$\alpha\beta\beta\beta\beta\beta\alpha\beta\beta\beta$	$\beta\beta\beta\alpha\alpha\beta$	$\beta\beta c\beta\beta\beta c\beta\beta\beta\beta\beta c\beta\beta\beta\alpha$	β	β	c	α	α

Table 4.1: Selected ψ backbone dihedral angles . The first three column refer to the three most flexible loops, which are the ones controlling the access to the catalytic pocket. The remaining columns refer to other isolated dihedrals, selected due to their high variability throughout the 18 states. For each row, the average over the configurations of the corresponding state is considered. For a better readability, we adopt a ternary labelling: if $-\pi/2 < \psi < pi/6$ the angle is labeled as α; if $\psi < -11/12\pi$ or $\psi > \pi/2$ as β; in all other cases the angle is labeled as c. Dihedrals whose label does not vary in any of the states of a given monomer are reported in light gray colour.

other few dihedrals which change significantly in the various states. Results are reported in Table 4.1. Notice that the labels α,β and c do not exactly refer to the peptide's secondary structure (α-helix, β-sheet and coil), since ψ dihedrals are not enough to univocally map the secondary structure geometry (and not even the Ramachandran plot[156] might be enough [164]); however, they provide a rapid, though approximate, indication, since the value of ψ is highly correlated with the secondary structure geometry.

4.4 Description of the metastable states

4.4.1 A little nomenclature

In this subsection we introduce some terms that will help us to describe the structural properties of the metastable states. As mentioned above, the catalytic dyad His^{41}-Cys^{145} is located in the pocket between the protein domains I and II. The access to this cavity is controlled by the flexible loop structures highlighted in Figure 4.3b. The two most flexible loops[25] involve residues from Ile^{43} to Pro^{52} (*left flap*) and from Phe^{185} to Tyr^{201} (*linker loop*). The left flap corresponds to the leftmost loop in Figure 4.3b, and opens and closes like a small door. No conformers from the second dimer m2 have the left flap wide open, consequently contact Glu^{47}-Leu^{57} is never formed. The linker loop closes the cavity from below in Figure 4.3b and links domains II and III. All the m2 states have a loosely structured linker loop, with contact Arg^{131}-Thr^{199} almost never formed and contact Asp^{197}&Thr^{198}-Asn^{238} always formed. The contacts controlling the distance between the β barrels of the I and II protein domain[207] (Asn^{28}-Tyr^{118} and Val^{18}-Gly^{120}), which are always formed in m1, are at times unformed in m2. The loop from Phe^{140} to Cys^{145} (we call it *upper flap*) is smaller and assumes mainly two conformations: tilted downwards (contacts Ans^{28}-Gly^{143}&Ser^{144} and Tyr^{118}-Asn^{142} not formed, dihedral ψ_{144} in β configuration), which hides the catalytic Cys^{145} or flat out (ψ_{144} in α configuration), which leaves more access to the dyad. Last, the β-sheet loop from Met^{162} to Gly^{170} delimits the cavity from the right in Figure 4.3b (we call it *right loop*); it is the least flexible, but it interacts with the N-finger of the other monomer and is crucial for shaping the substrate binding pocket[85].

4.4.2 Structural analysis

All m2 states except m2:5 have the upper flap not tilted down and retracted with respect to the pocket, with contact Tyr^{118}-Asn^{142} almost always formed and contact Gly^{138}-His^{172} almost never formed. These two contacts are almost always mutually exclusive, with exception of states m1:6 and m2:5, in which both contacts are formed at the same time. Another important difference among states, not related with the loops, is that dihedrals from Leu^{227} to Asn^{238} (bottom right in Figure 4.3b) in all states of m1 are arranged in α configuration, so that an α-helix is formed and contact Tyr^{239}-Leu^{287} is always formed; in m2 such α-helix structure is often defective. As for the contact between the N-finger and domain III (contact Gly^{2}-Asn^{214}), in m2 it is often formed, while it is broken in most m1 states.

We describe all the states in detail in Appendix C. Hereby, we focus on the most stable, the most open and the most closed according to the SASA and PDA observables. From the analysis of the maximum residence time it is clear that states 1 and 2 of both m1 and m2 are among the longest-lived metastable states. All four are in fact very similar to the crystallographic structure (PDB 6Y84[138]): they all have the left flap and the linker loop in contact between each other (cont. Met^{49}-Gln^{189}); the left flap is closed (cont. Glu^{47}-Leu^{57} broken, cont. Thr^{25}-Cys^{44} formed) and the linker loop stretched towards it (cont. Leu^{167}-Arg^{188} broken), covering the lower part of the binding pocket.

The two most open states are m2:4, which ranks the highest in both PDA and SASA, and m1:8. In m2:4 the upper flap is not tilted downwards and is far from the pocket and from the right loop, leaving cont. Gly^{138}-His^{172} not formed; the left flap is very open (although the dihedrals of this loop are quite variable among the configurations of such state); the linker loop is slightly contracted (cont. Arg^{131}-Thr^{199} and Pro^{132}-Thr^{196} not formed), not stretching towards the left flap as in other closed or partly-closed states; this leaves the catalytic dyad well exposed. State m1:8 also ranks very

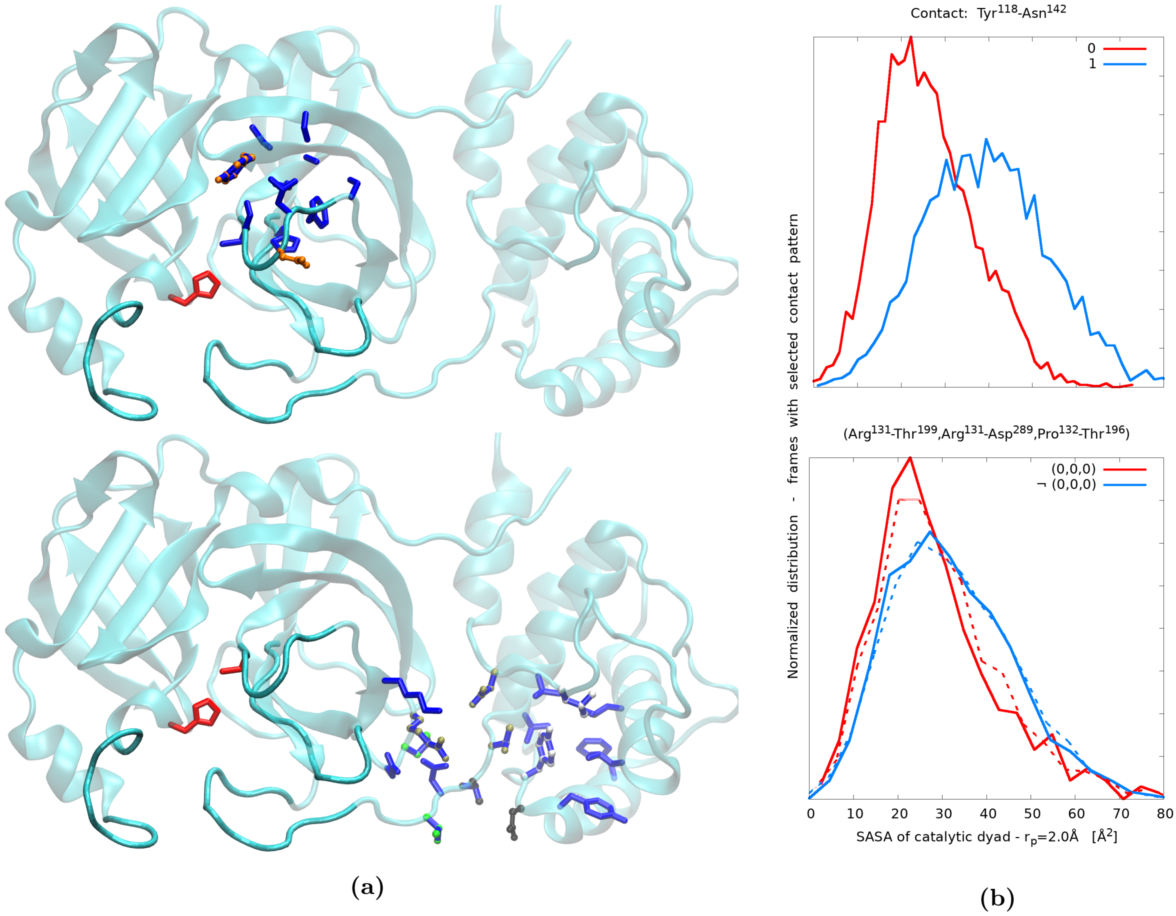


Figure 4.4: (a) Visualisation of monomeric M^{pro} in state m1:9[99]; in red the catalytic dyad; in dark blue the residues involved in the upper pocket (top) and the distal pocket (bottom) found by the software PockDrug[100]. **(b) SASA distibutions over configurations with selected contact patterns**. 0 indicates a contact surely not formed, 1 indicates a contact surely formed. Top: upper pocket. Bottom: distal pocket.

high in PDA and in SASA. The left flap is open, although dihedrals from Ile^{43} to Ser^{46} are not all in α configuration; their particular arrangement ($\alpha\beta\alpha c$), however, grants that the biggest sidechains of the left flap are not oriented towards the binding pocket. The linker loop is not stretched towards the left flap, but rather down, towards the interface with the solvent; it is quite open (dihedral of Gln^{189} in c instead of β configuration) in proximity of the pocket and all its sidechains do not obstruct the access to the cavity (in particular those of Arg^{188} and Gln^{189}, responsible for a low SASA in other states).

Among the most closed states we mention m1:7,m1:9,m2:3,m2:5. State m1:9 is very similar to m1:10 in its contact and backbone structure, with the exception of the left flap, which is more open in state m1:10. State m1:9 is also structurally similar to m1:7: the only difference among the contacts is Pro^{132}-Thr^{196}, which is formed in m1:7 and not in m1:9, allowing the lower loop to be more flexible. In both, the upper flap is tilted downwards, but the left flap backbone is open. In m1:9 the side-chains of the residues in the loops surrounding the binding pocket are oriented towards

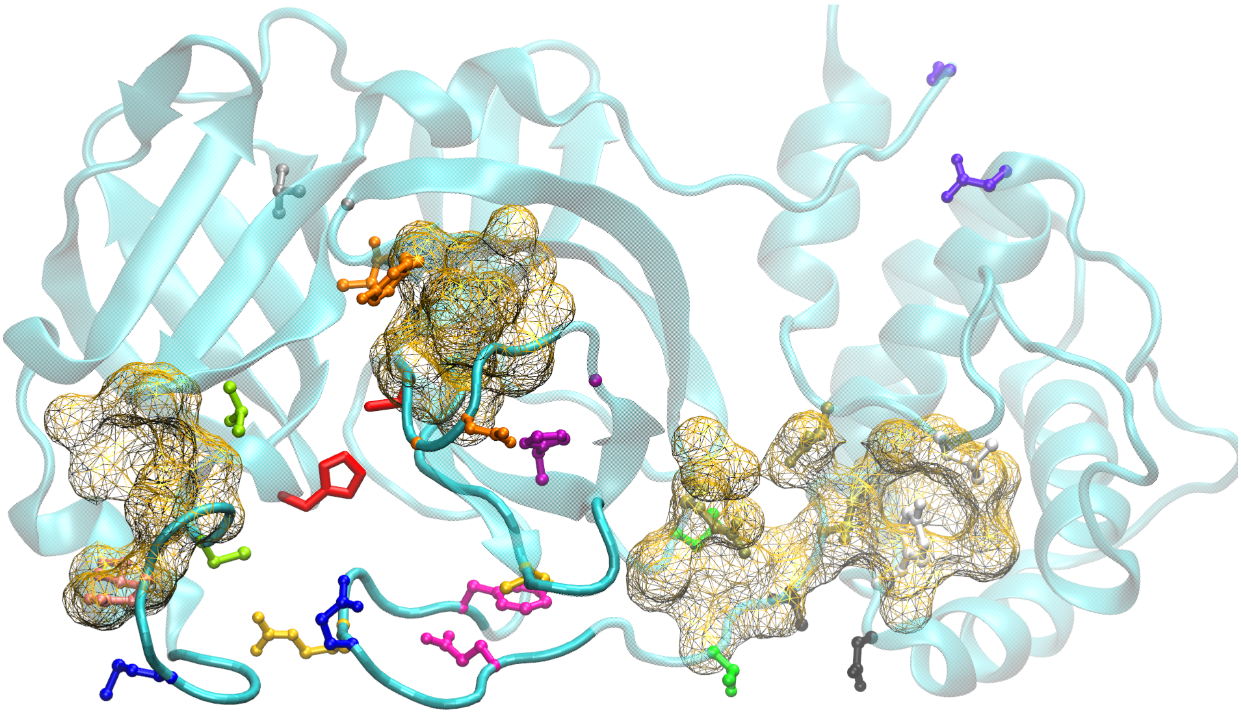


Figure 4.5: Representation of monomeric M^{pro} in state m1:1 with highlighted residues and candidate binding pockets. Residue couples involved in contacts are highlighted according to the colour code in the the table in Figure 4.3a. The three proposed binding sites are represented by a golden wire mesh.

the catalytic dyad, causing such state to rank among the lowest in SASA. State m1:7 ranks among the lowest in PDA and as the lowest in SASA; the reason lies in the sidechains of the lower and left flaps, in particular of Thr^{45} and Gln^{189}, which form a contact and effectively close the access to the reactive site. State m2:3 ranks as the third lowest in both SASA and PDA. Cys^{145} is not well covered, but on the other hand His^{41} is less accessible than in most other states. As most m2 states, m2:3 has the upper flap bent upwards and contact Gly^{138}-His^{172} not formed. The linker loop is not stretched, leaving the contacts with Arg^{131} partly unformed. The left flap is closed and stretched towards the linker loop and its dihedrals are arranged in such a way that cont. Met^{49}-Gln^{189} is not formed. Finally, state m2:5 is the one with the lowest PDA and is among the lowest-ranked in SASA. Its conformation is quite peculiar: the linker loop is all retracted and coiled (it is the only state of m2 forming cont. Leu^{167}-Arg^{188}). The left flap is all stretched towards the linker loop (cont. Met^{49}-Gln^{189} formed) and almost completely covers the catalytic His^{41}. The upper flap, rather than being flat or tilted down, is oriented upwards, causing a deformation in domain II which allows cont. Gly^{138}-His^{172} to be formed. Remarkably, like m1:9, state m2:5 is one of the few states with cont. Ala^{285} - Ala^{285*} not tightly formed.

4.5 Looking for druggable targets

Our analysis shows that the accessibility to the catalytic dyad is reflected in the forming and breaking of few relevant contacts around the reactive cavity. For example, cont. Glu^{47}-Leu^{57} is not formed when the left flap is closed, a condition common to most states in which the catalytic dyad is not accessible. Similarly, the catalytic site (in particular Cys^{145}) is less exposed when the upper flap it tilted downwards, i.e. when cont. Tyr^{118}-Asn^{142} is not formed. The druggability analysis software PockDrug[100] finds one pocket in correspondence of the residues of each of the two contacts (respectively called left pocket and upper pocket) and assigns to them a druggability probability of 0.68 ± 0.08 and 0.95 ± 0.03. Targeting these two regions with drug-like compounds, blocking the formation of the mentioned contacts, might prove a successful strategy for the inhibition of the catalytic activity. The distribution of SASA over all configuration in which contact Tyr^{118}-Asn^{142} is not formed is significantly shifted towards lower SASA values than in the cases in which the contact is formed (see Figure 4.4b).

Our analysis on the relevant contacts also unveils the presence of another interesting pocket far from the catalytic site, in the interface region between domains II and III (right hand side of the table in Figure 4.3a). The five relevant contacts in this region are: Arg^{131}-Thr^{199}, Arg^{131}-Asp^{289}, Pro^{132}-Thr^{196}, Asp^{197}&Thr^{198}-Asn^{238}, Tyr^{239}-Leu^{287}. This region, which we call *distal pocket* has been previously identified and screened for docking and has been predicted as a potential druggable target[59, 183]. It has also been suggested as a target for allosteric inhibition of the catalytic activity[60, 209]. Coherently, the predicted druggability score is 0.65 ± 0.08. Experimental confirmation of the viability of the distal pocket as a target comes from crystallographic fragment screening[45, 59]. Among the hits that were identified, three are particularly interesting. Fragment Mpro-x0390, classified as “high confidence", is in contact with atoms from five different residues, among which four are involved in the relevant contacts mentioned above. Fragment Mpro-x0464, also classified as “high confidence", is in contact with eleven residues, among which six are involved in the relevant contacts. Fragment Mpro-x1163, classified as “correct ligand but with weak density", is in contact with nine residues, among which five are involved in the relevant contacts. With a completely different approach, the database Pocketome[9] identifies for the coronavirus M$^{\text{pro}}$ a bindable pocket in the distal region, with two possible ligands (entry R1AB_SARS2_P6); this pocket includes residues Pro^{132}, Thr^{196}, Thr^{198}, Asn^{238}, Tyr^{239}, all involved in the five relevant distal pocket contacts. Alternatively, many other algorithms have been developed for the detection and scoring of druggable pockets[67, 82, 109, 117, 142, 165, 196]. We decided to further benchmark our findings by running the pocket detection software fpocket[113]. While for most structures the analysis does not detect any pocket in the distal region, the structures in the core set of state

Source Organism	47	57	118	142	131	132	196	197	198	199	238	239	287	289
Human SARS-CoV2	E	L	Y	N	R	P	T	D	T	T	N	Y	L	D
Human SARS-CoV	E	L	Y	N	R	P	T	D	T	T	N	Y	L	D
Murine CoV	A	L	Y	C	R	S	Q	D	Y	T	G	F	L	D
Human CoV 229E	T	E	Y	N	R	T	A	N	Q	M	G	F	L	D
Feline CoV	T	E	Y	A	R	S	T	N	V	M	S	F	L	D
Avian inf. bronchitis	S	V	Y	A	R	S	P	D	N	L	G	F	F	D
Thrush CoV HKU12	K	I	Y	N	Q	T	T	F	Q	Y	S	F	F	C

Table 4.2: Mutation and conservation of relevant residues. Amino acid 1-letter code of relevant residues in the Human SARS-CoV2 3CL$^{\text{pro}}$ (from PDB 6Y84) and of the corresponding residues in the other proteins in the seed of the same Pfam family (Coronavirus endopeptidase C30, Pfam entry PF05409). The sequence IDs reported as column headers refer to the sequence of Human SARS-CoV2 3CL$^{\text{pro}}$, in the first non-header line of the table.

m1:9 display two pockets in contact with various residues in the distal region, even if with low druggability. Finally, we analyse the whole trajectory with the software MDpocket[168], which quantifies in terms of a frequency grid the points involved in accessible pockets: the frequency value ranges from 0 if a point is never found along the trajectory in an open pocket to 1 if it is always found. The software assigns low values to the distal pockets: this suggests that the distal pocket is observed as a transient site, which makes its detection non-trivial. With the aim of verifying the presence of allosteric effects involving the distal pocket, we focus on the above mentioned contacts in this region. We compute the distribution of the PDA and of the SASA restricted to the frames in which the contact pattern described above is present or not. Despite all considered residues being far from the binding pocket, the distributions of the PDA and of the SASA are sizably different in the two conditions (see Figure 4.4b). This suggests that if these five contacts could be forced to be formed or broken according to the desired pattern, e.g. by a drug-like compound, one could influence the PDA and the SASA, controlling indirectly the access to the reactive site. Comparing the table in Figure 4.3a and Figure 4.1b, a good candidate for allosteric drugging seems to be the contact pattern of state m1:9: $(0, 0, 0, 1, 1)$. Interestingly, the PDA and SASA distributions obtained by selecting only the first three of the five contacts, namely $(0, 0, 0)$, do not differ significantly from those with all five contacts involved (see e.g. Figure 4.4b).

4.6 Mutation and conservation of relevant residues within the same protein family

We finally analyse the conservation of the residues involved in all the proposed contact patterns in the sequences of proteins belonging to the same family as M$^{\text{pro}}$. We perform a multiple sequence alignment of our sequence (from PDB 6Y84[138]) with all the sequences in the Pfam[64] seed of the corresponding family, Coronavirus endopeptidase C30 (Pfam entry PF05409), obtained via multiple sequence alignment. Similarly to reference [192], we find that many of the residues involved in the proposed target sites are conserved in all or most of the sequences and furthermore all of them are conserved in the sequence of Human SARS coronavirus (SARS-CoV). In fact, as we see by looking at Table 4.2, all relevant contacts are conserved between the M$^{\mathrm{pro}}$ of Human SARS-CoV2 and Human SARS-CoV. Particularly stable within the protein sequences appear to be the residues corresponding to: Tyr^{118}, Arg^{131}, Asp^{289}, Leu^{287}. Furthermore, quite recurrent are Asn^{142}, Thr^{196}, Asp^{197}.

4.7 Discussion

Our data analysis approach allowed us to identify 18 putative metastable states of the M$^{\mathrm{pro}}$ of SARS-CoV-2. We characterised these states in terms of their structural differences, identifying some contacts which are selectively formed or broken in the different states. We believe that this analysis brings insight on the molecule's conformational changes which might prove useful for the design of farmaceutical inhibitors. Our analysis approach is useful especially for understanding (and eventually controlling) the global dynamics of a protein, since treats the region of the catalytic cavity and any other part of the protein within the same framework. We stress that the same kind of analysis can easily be applied to any other candidate target proteins, due to its extreme generality.

Based on this analysis we propose some possible target sites for the design of drug-like molecules, some of which directly in contact with the flaps regulating the access to the enzyme's active site, some located in the distal pocket at the interface between domains II and III of the monomers (see Figure 4.5). We provide evidence of allosteric effects connected to such pocket and we propose as drug target simply three contacts whose inhibition is correlated to a reduction in the access to the catalytic site; a more refined drug design could yield even stronger catalytic inhibition. We show that all three proposed target sites are comprised in pockets with high druggability score according to the software PockDrug. We find that all residues involved in the proposed target sites are conserved between the M$^{\mathrm{pro}}$ of Human SARS-CoV and Human SARS-CoV-2 and that many of them are conserved in most sequences in the seed of the Pfam family to which they both belong. We interpret this as a comforting indication for the validity of our proposed targets. Moreover, the conservation of all such residues might suggest that mutations are unlikely, thus hopefully the displayed allosteric mechanisms are resistant to possible future mutations. A further possible interesting way to validate the viability of the predicted pockets as potential drug targets, especially of the distal pocket, would be analysing the effect of mutations in that region on the catalytic activity. Finally, a dynamical docking simulation would be the next step to assess our findings from a more accurate biochemical standpoint.

To summarise, the added value provided by our analysis is twofold. First, and most importantly, we provide the structure of the state which should be targeted for drug design. This structure does not coincide with the crystallographic structure, and not even with the most likely configuration observed in the MD simulation: indeed some crucial tertiary contacts which are formed in the crystal are not formed in the structure we propose, and these contacts form and break dynamically along the trajectory. Available bioinformatic tools for searching druggable cavities do not normally provide hints on the structural rearrangement which should be induced by the drug to modify the properties of the catalytic cavity, as we are instead able to do. The second non-trivial insight provided by our analysis is that it unveils high mobility in the distal pocket region, excluding the presence of relevant conformational changes coupled with the accessibility of the catalytic dyad in other sites. Even if we cannot exclude that allosteric effects may arise even from other pockets, our findings suggest prioritising these targets among the wealth of putative binding sites found by automatic scanning. The structures of the putative metastable states described in this work are available in the Supporting Information of reference[30] for independent structural analysis and for targeted drug design which, we hope, will be performed by groups with the appropriate competences.

All this analysis, however, was made possible by the public availability of a very long MD trajectory of the sytem considered, generated on a powerful supercomputer[172, 173]. The availability of such large special-purpose machines is a prerogative of few groups in the world. In order to obtain a relevant MD simulation an arbitrary system of interest, most researchers have to resort to enhanced sampling methods. In the next chapter we introduce a framework that allows extending our free energy estimation approach to the case of biased MD simulations.

Chapter 5

bPAk: free energy estimates from statically biased simulations

In the previous chapter we have presented an analysis of a molecular dynamics trajectory including configurations sampled from the canonical distribution. However, when dealing with molecular systems or other systems characterised by rare transitions, e.g. the one presented in Chapter 4, samples representative of the underlying distribution, which is typically unknown, are difficult to obtain and not always trajectories generated by special-purpose supercomputers[172, 173] are available. To address this problem people resort to enhanced sampling techniques[15, 18, 112, 174, 184, 206].

The simplest manner to artificially force a simulation to visit the relevant states of the system in a short simulation time is adding an external bias potential. Since estimating the free energy surface in the sense of equation 2.5 from a finite sample requires, implicitly or explicitly, counting how many times the system is observed in a finite region, the effect of this bias must be taken into account to estimate the unbiased distribution.

In this chapter we introduce an approach to estimate the free energy as a simultaneous function of several descriptors starting from data generated in a biased simulation. The method exploits the properties of PAk, especially those discussed in section 3.2.4 which make PAk estimator *punctual*, in a sense that will be rigorously defined in this chapter. We show that punctuality allows removing the effect of the external bias in a simple and rigorous manner. The approach is validated on model systems for which the free energy is known analytically and on a small peptide for which the ground truth free energy is estimated in an independent unbiased run. In both cases the free energy obtained with our approach is an unbiased estimator of the ground-truth free energy, with an error whose magnitude is also predicted by the model. The results hereby presented are published in reference [29] with the title *Statistically unbiased free energy estimates from biased simulations* and parts of this chapter are freely taken from such publication.

5.1 The problem of reweighting in biased simulations

The prototype of many enhanced sampling methods is Umbrella Sampling[189], in which an external bias potential, constant in time, is added to the potential energy with the aim of accelerating the transitions between the local free energy minima and explore all the relevant metastable states of the system. We shall focus here on this specific approach, aware that the applicability of our method can be extended given that few conditions are satisfied.

Let us consider a biased simulation in the canonical ensemble: during the simulation, the bias $B(\mathbf{x})$ is added to the potential energy $U(\mathbf{x})$. This bias, is a function of a (possibly multidimensional) CV $\mathbf{s}(\mathbf{x})$, namely $B(\mathbf{x}) \equiv B(\mathbf{s}(\mathbf{x}))$. In the most general case, one might want to estimate the free energy as a function of another (also possibly multidimensional) CV $\boldsymbol{\sigma}(x)$. The unbiased free energy for a specific value $\tilde{\boldsymbol{\sigma}}$ can be estimated as:

$$F(\tilde{\boldsymbol{\sigma}}) = -\beta^{-1} \log \int \rho^B(\mathbf{x})\, e^{\beta B(\mathbf{s}(\mathbf{x}))}\, \delta(\tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma}(\mathbf{x}))\, \mathrm{d}\mathbf{x} \; + f^B \tag{5.1}$$

where $\rho^B(\mathbf{x}) = Z_B^{-1} e^{-\beta(V(\mathbf{x})+B(\mathbf{s}(\mathbf{x})))}$ is the biased probability distribution, Z_B is the biased canonical partition function and f^B is an additive constant that will from now on be neglected. In ordinary Umbrella Sampling the biasing CV is an explicit function of the the variables σ, namely $\mathbf{s}(\mathbf{x}) \equiv \mathbf{s}(\boldsymbol{\sigma}(\mathbf{x}))$. In this case equation 5.1 takes the simple known form $F(\tilde{\boldsymbol{\sigma}}) = F^B(\tilde{\boldsymbol{\sigma}}) - B(\mathbf{s}(\tilde{\boldsymbol{\sigma}}))$ where $F^B(\tilde{\boldsymbol{\sigma}}) = -\beta^{-1} \log \int \rho^B(\mathbf{x})\, \delta(\tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma}(\mathbf{x}))\, \mathrm{d}\mathbf{x}$. If $\mathbf{s}(\mathbf{x})$ is a generic function of the coordinates, instead, the exponential factor cannot be brought out of the integral and there is no easy manner to estimate the unbiased probability density $\rho(\boldsymbol{\sigma})$ from $\rho^B(\boldsymbol{\sigma})$.

Both in the case in which $\mathbf{s}$ is a function of $\boldsymbol{\sigma}$ or not, what is generally done in literature[15, 174, 184, 206] is to relax the delta function in equation 5.1 and instead use a kernel K_h that converges to it only when the limit to zero is taken on its scale parameter h, as described in section 2.2.2. In other words, instead of F in equation 5.1, the following quantity F_K is computed:

$$F_K(\tilde{\boldsymbol{\sigma}}) := -\beta^{-1} \log \int \rho^B(\mathbf{x}) e^{\beta B(\mathbf{s}(\mathbf{x}))} K_h(\tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma}(\mathbf{x})) \mathrm{d}\mathbf{x} \tag{5.2}$$

so that, by replacing the true biased density $\rho^B(\mathbf{x})$ by its sample estimator (see equation (2.15)) one obtains:

$$\hat{F}_K(\tilde{\boldsymbol{\sigma}}) \sim -\beta^{-1} \log \sum_{j=1}^{N} e^{\beta B(\mathbf{s}(\mathbf{x}_i))}\, K_h(\tilde{\boldsymbol{\sigma}} - \boldsymbol{\sigma}(\mathbf{x}_j)) \tag{5.3}$$

where $\hat{F}_K$ denotes the estimator of F_K in equation 5.2 and is defined up to an additive constant. As seen in section 2.2 there are many examples of such kernels K among nonparametric methods. The common feature of these estimators is that they are not punctual, namely they provide an estimate of the free energy on a finite-size region. Thus, in the limit $n \to \infty$ they converge to F_K in equation 5.2 but not to F in equation 5.1. Over this region the value of $e^{\beta B(\mathbf{s})}$ can be largely fluctuating, making the estimators ill-behaved[108, 175]. The estimators $\hat{F}_K$ only converge to F asymptotically, namely in the limit when both $h \to 0$ and $n \to \infty$. However, even when big, n will always be finite, so a finite parameter h may be required in order for the estimators $\hat{F}_K$ to be statistically meaningful. This problem becomes more and more severe in high dimension, due to the COD and can happen even in the trivial case $\boldsymbol{\sigma} = \mathbf{s}$, if $\mathbf{s}$ is multidimensional or the sample is too small.

As we discussed in 3.2.2 in the PAk estimator for each point i one estimates the free energy by maximising the likelihood of a model in which F is assumed to depend linearly on the the neighbourhood order l:

$$\hat{F}_i := \underset{F}{\operatorname{argmax}} \ \max_{a} \sum_{l=1}^{\hat{k}_i} \log(e^{-F+al} e^{-\exp(-F+al)\nu_{i,l}}) \tag{5.4}$$

For each i, this maximisation is equivalent to the solution of a log-linear regression modelin which the l observed responses are the random variables $\mathbf{Y}^i = \{\frac{1}{\nu_{i,l}}\}_l$ distributed exponentially with expected value $\langle Y_l^i \rangle = e^{-F_i+al}$. Therefore, F_i is the intercept of such model and this procedure makes the estimator (empirically) *punctual*, as the $l \to 0$ limit is practically equivalent to taking the limit for $h \to 0$. Under these conditions, it is not anymore necessary to take the average of the bias factors $e^{\beta B(\mathbf{s}(\mathbf{x}))}$ over the neighbourhood of $\boldsymbol{\sigma}(\mathbf{x}_i)$ set by a finite h: the reweighting involves only the punctual value of the bias applied when generating a datapoint i, without even necessity to specify the underlying CV $\mathbf{s}(\mathbf{x}_i)$. As we will show, this allows removing the effect of the bias in a simple, numerically robust and theoretically well-founded manner, also in the case in which the CV on which the bias is applied is not an explicit function of the variables $\boldsymbol{\sigma}$.

5.2 Punctual reweighting using PAk

5.2.1 Analytic conditions for punctual reweighting in Umbrella Sampling

Looking at equation 5.1 we see that if there exists a map $\boldsymbol{\sigma} \mapsto \hat{\mathbf{s}}(\boldsymbol{\sigma})$ associating to each point $\boldsymbol{\sigma}(\mathbf{x})$ a unique value $\mathbf{s}(\mathbf{x}) = \hat{\mathbf{s}}(\boldsymbol{\sigma}(\mathbf{x}))$ then $B(\mathbf{s}(\mathbf{x}))$ can be formally expressed as an explicit function of $\boldsymbol{\sigma}(\mathbf{x})$ and the exponential factor $e^{\beta B(\hat{\mathbf{s}}(\boldsymbol{\sigma}))}$ can be brought out of the integral. Hence Equation 5.1 for $\tilde{\boldsymbol{\sigma}} = \boldsymbol{\sigma}(\mathbf{x}_i)$ takes the form

$$F(\boldsymbol{\sigma}(\mathbf{x}_i)) = F^B(\boldsymbol{\sigma}(\mathbf{x}_i)) - B(\mathbf{s}(\mathbf{x}_i)). \tag{5.5}$$

For future reference, we call the existence of such map the Map-Existence Condition (MEC). A consequence of the MEC is that if for two configurations $\mathbf{x}_1, \mathbf{x}_2$ we have $\sigma(\mathbf{x}_1) = \sigma(\mathbf{x}_2)$ then one cannot have $\mathbf{s}(\mathbf{x}_1) \neq \mathbf{s}(\mathbf{x}_2)$. Again, we stress the MEC is required only for the configurations $\mathbf{x}$ in the thermal ensemble. In fact, in molecular systems the interactions among atoms strongly reduce the independent directions in which the system can move. For this reason the ensemble density $\rho(\mathbf{x})$ is almost vanishing on a big portion of $\mathbb{R}^N$. The simplest case in which the MEC is verified is for $\mathbf{s}(\mathbf{x}) = \mathbf{s}(\boldsymbol{\sigma}(\mathbf{x}))$, namely when $\mathbf{s}$ is an explicit functions of the coordinates $\boldsymbol{\sigma}$. In this case $\hat{\mathbf{s}} \equiv \mathbf{s}$. However, equation 5.5 can also be valid if ρ^B is estimated on the $\boldsymbol{\sigma}$ but $\mathbf{s}$ is an explicit function of different coordinates σ', as long as these can be expressed as function of σ. This is true if all relevant $\boldsymbol{\sigma}'$ can be parametrised by an explicit function $\boldsymbol{\sigma}' = \varphi(\boldsymbol{\sigma})$. In this case $\mathbf{s}(\mathbf{x}) \equiv \mathbf{s}(\boldsymbol{\sigma}'(\mathbf{x})) \equiv \mathbf{s}(\varphi(\boldsymbol{\sigma}(\mathbf{x})))$ and $\hat{\mathbf{s}} \equiv \mathbf{s} \circ \varphi$.

5.2.2 Reweighting with a punctual estimator

In the cases where equation 5.5 holds, if one is able to estimate $F^B(\boldsymbol{\sigma}(\mathbf{x}_i))$ via an unbiased punctual estimator $\hat{F}_i^B$, the unbiased free energy at point i can be estimated as:

$$\hat{F}_i := \hat{F}_i^B - B_i \tag{5.6}$$

where $B_i := B(\mathbf{s}(\mathbf{x}_i))$ is simply the numerical value of the bias applied when generating datapoint $\mathbf{x}_i$. Equation 5.6 applies in principle to any punctual estimator of the biased free energy. By choosing a suitable $\hat{F}_i^B$, the meaning of $\hat{F}_i$ becomes operatively clear. We propose to estimate $\hat{F}_i^B$ with PAk, because, as proven in section 3.2.2, it is punctual without explicitly taking the limit $h \to 0$, thanks to the extrapolating features of its likelihood optimisation, depicted in Figure 3.2. Rephrasing it, for all points $\{\boldsymbol{\sigma}_i\}_i$ in a sample, it provides an unbiased estimate of the free energy $F(\boldsymbol{\sigma}_i)$ in equation 5.1 rather than the one in equation 5.2.

With this specification, equation 5.6 defines a simple punctual reweighting scheme to estimate point by point the unbiased free energy of a set of data generated in a biased simulation. From now on we shall for brevity refer to this procedure as **bPA*k***. We will show that the procedure defined in equation 5.6 gives consistent result even when the MEC is slightly violated, i.e. when $\mathbf{s}$ is not an explicit function of the $\boldsymbol{\sigma}$, but there exists a parametrisation φ that, given some $\boldsymbol{\sigma}$, is able to capture most relevant features in the space of the $\boldsymbol{\sigma}'$s.

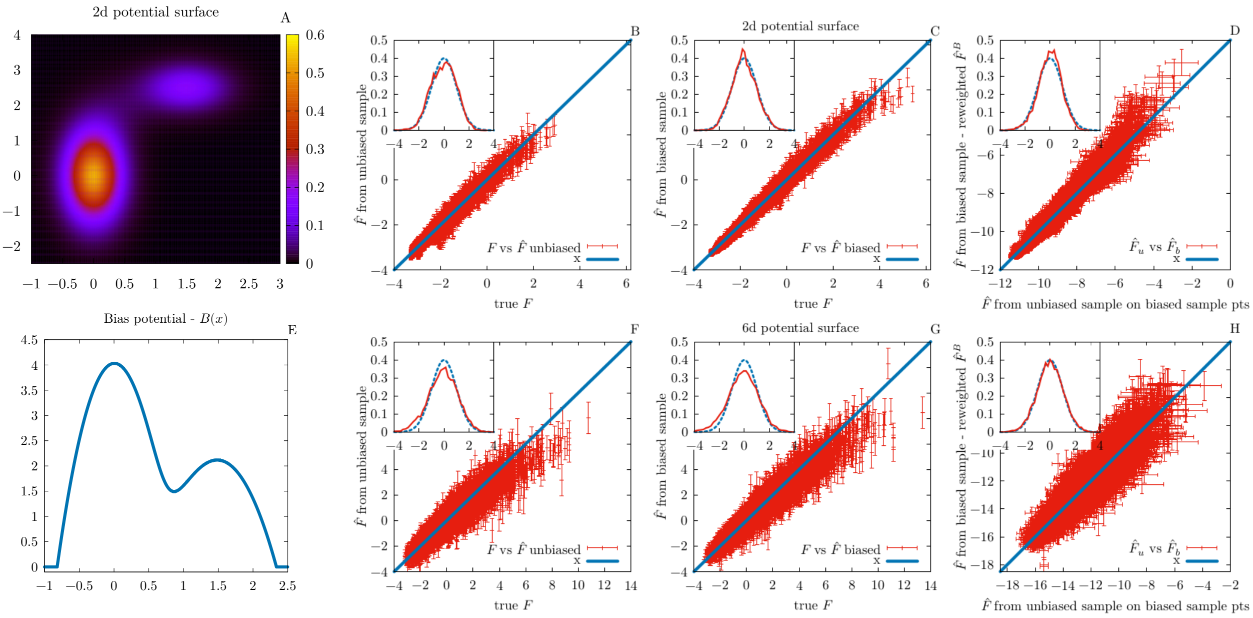


Figure 5.1: Comparison of bPAk performance to PAk's on analytic potentials. We consider two functional forms: the 2-dimensional double well potential and the 6-dimensional potential in Appendices A.2 and A.6. In both cases we sample 10.000 points from a biased and an unbiased simulation. First column: (A) 2-dimensional double-well potential surface; (E) bias potential used along x coordinate in the biased run for both analytic potentials. Second and third column: comparison of PAk and bPAk estimates against the analytic free energy showing correlation plots and pull distribution; (B),(C) respectively unbiased and biased case for 2d potential; (F),(G) respectively unbiased and biased case for 6d potential. (D),(H) statistical tests comparing bPAk to PAk on the points of the biased samples for 2d and 6d potentials respectively.

5.3 Validation of bPAk

In order to validate the robustness of the punctual reweighting procedure that we called bPAk we compare its performance to that of PAk on unbiased samples, since PAk is already established as a good free energy estimator on multidimensional data[157]. Thus, we choose our test systems such that we are able to generate both an unbiased and a biased equilibrium sample. In order to assess the statistical compatibility of the results obtained with PAk and bPAk, we use, as done in section 3.2.3.2, the correlation plot between estimated and ground truth free energies and the distribution of the pull in equation (3.16). Using these tools we first of all compare the estimates of free energy in the case of PAk and bPAk directly to the true known value in the case of multidimensional analytic potentials that we sample numerically. Secondly, still in the analytic case, we compare directly the two estimators. Finally, we consider as realistic case MD simulations of the CLN025 decapeptide; in this case there is no known ground truth free energy for the system, therefore the only sensible test is to directly compare the unbiased and the biased estimators.

5.3.1 Data sampled from multidimensional analytic potential surfaces

As a first step, we test bPAk on systems for which the ground truth potential is known analytically. We consider two functional forms: the bidimensional double well potential shown in Figure 5.1A, and the 6-dimensional potential which is identical to the bidimensional one in the first two directions. Thus, to bias the dynamics in both cases we can apply as bias potential the inverse of the analytic free energy along the x axis, cutoffed such that it is non-zero only on a finite interval (Figure 5.1E). We apply PAk to the unbiased sample and bPAk to the biased sample, obtaining two estimates of the free energies, which in these two cases should coincide with the analytic potential energy. We test the two estimators directly against the analytically known ground truth with the statistical tests described in section 5.3.

Looking first at the 2-dimensional case, Figures 5.1B and 5.1F, the normal distribution with unitary variance and zero mean seems in both cases well approximated; the correlation is good, with the difference that the biased simulation explores regions at higher free energy, as expected. Also in the 6-dimensional case, Figures 5.1C and 5.1G, the agreement with the ground truth potentials is good. While for PAk (unbiased case) this had previously been shown, also for bPAk we can conclude that the method performs excellently, at least in these two simple cases, and gives a statistically unbiased estimate of the correct free energies.

We now consider the case in which the ground truth unbiased free energy is not known explicitly, but one is able to generate a sample of data from the unbiased distribution. In this case, the unbiased and biased simulations sample different sets of points. Thus, since PAk only gives a punctual estimate of the free energy, we need to resort to the interpolation scheme described in section 3.2.6 of Chapter 3, PAk interpolator, in order to directly compare PAk and bPAk results. We apply PAk interpolator from the unbiased sample to compute the interpolated free energy on the biased sample points. Then we apply bPAk procedure to biased points, i.e. we estimate the biased free energy with PAk and then reweight punctually according to equation (5.2). This procedure works even in the case we only know the value of the bias potential $\{B_i\}_i$ on sample points.

Figure 5.1 (D),(H) shows the results of this test for the 2d and the 6d potentials; all samples generated have 10.000 points. Excellent compatibility between PAk and bPAk for both potentials is displayed. The example illustrates that the compatibility between a free energy estimated with and without the bias can be demonstrated also if the free energy is not known analytically. This will be used to analyse the simulation of the peptide.

5.3.1.1 Comparison with standard reweighting on a finite neighbourhood

We compare the performance of bPAk with that of other nonparametric methods in the estimate of the free energy from biased data. Firstly, we compare the results of bPAk to those of k-NN reweighted in the standard way illustrated in equation 5.3, i.e. by subtracting to the estimated biased free energy the quantity $\log\langle e^{\beta B(\mathbf{s}_j)}\rangle_j$. Secondly, we apply the punctual reweighting also

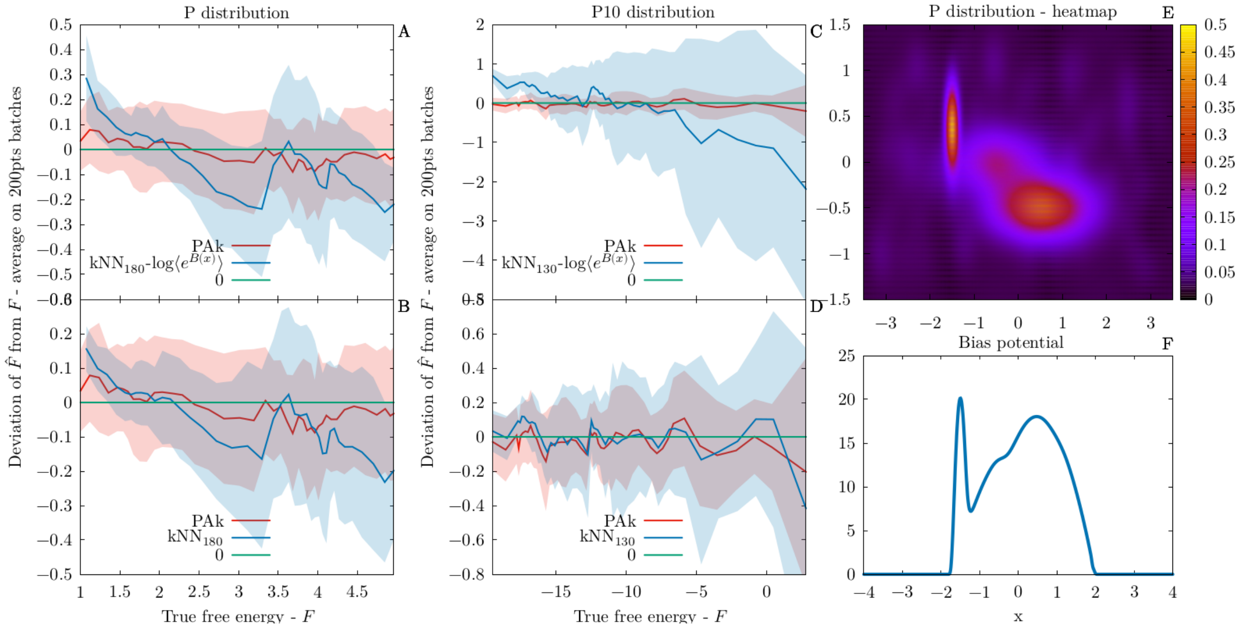


Figure 5.2: Comparison of reweighting protocols on rough analytic potentials. Comparison of different biased free energy estimators and reweighting protocols on datasets sampled from two distributions: p and p^{10}. (E) Heatmap representation of unbiased p. (F) Bias applied to p^{10}, which is exactly ten times the one applied to p. In both cased $d = D = 2$, so what we call unbiased free energy corresponds identically to the potential energy entering the Boltzmann factor if p and p^{10} are interpreted as canonical p.d.f.'s. (A-D) . All estimates are performed on 10000 data points. The values represented on the vertical axes are averages over batches of 200 points. In solid transparent bars the sample standard deviation of the batch. (A),(B) refer to p; (C),(D) refer to p^{10}. The chosen values of 180 and 130 for the k-NN estimators are chosen as the average optimal value of k predicted by PAk for the datasets. (A),(C) bPAk compared to k-NN with standard reweighting. (B),(D) bPAk compared to k-NN, both with punctual reweighting

to k-NN, for which this ansatz is not justified by a demonstration of punctuality of the density estimator, but becomes correct only in the limit $h \to 0$.

We use two datasets in 2 dimensions: one sampled from the probability density function p represented in Figure 5.2E; the other one sampled from p^{10} re-normalised to 1, which for further reference we simply indicate by p^{10}. The two systems display metastability between the two main basins. By construction, the potential barrier in p^{10} is exactly 10 times as high as the one in p. The simulations are biased along the x coordinate, with a bias obtained from the histogram along x of the unbiased sampling (see Figure 5.2F).

For both distributions, bPAk drastically outperforms the k-NN method reweighted in a standard manner (Figures 5.2A,5.2C). With punctual reweighting, the quality of k-NN estimates of the unbiased free energy improve. However, bPAk still shows a better performance along the whole range of free energy values, especially for high values (Figures 5.2C,5.2D). Furthermore, bPAk yields much better relults also looking at the pull distributions between the estimated and the ground truth free energies in all reweighting schemes, a sign that bPAk estimator is preferable to standard k-NN also in terms of error estimate.

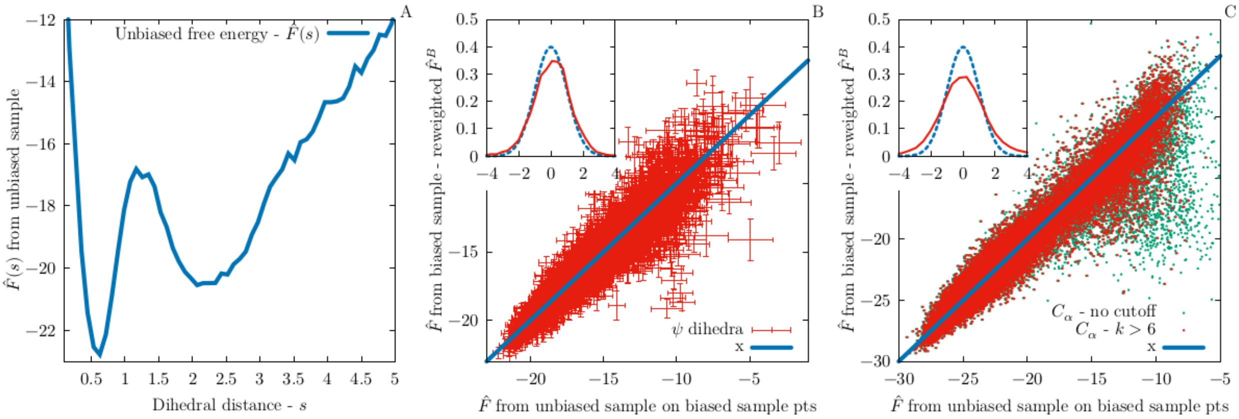


Figure 5.3: Performance of bPAk on a realistic system: CLNO25. All free energies in this figure are measured in $kJmol^{-1}$. (A) Free energy of the CLN025 peptide computed as histogram of the collective variable s. (B),(C) statistical tests comparing bPAk to PAk on the points of the biased sample in the case of the ψ-dihedral angles and of the C_α distances respectively. In (C) the error bars have been omitted from the correlation plot for a better readability; green dots represent all the points in the biased dataset; the red dots neglect all points with $\hat{k}_i \leq 6$.

5.3.2 Application of bPAk on an all-atom based simulation of a peptide

We test our method also on a realistic system, namely the MD simulation of decapeptide CLN025 discussed in Appendix A.7. In order to bias the trajectory, we choose as collective variable the ψ-dihedral distance from an equilibrium configuration, defined as:

$$s = \sum_{n=1}^{9} \frac{1 - cos(\psi_n - \psi_n^{ref})}{2} \tag{5.7}$$

where ψ_n denotes the n-th backbone ψ-dihedral angle of the peptide in the present configuration and ψ_n^{ref} is the value of the same dihedral angle in a chosen reference equilibrium configuration (in our case we chose the crystal structure[98]); this CV takes values from $s = 0$, in the reference configuration, to $s = 9$. We evaluate s along the trajectory and compute the free energy $F(s)$ from a histogram (Figure 5.3A). We fit the lowest part ($\sim 10\ kJmol^{-1}$) of the free energy with a sum of Gaussians $\tilde{F}(s) + c$ and use the negative of such sum as bias potential $B(s) = -\tilde{F}(s)$. Using PLUMED[22] we run a umbrella sampling biased REMD simulation in the same setting of the unbiased one.

We analyse the two trajectories in the 9-dimensional ψ-backbone-dihedra space (see Appendix A.7.1). This choice implies of course a drastic dimensional reduction on the over-400-dimensional raw coordinate space, the atomic configuration space; still, even after this huge projection the system will show complex features and a reasonably high dimensionality, so that we are entitled to consider it a realistic case. We extract 9.500 points from the unbiased trajectory to use them as reference sample sample for PAk Interpolator and 3.700 points from the biased one to be used as test sample. The estimated intrinsic dimension of the dataset is $d \sim 7$. The comparison between

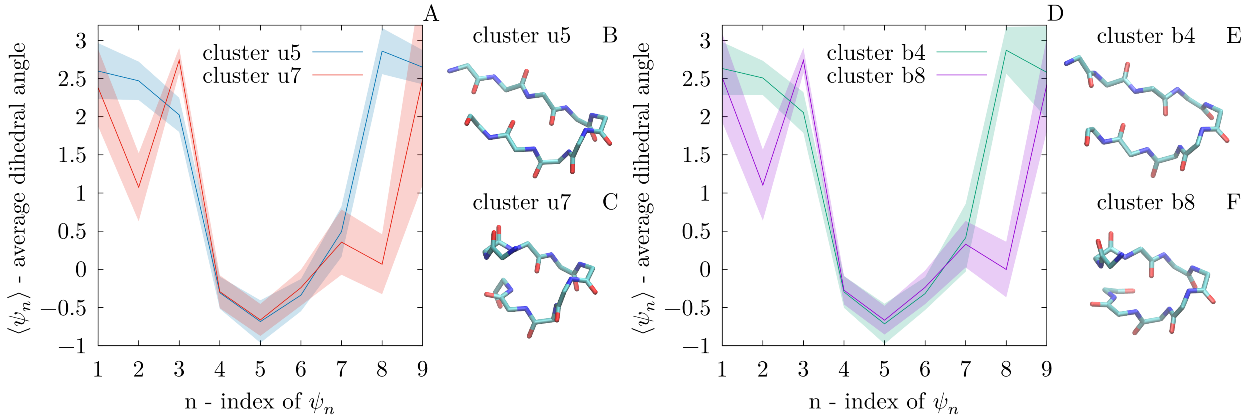


Figure 5.4: Preservation of cluster structure obtained from the biased and the unbiased trajectories of a small peptide. Comparison between the two main clusters of the unbiased and of the biased trajectories. (A),(D) Average dihedral angle for each of the backbone dihedra for each cluster. (B),(C),(E),(F) backbone visualisation of the configurations closest in dihedral distance to the cluster average.

bPAk estimate and our ground truth free energy (output of PAk Interpolator), in Figure 5.3B, shows excellent agreement, despite the high dimensionality.

5.3.2.1 Robustness of bPAk under change of metric

We finally test the robustness of bPAk using a different coordinate system, in which $\mathbf{s}$ is not anymore an explicit function of the $\boldsymbol{\sigma}$. We choose as coordinates the distances among alpha carbons (C_α), as discussed in Appendix A.7.2, thus the embedding dimension of the chosen space is now $D = 45$. This time, due to the much higher dimensionality of the embedding space, we feed our estimators with higher statistics: we take 37.000 reference points from the biased sample and 38.000 from the unbiased one. The estimated intrinsic dimension is $d = 9$. We carry out our protocol and show the result of the statistical tests in Figure 5.3C, green dots. Looking at the correlation plot we see a divergence from linearity especially at higher values of the free energy. We know that such values are associated to low densities in configuration space, corresponding either to low $\hat{k}_i$ or to high values for the distance of the $\hat{k}_i$-th neighbour. We notice (Figure 5.3C, red dots) that neglecting all points with $\hat{k}_i \leq 6$, corresponding to $\sim 10\%$ of the data, the correlation plot improves sensibly. Our explanation is that C_α-distance and ψ-dihedra metrics are equivalent in the sense defined in section 5.2.1 for low free energy configurations, more structured and dense in phase space, while the MEC is partly violated at high free energy values. As a sidenote, the choice of $\hat{k}_i \leq 6$ as cutoff value is justified by quantitatively measuring the percentage of outliers at 2 sigma (pull > 2) as a function of $\hat{k}_i$: systematic errors stop showing dependence on $\hat{k}_i$ around the value 6.

5.3.2.2 Cluster analysis

As a last test to assess whether bPAk captures the correct features of the free energy landscape we compare a cluster analysis in PAk and bPAk. As a clustering method we use Density Peak[158],[61]. In both cases we found eight clusters, but the two main clusters contain together more than 70% of all the configurations. These clusters contain the most structured configurations, closest to the native state; they correspond in fact to the leftmost basin in the free energy of Figure 5.3A. Looking at Figure 5.4 we see both quantitatively and qualitatively that cluster u5 almost perfectly matches cluster b4 and the same happens with u7 and b8. Both couples of clusters present a structured α-helix turn in the central region of the peptide (dihedra 4-6); in the case of u5 and b4 the tails (dihedra 1-3 and 8-9) lie in the β-domain, hence the β-hairpin is fully folded; in clusters u7 and b8 the tails are unstructured; as expected, the first couple is also energetically slightly favoured. As for our initial purpose, we can conclude that bPAk preserves also the cluster structure of the free energy.

5.4 Discussion

We have presented bPAk, a procedure to estimate the free energy in high-dimensional spaces starting from a sample of points generated in a biased simulation. The approach is based on PAk free-energy estimator. Besides all PAk aspects discussed in section 3.2.4, such as e.g. the nonnecessity of defining CVs and the restriction to the intrinsic data manifold, the features that are crucial in order to have a punctual reweighting are those that make PAk estimates more local than in the case of other nonparametric methods. In particular, on one hand PAk optimally selects for every point in the dataset the size of the neighbourhood considered in the free energy estimate; on the other hand, its likelihood maximisation extrapolates the value of the free energy in the limit of neighbourhood size going to zero, which makes the estimate punctual, differently from other kernel-based methods.

The bPAk protocol consists of computing the biased free energy at all points in the dataset applying PAk to the biased sample and then reweighting this quantity point by point simply subtracting the numerical value of the applied bias to retrieve an unbiased estimate. The simple additive form of this reweighting procedure is a nontrivial result. First of all, it crucially relies on the punctuality of PAk. Second, since it involves integration over degrees of freedom which are not necessarily explicitly orthogonal to those involved in the computation of the bias potential, some reasonable but necessary requirements must be satisfied. We have described the condition under which it is possible to reweight in an Umbrella-Sampling fashion the biased free energy over the coordinates $\boldsymbol{\sigma}(\mathbf{x})$ when the applied bias potential is a function of some possibly different CVs $\mathbf{s}(\mathbf{x})$. In short, this is possible if all the information necessary to define the biasing CVs $\mathbf{s}$ is encoded in the coordinates $\boldsymbol{\sigma}$ over which the free energy is computed; in other words, if the manifold the $\{\mathbf{s}_i\}_i$ can be mapped to a submanifold of the manifold the $\{\boldsymbol{\sigma}_i\}_i$ lie on. A case in which this condition is violated on a small subset of configurations has been presented in the case of the CLN025 peptide.

We have tested bPAk comparing its results to various unbiased ground truth free energies, some known analytically, some estimated with PAk from an unbiased simulation. In all tested cases, the pull distribution proved bPAk to be an unbiased estimator of the ground truth values. In the case of two analytically-known distributions, we have also compared the performance of bPAk to that of other estimators, both with standard and with punctual reweighting. While bPAk is confirmed to be our best choice, punctual reweighting visibly improved the estimates also in the case of finite-size-kernel estimators, where in principle one should take a suitable average of the bias in the neighbourhood defined by the kernel width. The opportunity of adopting punctual reweighting even with non-punctual estimators could be further investigated, but this is beyond the scope of this work.

Chapter 6

Beyond PAk: including free energy derivatives information

In this chapter we present a free energy estimation method developed by us in order to improve the performance of the PAk estimator, introduced in Chapter 3. In section 3.2.5 we have discussed some of the drawbacks affecting this estimator. In section 3.2.6 of the same chapter we have proposed a scheme to overcome the first of its limitations, the fact that PAk does not generalise outside the data sample. In Chapter 5 we have presented a framework to apply PAk with an efficient reweighting scheme for the analysis of samples generated under the action of an external bias, which was lacking in PAk's original formulation. We here focus on the problem of the roughness of PAk, induced, we recall, by the fact that the free energy is estimated independently for each data point, even when they are neighbours. The approach described here, still unpublished, is based on the estimation of free energy differences δF within small neighbourhoods over which the PDF is approximately quadratic. This estimation, in turn, relies on an accurate estimation of the free energy gradient. The estimation of the free energy gradient at a given point is illustrated in section 6.1, while the estimation of the δFs is dealt with in section 6.2. In section 6.3 we introduce the free energy calculation method that we named *Binless Multidimensional Thermodynamic Integration*.

6.1 Estimating the free energy gradient via a $\hat{k}$NN kernel

Let us first consider a distribution $\tilde{\rho}$ varying linearly along a direction (indicated by its gradient) in a given region of configuration space Ω_i centred around point $\mathbf{x}_i$. For any point $\mathbf{x}$ in Ω_i:

$$\tilde{\rho}(\mathbf{x}) = \tilde{\rho}(\mathbf{x}_i) + \nabla_{\mathbf{x}}\tilde{\rho}(\mathbf{x})|_{\mathbf{x}_i}(\mathbf{x} - \mathbf{x}_i) \ . \tag{6.1}$$

In these conditions the gradient of the density is proportional to the mean shift[75] around the central point:

$$\nabla_{\mathbf{x}}\tilde{\rho}(\mathbf{x}_i) := \nabla_{\mathbf{x}}\tilde{\rho}(\mathbf{x})|_{\mathbf{x}_i} \propto \langle(\mathbf{x}-\mathbf{x}_i)\rangle_{\tilde{\rho}} = \frac{\int \tilde{\rho}(\mathbf{x})(\mathbf{x}-\mathbf{x}_i)\,\mathrm{d}\mathbf{x}}{\int \tilde{\rho}(\mathbf{x})\,\mathrm{d}\mathbf{x}}. \tag{6.2}$$

We now show how accurate is the approximation 6.2 for a generic PDF, in which also quadratic or terms are present. Let us consider the Taylor expansion of a density $\rho(\mathbf{x})$ around point $\mathbf{x}_i$:

$$\rho(\mathbf{x}) = \rho(\mathbf{x}) = \rho(\mathbf{x}_i) \;+\; \nabla^{\mathrm{T}}_{\mathbf{x}}\rho(\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i) \;+\; \frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}\,\nabla^2_{\mathbf{x}}\,\rho(\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i) + \mathcal{O}\left((\mathbf{x}-\mathbf{x}_i)^3\right)\;. \tag{6.3}$$

The mean shift around point $\mathbf{x}_i$ within region $\Omega_i := B^d(r_i, \mathbf{x}_i)$ of volume V_d is defined as:

$$\langle(\mathbf{x}-\mathbf{x}_i)\rangle_{\Omega_i,\rho} := \frac{\int_{\Omega_i}\rho(\mathbf{x})(\mathbf{x}-\mathbf{x}_i)\,\mathrm{d}\mathbf{x}}{\int_{\Omega_i}\rho(\mathbf{x})\,\mathrm{d}\mathbf{x}}\;. \tag{6.4}$$

For a lighter notation we choose the specific case $\mathbf{x}_i = \mathbf{0}$, but the derivation remains valid also in the more general case. Inserting the expansion (6.3) into equation (6.4) and taking into account the results (D.3), (D.4) and (D.5):

$$\begin{aligned}
\langle(\mathbf{x}-\mathbf{x}_i)\rangle_{\Omega_i,\rho} &= \frac{\int_{\Omega_i}\rho(\mathbf{x})\,\mathbf{x}\,\mathrm{d}\mathbf{x}}{\int_{\Omega_i}\rho(\mathbf{x})\,\mathrm{d}\mathbf{x}} \\
&= \frac{\rho(\mathbf{x}_i)\cancelto{0}{\int_{\Omega_i}\mathbf{x}\,\mathrm{d}\mathbf{x}} + \;\nabla^{\mathrm{T}}_{\mathbf{x}}\rho(\mathbf{x}_i)\int_{\Omega_i}\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\,\mathrm{d}\mathbf{x} \;+\; \frac{1}{2}\,\nabla^2_{\mathbf{x}}\,\rho(\mathbf{x}_i)\cancelto{0}{\int_{\Omega_i}\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\,\mathbf{x}\,\mathrm{d}\mathbf{x}}}{\rho(\mathbf{x}_i)\int_{\Omega_i}1\,\mathrm{d}\mathbf{x} \;+\; \nabla^{\mathrm{T}}_{\mathbf{x}}\rho(\mathbf{x}_i)\cancelto{0}{\int_{\Omega_i}\mathbf{x}\,\mathrm{d}\mathbf{x}} + \;\frac{1}{2}\;\mathrm{Tr}\left[\nabla^2_{\mathbf{x}}\,\rho(\mathbf{x}_i)\int_{\Omega_i}\mathbf{x}\,\mathbf{x}^{\mathrm{T}}\mathrm{d}\mathbf{x}\right]} \;+\; \mathcal{O}(V_d\,r_i^4) \\
&= \frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}_i)\,\cancel{V_d}\,\frac{r_i^2}{d+2}}{\rho(\mathbf{x}_i)\,\cancel{V_d} \;+\; \frac{1}{2}\;\mathrm{Tr}\,\nabla^2_{\mathbf{x}}\,\rho(\mathbf{x}_i)\,\cancel{V_d}\,\frac{r_i^2}{d+2}} \;+\; \mathcal{O}(\cancel{V_d}\,r_i^4) \\
&= \frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}_i)\;\frac{r_i^2}{d+2}}{\rho(\mathbf{x}_i)\;\left(1 \;+\; \frac{\mathrm{Tr}\,\nabla^2_{\mathbf{x}}\,\rho(\mathbf{x}_i)}{2\,\rho(\mathbf{x}_i)}\;\frac{r_i^2}{d+2}\right)} \;+\; \mathcal{O}(r_i^4) \\
&= \frac{r_i^2}{d+2}\frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}_i)}{\rho(\mathbf{x}_i)}\left(1 - \frac{\mathrm{Tr}\nabla^2_{\mathbf{x}}\rho(\mathbf{x}_{\mathrm{i}})}{2\rho(\mathbf{x}_i)}\frac{r_i^2}{d+2}\right) \;+\; \mathcal{O}(r_i^4)
\end{aligned} \tag{6.5}$$

where the neglected integrals vanish for integration of an odd function on a symmetric domain. Therefore, having $\nabla_{\mathbf{x}}F(\mathbf{x}_i) = -k_BT\frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}_i)}{\rho(\mathbf{x}_i)}$ according to equation (3.8), the gradient of the free energy can be well approximated by:

$$\nabla_{\mathbf{x}}F(\mathbf{x}_i) \approx -k_BT\;\frac{d+2}{r_i^2}\;\langle(\mathbf{x}-\mathbf{x}_i)\rangle_{\Omega_i,\rho}\;. \tag{6.6}$$

Operatively, the mean shift on the right-hand side of equation (6.6) can be easily estimated as a sample average of the shift observable $(\mathbf{x}-\mathbf{x}_i)$ over the first $\hat{k}_i - 1$ NNs of $\mathbf{x}_i$ (see Appendix D.2.1), equation (D.9):

$$\langle(\mathbf{x}-\mathbf{x}_i)\rangle_{\Omega_i,\hat{\rho}} := \frac{1}{\hat{k}_i}\sum_{j=1}^{\hat{k}_i-1}(\mathbf{x}_j-\mathbf{x}_i)\ . \tag{6.7}$$

In these conditions, due to the specific neighbourhood selection, we expect the approximation in equation (6.5) to hold well. Notice that other radially-symmetric kernels can be employed in alternative to kNN[39, 46, 74]). By putting together equations (6.6) and (6.7) we recover a sample free energy gradient estimator:

$$\hat{\mathbf{g}}_i \;=\; \hat{\nabla_{\mathbf{x}}F}(\mathbf{x}_i) \;:=\; -k_BT\,\frac{d+2}{r_i^2}\,\frac{1}{\hat{k}_i}\sum_{j=1}^{\hat{k}_i-1}(\mathbf{x}_j-\mathbf{x}_i)\ . \tag{6.8}$$

A similar estimator was first proposed in reference [75] based on intuitive arguments. To the best of our knowledge, however, the explicit derivation of the the expression in equation (6.8) is not present in literature. Moreover, it is worth to stress that $\hat{\mathbf{g}}_i$ in our approach is adaptive in the same sense of the $\hat{k}$NN estimator (cfr. section 3.1), since it restricts to the intrinsic manifold of dimension $d \ll D$ and operates a doubly-adaptive bandwidth selection; this is probably the main reason of its successful performance, which will be illustrated in what follows. As a final remark, notice that, due to the finite size of the kernel bandwidth, $\hat{\mathbf{g}}_i$ is not exactly an unbiased estimator of the punctual free energy gradient $\nabla_{\mathbf{x}}F(\mathbf{x}_i)$, but in fact an estimator of the average quantity $\langle\nabla_{\mathbf{x}}F(\mathbf{x}_i)\rangle_{\Omega_i}$ over the region Ω_i, up to cubic order in the Taylor expansion of $\rho(\mathbf{x})$ (cfr. Appendix D.2.1.3); the latter quantity converges asymptotically to the true value $\nabla_{\mathbf{x}}F(\mathbf{x}_i)$ in the limit $r_i \to 0$.

6.1.1 Variance-covariance matrix of the gradients

The estimator $\hat{\mathbf{g}_i}$ is the sample average of a set of i.i.d random variables $\left\{-k_BT\,\frac{d+2}{r_i^2}\,(\mathbf{x}_j-\mathbf{x}_i)\right\}_{j=1}^{\hat{k}_i}$ whose mean value is $\mathbf{g}_i$. From the central limit theorem we know that the distribution of $\hat{\mathbf{g}}_i$ is a D-variate normal distribution whose variance-covariance matrix $\boldsymbol{\sigma}_i^2 := \boldsymbol{\sigma}^2[\hat{\mathbf{g}}_i] = \mathbf{cov}[\hat{\mathbf{g}}_{\mathrm{i}},\hat{\mathbf{g}}_{\mathrm{i}}]$ is proportional to $1/\hat{k}_i$ times the variance-covariance matrix of $(\mathbf{x}-\mathbf{x}_i)$ in Ω_i. The standard deviation of a gradient component $\hat{g}_{i,\alpha}$ is simply the squared root of the marginal of $\boldsymbol{\sigma}_i^2$ over the component α[62]. This marginal is estimated by:

$$\hat{\sigma}[\hat{g}_{i,\alpha}] = \frac{1}{\sqrt{\hat{k}_i}-1}\ \hat{\sigma}\left[\frac{d+2}{r_i^2}(x_\alpha - x_{i,\alpha})\right]\ , \tag{6.9}$$

Here, and in the following, by sample standard deviation of an observable $O(\mathbf{x})$ over a sample of

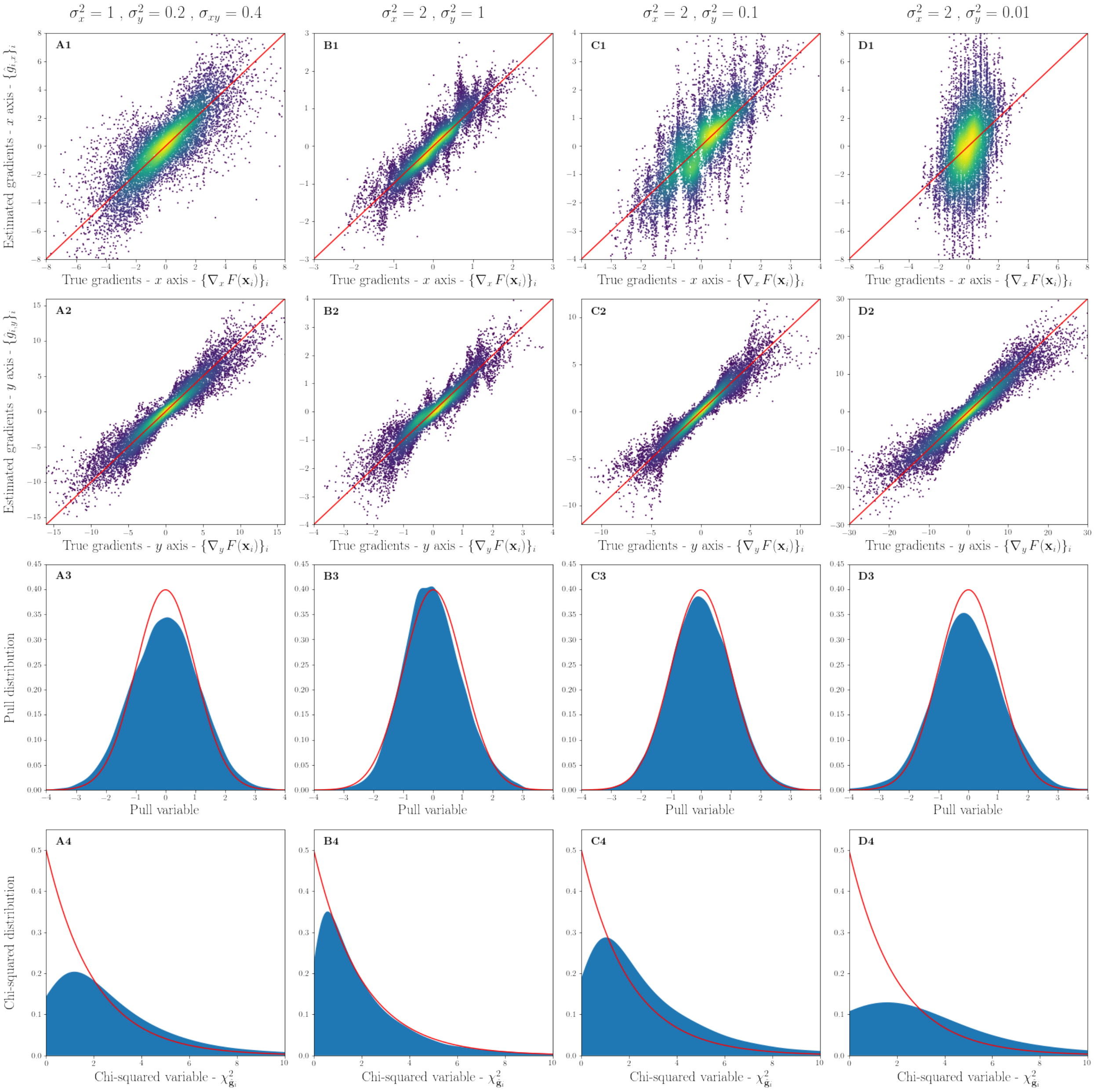


Figure 6.1: Free energy gradient components estimator performance tested on various bivariate Gaussian systems. All four systems considered, one for each column, have a bivariate normal PDF centered at the origin of the Cartesian plane (see Appendix A.5) sampled 10000 times. The entries of each system's covariance matrix are indicated in the column header. First row: correlation plots of estimated x gradient components against true values. In red the line $\hat{g}_{i,y} = g_{i,y}$. Second row: correlation plots of estimated y gradient components against true values. In red the line $\hat{g}_{i,y} = g_{i,y}$. Third row: distribution of the pull of gradient components from equation (6.10). In red the standard normal distribution $\mathcal{N}(0,1)$. Bottom row: distribution of the chi-squared variables $\hat{\chi}^2_{\hat{\mathbf{g}}_i}$ defined using equation (6.11) and the auto-covariance matrix estimated from the sample by $\hat{\boldsymbol{\sigma}}_i^2$ in equation (D.15). In red the analytical chi-squared distributions with 2 DOFs.

N_s datapoints we mean the square root of the sample variance defined:

$$\hat{\sigma}^2[O(\mathbf{x})] := \frac{N_s}{N_s - 1}\left[\frac{1}{N_s}\sum_{l=1}^{N_s} O^2(\mathbf{x}_l) - \left(\frac{1}{N_s}\sum_{l=1}^{N_s} O(\mathbf{x}_l)\right)^2\right] \approx \langle O^2(\mathbf{x})\rangle_\Omega - \langle O(\mathbf{x})\rangle_\Omega^2 = \hat{\sigma}^2[O(\mathbf{x})]_\Omega \ ,$$

where the prefactor $\frac{N_s}{N_s-1}$ is the so-called Bessel's correction for the unbiased sample variance estimator[193]. The variance-covariance of the estimator hence decreases with the number of sampled points as $1/\hat{k}_i$, while, as mentioned before, the estimator bias grows with the neighbourhood size r_i. We face also here the bias-variance trade-off problem. $\boldsymbol{\sigma}_i^2$ is a $D \times D$ matrix, where D can be typically large, while $\hat{k}_i$ can become quite small, especially in high-dimensions, so the estimates for the gradient covariance is typically very noisy. Luckily, we are often interested only on the values on the diagonal, corresponding to equation (6.9), which, as we will see in section 6.1.2, appear to be estimated reliably also in realistic settings. In Appendix D.2.2.1 one can find the a derivation of $\boldsymbol{\sigma}_i^2$ (cfr. equation (D.14)) and an expression for the sample estimator $\hat{\boldsymbol{\sigma}}_i^2$ (cfr. equation (D.15)).

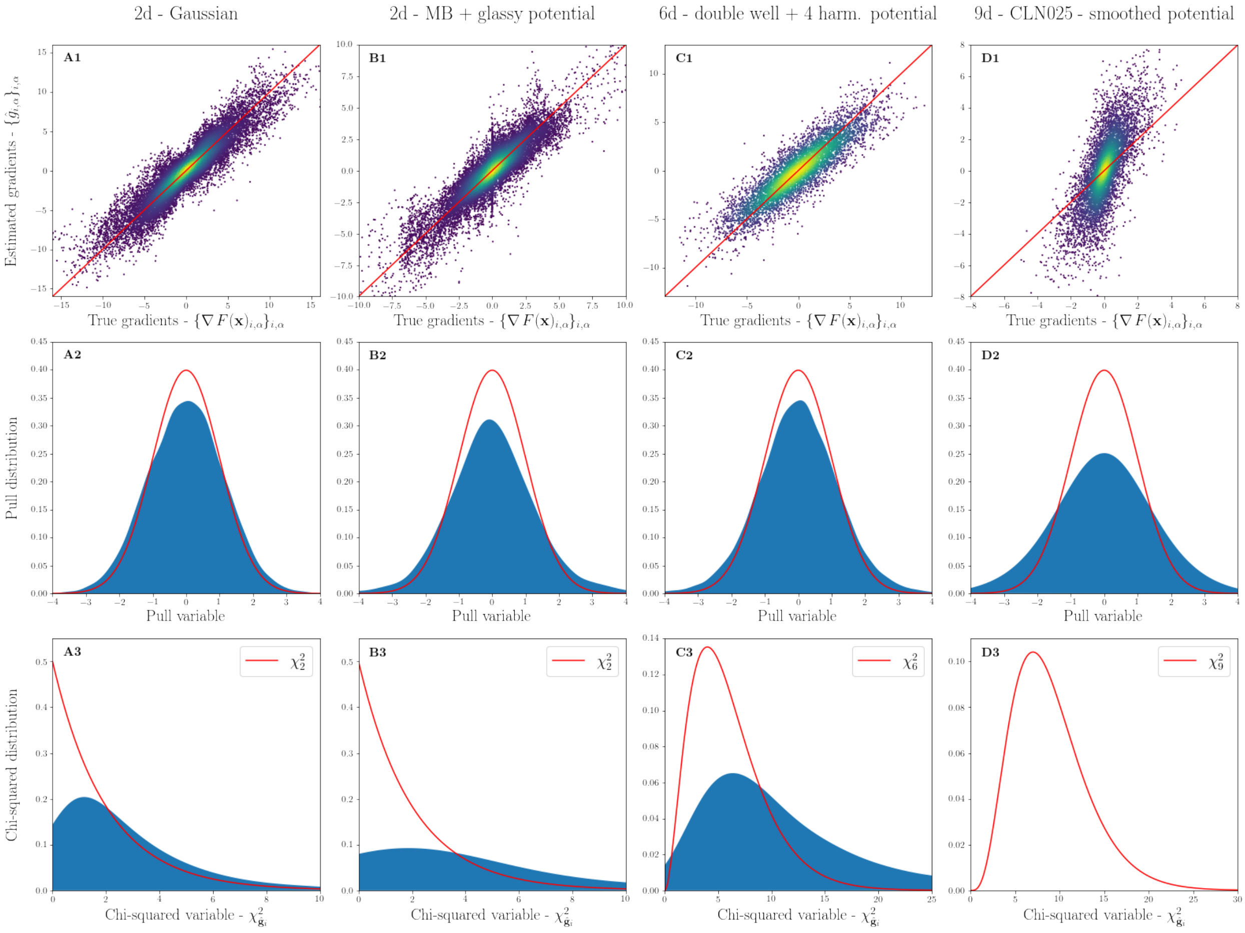
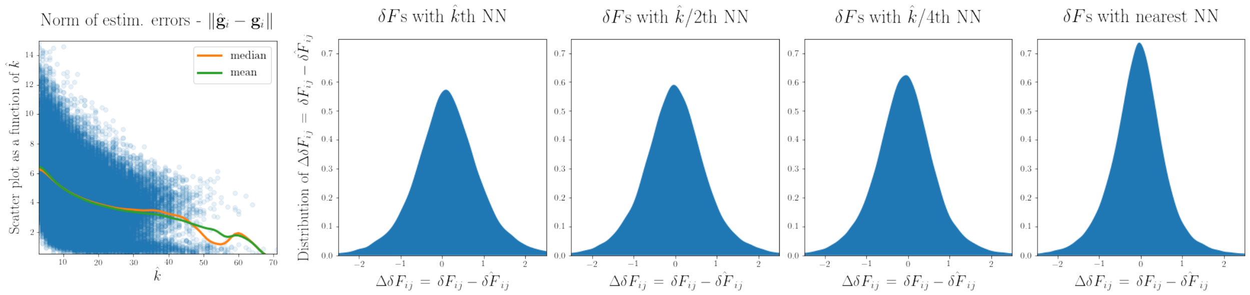


Figure 6.2: Free energy gradient estimator performance tested on various systems. The four systems, one for each column, are indicated in the column header; they are all described in Appendix A; their dimensionality goes from 2 to 9. For all of them, the analytic expression of the free energy gradient is known. In the fourth and last column, the nine-dimensional case, 80000 sample points are considered; for all other system the sample size is 10000. In the first column the system is the same considered in the first column of Figure 6.1. Top row: correlation plots of estimated gradient components against true values. In red the line $\hat{g}_{i,\alpha} = g_{i,\alpha}$. Middle row: distribution of the pull of gradient components from equation (6.10). In red the standard normal distribution $\mathcal{N}(0,1)$. Bottom row: distribution of the chi-squared variables $\hat{\chi}^2_{\hat{\mathbf{g}}_i}$ defined using equation (6.11) and the auto-covariance matrix estimated from the sample by $\hat{\boldsymbol{\sigma}}_i^2$ in equation (D.15). In red the analytical chi-squared distributions with number of DOFs equal to the embedding dimension of the system $\nu = D$. Panel D3 does not show the measured distribution because the computed auto-covariance matrix is too noisy to provide a sensible inverse.



(a)

(b)

Figure 6.3: Effect of the neighbourhood size $\hat{k}$ on the accuracy of the estimators of the gradients and of the δFs. In all panels, the system considered is the 9-dimensional smoothed FES of the CLN025 decapeptide in the ψ-dihedrals space (cfr. Appendix A.7.1.1). **(a)** We plot for each point i the norm of the vector difference between the estimated gradient $\hat{\mathbf{g}}_i$ and the true one $\mathbf{g}_i$ as a function of the neighbourhood size $\hat{k}_i$: $\|\hat{\mathbf{g}}_i - \mathbf{g}_i\|$. In yellow and green the median and mean of this quantity for all points having the same $\hat{k}$. **(b)** Distribution of the error on δF estimates w.r.t to true values $\Delta\delta F_{ij} := \widehat{\delta F}_{ij} - \delta F_{ij}$ as a function of the neighbourhood rank of point j w.r.t. i. The rightmost panel shows the distribution of the error in the prediction of the free energy differences between nearest neighbours. The rank increases going from left to right until the $\hat{k}$th nearest neighbour is considered in the leftmost panel of subfigure (b).

6.1.2 Performance of the gradient estimator

In order to illustrate the performance of the gradient estimator we use two representations we have already adopted in Chapters 3 and 5: the correlation plot and the Gaussianity test in form of the pull distribution. Moreover, we compare the empirical chi-squared of the estimated gradients to the theoretical distribution, as we will explain briefly. All these tools allow us to condensate a lot of information in simple plots. This is particularly useful in the case of gradients, since we are estimating a vector quantity and moreover it has the dimensionality of the embedding space D.

In the case of the correlation plots we can either condense all couples of points:

$$\left\{\{(g_{i,1}, \hat{g}_{i,1})\}_{i=1}^{N}, \{(g_{i,2}, \hat{g}_{i,2})\}_{i=1}^{N}, \ldots, \{(g_{i,D}, \hat{g}_{i,D})\}_{i=1}^{N}\right\}$$

on a single plot, as done in Figure 6.2, or inspect the behaviour along the various directions in D different plots, as done for various two-dimensional Gaussian systems in Figure 6.1, choosing Cartesian coordinates or any other suitable coordinates system.

As per the pull distribution, we use it as the simplest way to assess the performance of the gradient estimator and of its error. In particular, since the estimated gradient components $\hat{g}_{i,\alpha}$ are all IID RVs, the pull for the gradient components:

$$\chi_{i,\alpha} := \frac{g_{i,\alpha} - \hat{g}_{i,\alpha}}{\sigma[\hat{g}_{i,\alpha}]} \tag{6.10}$$

should be distributed as a standard normal variable: $\chi_{i,\alpha} \sim \mathcal{N}(0,1)$. We are able to estimate the gradient standard deviation via $\hat{\sigma}[\hat{g}_{i,\alpha}]$ in equation (6.9), so we can use it in equation (6.10) to check its accuracy.

An alternative test for the estimated vector quantities $\hat{\mathbf{g}}$ is looking at the distribution of the chi-squared variable:

$$\chi^2_{\hat{\mathbf{g}}_i} := (\hat{\mathbf{g}}_i - \mathbf{g}_i)^{\mathrm{T}} \cdot \mathbf{cov}^{-1}[\hat{\mathbf{g}}_{\mathrm{i}}, \hat{\mathbf{g}}_{\mathrm{i}}] \cdot (\hat{\mathbf{g}}_{\mathrm{i}} - \mathbf{g}_{\mathrm{i}}) \ \sim \ \chi^2_{\nu=\mathrm{D}} \ , \tag{6.11}$$

which, as indicated, should have a chi-squared distribution with D DOFs, since the standardised multivariate pull variable $\chi_{\hat{\mathbf{g}}_i} := \mathbf{cov}^{-\frac{1}{2}}[\hat{\mathbf{g}}_{\mathrm{i}}, \hat{\mathbf{g}}_{\mathrm{i}}] \cdot (\hat{\mathbf{g}}_{\mathrm{i}} - \mathbf{g}_{\mathrm{i}})$ should be distributed as a standard D-variate normal $\mathcal{N}(0, \mathbb{1}_D)$[86, 123]. If, instead of the analytical auto-covariance of the gradient estimators, we use the estimated one $\hat{\boldsymbol{\sigma}}^2_i$, we obtain $\hat{\chi}^2_{\hat{\mathbf{g}}_i}$. We use this approach to assess the quality of the estimated covariance.

Figure 6.1 illustrates the performance of the gradient estimator on four two-dimensional Gaussian probability distributions, with variances and covariances defined in the titles. In the top two rows we can see the correlation plots of the two estimated gradient components along the x and y axes against the true values. Looking at the parameters defining the distributions, in the column headers, we see that only the Gaussian in the first column has a non-diagonal covariance matrix. In the remaining columns the width of the Gaussian is kept fixed along the x direction, while it is reduced more and more going from left to right. Along the y axis we see that all estimates correlate well with the true values. Along the x axis, instead, estimates are noisier and noisier going from left to right, namely towards smaller variance along the y axis. Indeed, the gradient estimated via the sample mean shift (6.8) is good at capturing the gradient direction, but in these e datasets the gradient is mostly oriented along the y direction, so the relative error on the transverse direction is larger. Another way to understand this effect is that we are considering circular regions $\{\Omega_i\}_i$ in a anisotropic landscapes; in these conditions the approximation leading to the mean shift equivalence in equation (6.5) is violated and higher order corrections play a role, with a higher visible impact on the direction where the free energy varies more slowly. As for the Gaussian in the first column, since is orientation tilted w.r.t. and not aligned with any axis, the noise is present but is less structured in the correlation plot A1 with respect to the other examples. For all datasets the computed pull distribution for the gradient components, in the third row, is in good agreement with the standard normal distribution, a sign that our estimates are unbiased and that we correctly estimate their variance. Instead the chi-squared distributions in the bottom row display a fair agreement only in the second system; it gives poorer results, in order, in panels C4, A4 and D4, where fatter tails are observed w.r.t. the predicted distribution. This means that our estimates $\hat{\boldsymbol{\sigma}}^2_i$ of the gradients' covariance are very noisy. However, we anticipate that these estimated auto-covariances are sufficiently accurate to quantify the error of free energy *differences* between neighbouring points, possibly due to subtle error cancellation. As a side comment, the reason why in terms of quality of both the pulls and the chi-squared the system in the first column appears to underperform columns 2 and 3 is that the "aspect ratio" of the first Gaussian is somewhere in between the ones of the third and fourth.

Let us now turn to Figure 6.2, which shows the performance of the gradient estimator on four different model free energy landscapes (see Appendix) in terms of the correlation plot of estimated and true gradient components, the distribution of the pull of gradient components from equation (6.10) and the distribution of the chi-squared variables $\chi^2_{\hat{\mathbf{g}}_i}$ defined using equation (6.11) together with the sample autocovariance $\hat{\boldsymbol{\sigma}}^2_i$ defined in equation (D.15). In the correlation plots (top row), differently from Figure (6.1), all gradient components, from 1 to D, are plotted together. We can see that gradient estimates correlate quite well with the true analytical values. Only in the 9-dimensional case, in panel D1, there is a visible bias: it can happen in fact that the gradient modulus is overestimated for some points, which results in a correlation plot slightly tilted w.r.t. to the identity line. Taking a closer look, we notice that this happens for points with few neighbours: in Figure 6.3a we see how both the mean and the median norm of the vector difference between the estimated and the true gradient (the two curves are almost coincident) are decreasing functions of the neighbourhood size $\hat{k}_i$ of the points; the same is true for the biggest value of the quantity $\|\hat{\mathbf{g}}_i - \mathbf{g}_i\|$ as a function of $\hat{k}_i$. Indeed, the gradient of points with smaller neighbourhoods is affected by a large variance. The quality of the pull distributions in the second row testify that even in high dimensionality our error estimates are quite good. This will become even clearer in the next section. The reason why on the 2-dimensional potential in panel B2 the gradient estimator performs worse than in the 6-dimensional case, in panel C2, is because the former is designed to put a strain on estimators, being rugged and spiky, so that the selected neighbourhood size $\hat{k}_i$ is for many points quite small. In the bottom row, the observed chi-squared distributions are not satisfying; the same considerations made commenting Figure (6.1) apply.

6.2 Free energy differences between neighbouring points

We will now show that using the estimator $\hat{\mathbf{g}}$ in equation (6.8) it is possible to compute free energy differences δF_{ij} between neighbouring points $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$\delta F_{ij} := F_j - F_i \ . \tag{6.12}$$

One could be tented to express their free energy difference as the contraction between the estimated gradient and their vector difference $\mathbf{r}_{ij} := \mathbf{x}_j - \mathbf{x}_i$:

$$\delta F^i_{ij} := \nabla^{\mathrm{T}}_{\mathbf{x}} F(\mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i) = \mathbf{g}_i \cdot \mathbf{r}_{ij} \tag{6.13}$$

and the estimator version[1]:

$$\hat{\delta F}^i_{ij} := \hat{\mathbf{g}}_i \cdot \mathbf{r}_{ij} \ . \tag{6.14}$$

However, the gradients in the two points $\mathbf{g}_i$ and $\mathbf{g}_j$ can be different, so in principle $\delta F^i_{ij} \neq \delta F^j_{ij}$. The right quantity to contract with $\mathbf{r}_{ij}$ in order to obtain exactly δF_{ij} would be the average free energy gradient along their connecting segment $\int_0^1 F(\lambda(t))\|\lambda'(t)\|dt$ where $\lambda(t)$ is a parametrisation of the vector $\mathbf{r}_{ij}$; such quantity is well approximated, until third order terms in the Taylor expansion of the free energy become relevant, by the semisum of the gradients in the two neighbouring points, so that the free enegy difference δF_{ij} can be estimated as:

$$\hat{\delta F}_{ij} := \frac{\hat{\mathbf{g}}_i + \hat{\mathbf{g}}_j}{2} \cdot \mathbf{r}_{ij} \ . \tag{6.15}$$

6.2.1 Error estimates on the δFs

Defining $\boldsymbol{\sigma}^2_{ij} := \boldsymbol{\sigma}^2[\frac{\hat{\mathbf{g}}_i + \hat{\mathbf{g}}_j}{2}]$, simple error propagation from equation (6.12) gives $\varepsilon^2_{ij} := \sigma^2[\hat{\delta F}_{ij}] = \mathbf{r}^{\mathrm{T}}_{ij} \cdot \boldsymbol{\sigma}^2_{ij} \cdot \mathbf{r}_{ij}$, with:

$$\boldsymbol{\sigma}^2_{ij} = \frac{1}{4}(\boldsymbol{\sigma}^2_i + \boldsymbol{\sigma}^2_j + 2\,\mathbf{cov}[\hat{\mathbf{g}}_{\mathrm{i}}, \hat{\mathbf{g}}_{\mathrm{j}}]) \ , \tag{6.16}$$

where, recall, $\boldsymbol{\sigma}^2_i = \mathbf{cov}[\hat{\mathbf{g}}_{\mathrm{i}}, \hat{\mathbf{g}}_{\mathrm{i}}]$. However, we do not have a solid cross-covariance model for $\mathbf{cov}_{\hat{\mathbf{g}}_{\mathrm{i}}, \hat{\mathbf{g}}_{\mathrm{j}}}$ and it is impossible to estimate it rigorously from a single sample, which makes estimating the uncertainty of $\hat{\delta F}_{ij}$ far from trivial (refer to Appendix D.2.2.2 for a further discussion and a tentative model for the cross-covariance). Nonetheless, by calling ${\varepsilon^i_{ij}}^2 := \mathbf{r}^{\mathrm{T}}_{ij} \cdot \boldsymbol{\sigma}^2_i \cdot \mathbf{r}_{ij}$, we can express the uncertainty on $\hat{\delta F}_{ij}$ as:

$$\varepsilon^2_{ij} = \frac{1}{4}({\varepsilon^i_{ij}}^2 + {\varepsilon^j_{ij}}^2 + 2\,p_{ij}\,\varepsilon^i_{ij}\,\varepsilon^j_{ij}) \ , \tag{6.17}$$

where p_{ij} is the Pearson correlation coefficient between the estimators $\hat{\delta F}^i_{ij}$ and $\hat{\delta F}^j_{ij}$ seen as random variables, and specifically:

$$p_{ij} = \frac{\mathbf{r}^{\mathrm{T}}_{ij} \cdot \mathbf{cov}[\hat{\mathbf{g}}_{\mathrm{i}}, \hat{\mathbf{g}}_{\mathrm{j}}] \cdot \mathbf{r}_{\mathrm{ij}}}{\varepsilon^i_{ij}\,\varepsilon^j_{ij}} \ . \tag{6.18}$$

Like for the cross-covariance of the gradients, we cannot straightforwardly compute p_{ij} from a single

[1] A possible rewriting solely in terms of distances between points of the estimator in equation (6.14), which can be useful in a computational perspective, is reported in equation (D.19) of Appendix D

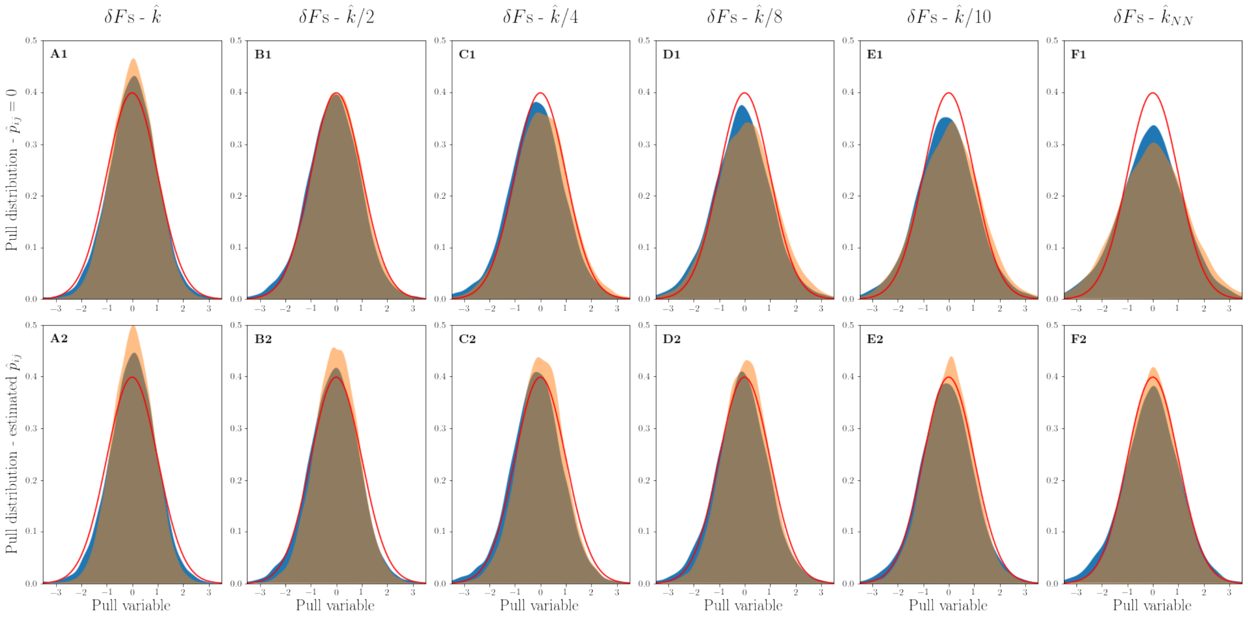


Figure 6.4: Effect of the estimated Pearson correlation coefficient $\hat{p}_{ij}$ on the pull distribution of the δFs at fixed neighbourhood rank values. Distribution of the pull variables $\hat{\chi}_{ij}$ in equation (6.21) for points j at fixed neighbourhood rank values w.r.t. point i. The systems considered are in blue the 6-dimensional potential (see Appendix A.6) and in yellow the bivariate Gaussian in column B of Figure 6.1. In red the standard normal distribution $\mathcal{N}(0,1)$. On the rightmost column, for each couple (i,j) considered, j is the NN of point i. The neighbourhood rank increases going left, so that e.g. the second column from the left considers points j whose neighbourhood rank w.r.t. i is $\hat{k}_i/2$ (rounded to the closest integer). Finally, on the leftmost column point j is the furthermost point from i within its neighbourhood. Top row: no Pearson correlation coefficient is used: $\hat{p}_{ij} = 0$. Bottom row: $\hat{p}_{ij}$ is the one in equation (6.19).

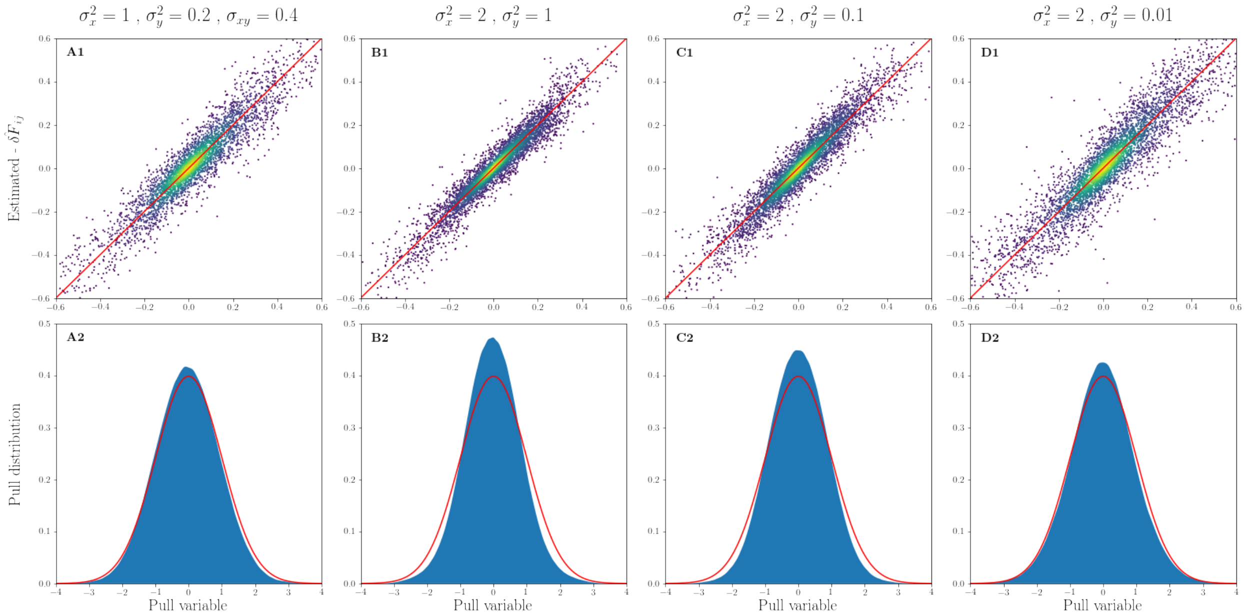


Figure 6.5: $\hat{\delta F}$ estimator performance tested on various bivariate Gaussian systems. The four systems considered, one for each column, are the same bivariate Gaussians considered in Figure 6.1. The entries of each system's covariance matrix are indicated in the column header. Top row: correlation plots of estimated $\{\hat{\delta F}_{ij}\}_{ij}$ against true values $\{\delta F_{ij}\}_{ij}$. In red the line $\hat{\delta F}_{ij} = \delta F_{ij}$. Bottom row: distribution of the pull variables $\hat{\chi}_{ij}$ in equation (6.21). In red the standard normal distribution $\mathcal{N}(0,1)$.

data sample, but we propose to estimate it giving it a geometrical interpretation: since p_{ij} must be 0 for uncorrelated estimates (which happen if the the d-hyperspheres Ω_i and Ω_j do not overlap at all), nonzero in case of overlapping neighborhoods $\Omega_i \cap \Omega_j \neq \emptyset$ and 1 if $\Omega_i \equiv \Omega_j$, we assume it to be well approximated by the ratio between the squared volume of their intersection $\Omega_i \cap \Omega_j$ and the product of the two volumes. Moreover, since $\hat{k}_i$ is proportional to the d-volume $V_i = \omega_d r_i^d$ of Ω_i via the relation $\hat{k}_i = V_i \, \hat{\rho}_i^{\hat{k}\mathrm{NN}}$, we define a proxy for p_{ij} as:

$$\hat{p}_{ij} = \frac{\hat{k}_{ij}^2}{\hat{k}_i \, \hat{k}_j} \ , \tag{6.19}$$

where $\hat{k}_{ij}$ is the number of points in common between the neighbourhoods of $\mathbf{x}_i$ and $\mathbf{x}_j$. Other possible estimators for p_{ij} are discussed in Appendix D.2.2.3. Together with the specification in equation (6.19), equation (6.17) guarantees that the uncertainty estimation on $\widehat{\delta F}_{ij}$ is accurate if its pull χ_{ij} is normally distributed:

$$\chi_{ij} := \frac{F_j - F_i - \widehat{\delta F}_{ij}}{\varepsilon_{ij}} \sim \mathcal{N}(0,1) \ . \tag{6.20}$$

The quantity in equation (6.20) is computed using estimated gradient auto-covariances in equation (6.17), so that $\hat{\varepsilon}_{ij}^{i\;2} := \mathbf{r}_{ij}^{\mathrm{T}} \cdot \hat{\boldsymbol{\sigma}}_i^2 \cdot \mathbf{r}_{ij}$ and $\hat{\varepsilon}_{ij}^2 := \frac{1}{4}(\hat{\varepsilon}_{ij}^{i\;2} + \hat{\varepsilon}_{ij}^{j\;2} + 2\,\hat{p}_{ij}\,\hat{\varepsilon}_{ij}^{i}\,\hat{\varepsilon}_{ij}^{j})$, thus

$$\hat{\chi}_{ij} := \frac{F_j - F_i - \widehat{\delta F}_{ij}}{\hat{\varepsilon}_{ij}} \ . \tag{6.21}$$

We examine the Gaussianity of this quantity in the next section.

6.2.2 Performance of the $\widehat{\delta F}$ estimator

Once again, in order to test our estimator $\widehat{\delta F}_{ij}$ we resort to correlation plots of $\widehat{\delta F}_{ij}$ against δF_{ij} and distributions of the pull variables in equation (6.20). Both in Figures 6.5 and 6.6 we see that all correlation plots and pull distributions are in excellent agreement with the predictions for unbiased estimators. From Figure 6.5 we see that the noise present in the gradient components estimates is strongly dumped. Also in Figure 6.6 we observe better pull distributions than for the gradient components. In this figure we have also represented, in columns B, C and E, three systems (cfr. Appendix A.8) for which we know the ground truth free energy values on the sample points, but we do not know the analytical FES: thus we can test the $\widehat{\delta F}$s but not the gradient estimators. These tests demonstrate that the estimator $\widehat{\delta F}_{ij}$ is more robust than the estimator $\hat{\mathbf{g}}_i$; we explain this fact by considering that by taking the semisum of two gradient estimates as in equation (6.15), errors compensate at second order, bringing the leading-order corrections in the estimator $\widehat{\delta F}_{ij}$ to fourth order.

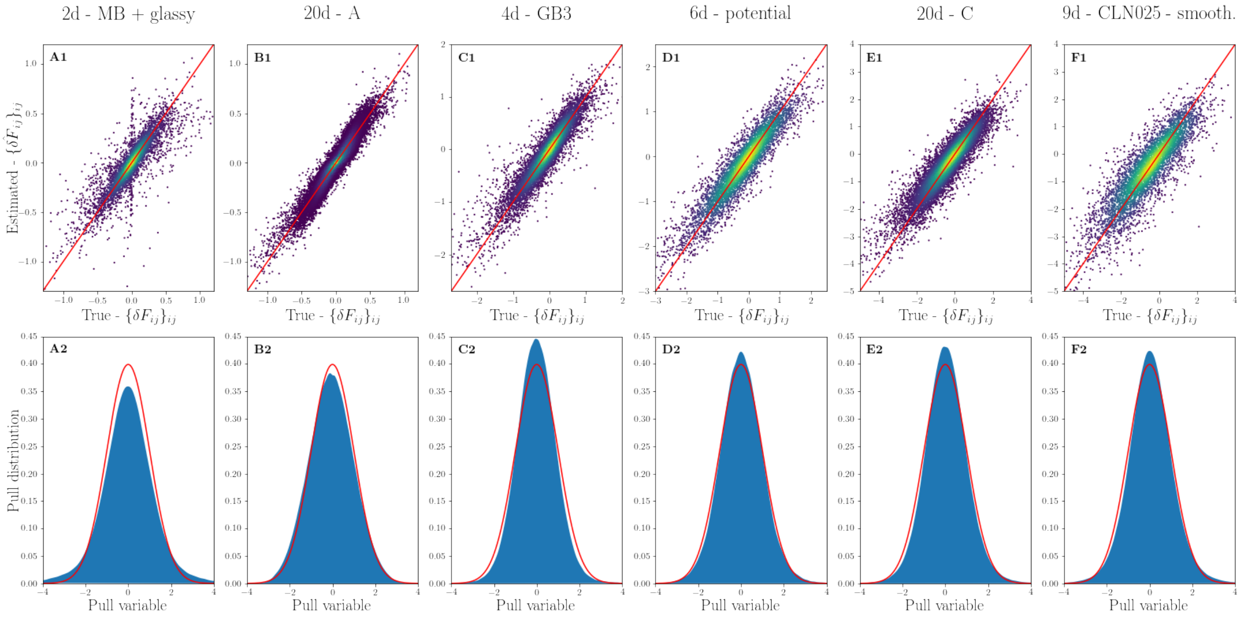


Figure 6.6: $\hat{\delta F}$ estimator performance tested on various systems. The six systems considered, one for each column, are indicated in the column header (cfr. Appendix A). Top row: correlation plots of estimated $\{\hat{\delta F}_{ij}\}_{ij}$ against true values $\{\delta F_{ij}\}_{ij}$. In red the line $\hat{\delta F}_{ij} = \delta F_{ij}$. Bottom row: distribution of the pull variables $\hat{\chi}_{ij}$ in equation (6.21). In red the standard normal distribution $\mathcal{N}(0, 1)$.

Finally, Figure 6.4 illustrates the effect of estimating the Pearson correlation coefficient $\hat{p}_{ij}$ in different manners. In the top row we see the behaviour when $\hat{p}_{ij}$ is set to zero for all points: the error is visibly underestimated when i and j are very close (panels on the right), generating pull distributions with variance greater than one; when considering couples of further points (going from right to left) the effect of correlations between gradient estimates is less pronounced, since they do not have relevant parts of their neighbourhoods in common, and the pulls resemble more the standard normal distribution. In the bottom row we see how, by using our geometrical proxy in equation (6.19), the bias at low neighbour ranks is corrected. The effect at all ranks is more visible for the lower-dimensional system, in yellow, w.r.t. the higher-dimensional in blue; this is in agreement with our understanding that correlations between sample points are more important in low dimension. Actually, by looking at the rightmost panels, we see that the pull in yellow goes from being too spread to being slightly too narrow. This might suggest that $\hat{p}_{ij}$ slightly overestimates the real Pearson correlation coefficient. However, panel A1 shows that the pull χ_{ij} when point j has a neighbourhood rank close to $\hat{k}_i$ w.r.t. point i is slightly more concentrated than predicted, even when $\hat{p}_{ij}$ is zero; all of this despite the fact that, as seen in Figure 6.3b, the absolute error $\Delta\delta F_{ij} := \delta F_{ij} - \hat{\delta F}_{ij}$ is higher at higher ranks. One possible explanation, then, is that variances propagated from single gradient estimates $\hat{\varepsilon}^{i}_{ij}{}^{2}$ are slightly overestimated, thereby narrowing the pull.

Possibly, improvements could be introduced by estimating more rigorously the variance-covariance matrix as proposed in Appendices D.2.2.2 and D.3.4. These are quite recent theoretical results and still need to be implemented and tested. However it could well be that all our attempts are limited by noise due to the small statistics. In fact, the tested distributions of the observed chi-squared $\chi^2_{\hat{\mathbf{g}}_i}$ seem to suggest that the gradients' covariance matrices estimates rapidly become unreliable with increasing dimensionality.

In conclusion, Figures 6.5 and 6.6 display an excellent overall performance in a wide range of systems, embedding dimensionalities and IDs for both the estimators of the neighbours free energy difference $\hat{\delta F}_{ij}$ and of its error $\hat{\varepsilon}_{ij}$, which includes our empirical correction $\hat{p}_{ij}$. We consider these estimators satisfying enough to build upon them.

6.3 Binless multidimensional thermodynamic integration: BMTI

So far we have illustrated how we estimate the free energy difference between neighbouring points of the data sample and its statistical error. These are, in turn, based on a reliable procedure to define the extent of points' neighbourhoods and on an accurate gradient estimator with its uncertainty. By choosing a path connecting point $\mathbf{x}_i$ to point $\mathbf{x}_f$ via couples of neighbouring points $\{(\mathbf{x}_i, \mathbf{x}_{i+1}), (\mathbf{x}_{i+1}, \mathbf{x}_{i+2}), \ldots, (\mathbf{x}_{f-2}, \mathbf{x}_{f-1}), (\mathbf{x}_{f-1}, \mathbf{x}_f)\}$ one should in principle be able to reconstruct the free energy difference $\Delta F_{if} := F_f - F_i$ by summing all the small estimated $\hat{\delta F}$s. This idea is the same one underpinning thermodynamic integration[107, 126]: estimating the gradient (or derivative in the univariate case) of the free energy and integrating it along a path to retrieve free energy differences.

Fixing a set of neighbourhood sizes $\{\hat{k}_i\}_i$ for each point i, as discussed in sections 2.2.3 and 3.1.3, defines a sparse directed connection graph, the NG, on which a couple of points is typically connected by multiple paths. This fact calls for a procedure that estimates free energy differences among points considering contributions from all the possible paths. If one were able to do so, as long as the NG is connected, all the relative free energies among points would be coherent, getting rid of the spurious correlations due to redundant counting of points discussed in section 3.2.5.3. What is most commonly done in literature when dealing with TI[40] is either integrating the so-called mean force along a single curve[53] or solving the problem on a grid where estimates of the free energy gradient have been collected at each node[54, 94]. In the first family of methods estimated free energies are by construction one-dimensional. In a high-dimensional space, the free energy difference between pair of data points can turn out to be path-dependent; in other words, closed paths in configuration space typically start and end at different free energy values. In the second family of methods, instead, the requirement that a dense grid of estimates is populated is quite demanding making the approaches computationally heavy in dimensionality greater than 2 or 3.

In analogy with PAk, but in a completely different setting, we propose to obtain the free energy estimates via a log-likelihood maximisation. However, in this case all the free energy at sampled points are computed simultaneously.

6.3.1 Distribution of the δFs

First of all, let us now, for a lighter notation, label any couple of neighbouring points in the sample by a single index $a := (i, j)$, so that $\{a, b, \dots\}$ represent the edges of the NG $\{(i, j), (l, m), \dots\}$ (cfr. Appendix D.3.4). How many of these couples are there? If one wants to consider directed edges, then they are as many as the non-zero entries of the sparse connectivity matrix of the NG, that is:

$$N_{\text{spar}} = \sum_{i=1}^{N} (\hat{k}_i - 1) \approx N(< \hat{k} > -1) \ . \tag{6.22}$$

Then, let us us recall the considerations of section 6.2.1 and in particular equation (6.20). We have proven in section 6.2.2 that χ_{ij} is well estimated by $\hat{\chi}_{ij}$ in equation (6.21). This means that the random variable $\hat{\delta F}_a$ has mean value $\mu_a := \langle \hat{\delta F}_a \rangle = \delta F_{ij} = F_j - F_i$ and can be seen as marginal variable of a multivariate normal distribution: $\hat{\delta F}_a \sim \mathcal{N}(\mu_a \ , \ \varepsilon_a^2)$. Therefore, the vector containing the $\hat{\delta F}_a$'s for all couples of neighbours $\hat{\boldsymbol{\delta F}} = \{\hat{\delta F}_a\}_a$ is distributed as:

$$\hat{\boldsymbol{\delta F}} \sim \mathcal{N}(\boldsymbol{\delta F} \, , \, \mathbf{C}) \ \propto \ \exp\left[-\frac{1}{2} \sum_{a,b} (\delta F_a - \hat{\delta F}_a)^{\mathrm{T}} \ \mathbf{C}^{-1}_{a,b} \ (\delta F_b - \hat{\delta F}_b) \right] \ . \tag{6.23}$$

The covariance matrix $\mathbf{C}$ has size $N_{\text{spar}} \times N_{\text{spar}}$, with N_{spar} defined in equation (6.22). The diagonal of matrix $\mathbf{C}$ is $C_{aa} = \varepsilon_a^2$ for all N_{spar} couples labelled by a. However $\mathbf{C}$ is generally not diagonal due to correlations between couples $\delta F_a, \delta F_b$ estimated on at-least-partially overlapping regions of configurational space. Again, we refer the reader to an appendix, namely Appendix D.3.4, for a discussion on the matter and a proposal on how to address it.

Looking at equation (6.23) and bearing in mind the definition of the δFs in equation (6.12), we can see that the argument of the exponential, in square brackets, can be recast into a quadratic form for the free energies. By calling $\mathbf{F}$ the vector of all free energies at sample points $\{F\}_i$, this quadratic form reads:

$$\mathbf{F}^{\mathrm{T}} \cdot \mathbf{A} \cdot \mathbf{F} \ + \ \mathbf{b}^{\mathrm{T}} \cdot \mathbf{F} \ + \ c \tag{6.24}$$

where the $N \times N$ matrix $\mathbf{A}$, the N-vector $\mathbf{b}$ and the scalar c depend on the estimated free energy differences $\hat{\boldsymbol{\delta F}}$ and on their covariance matrix $\mathbf{C}$. Therefore, we can take the logarithm of this N_{spar}-variate Gaussian and interpret it as a log-likelihood for the error-affected observations $\{\delta \hat{F}_{ij}\}_{ij}$s as a function of the parameters $\{F_i\}_i$:

$$\mathcal{L}(\mathbf{F} \mid \hat{\boldsymbol{\delta F}}\,, \mathbf{C}) \;\propto\; \log \mathcal{N}(\boldsymbol{\delta F}\,, \mathbf{C})\;. \tag{6.25}$$

This theoretical framework allows to obtain estimates for the free energies maximising the log-likelihood over the parameters and also provides a well-defined procedure to compute their error.

6.3.2 Uncorrelated free energy differences approximation

As a first approach to the solution of the model in equation (6.25) we can make the simplifying assumption that the δF's are uncorrelated, namely that $\mathbf{C}$ is approximated by a matrix D retaining only its diagonal part: $D_{ab} = \delta_{ab}\, C_{aa} = \varepsilon_a^2$. With this assumption, the log-likelihood in equation (6.25) becomes:

$$\log \mathcal{N}(\boldsymbol{\delta F}\,, \mathbf{D}) \;\propto\; \mathcal{L}(\mathbf{F} \mid \hat{\boldsymbol{\delta F}}\,, \mathbf{D}) \;:=\; -\sum_{i=1}^{N} \sum_{j\in\Omega_i} \frac{(F_j - F_i - \hat{\delta F}_{ij})^2}{2\varepsilon_{ij}^2}\;. \tag{6.26}$$

This formulation makes it clearer that we are dealing with a weighted least squares model, while equation (6.25) corresponds to a generalised least squares model[181]. The maximisation of $\mathcal{L}(\mathbf{F} \mid \hat{\boldsymbol{\delta F}}\,, \mathbf{D})$ with respect to the parameters $\{F_i\}_i$ to obtain the optimal free energy estimators $\{\hat{F}_i\}_i$ provided by this model can be recast (see Appendix D.3.1) into the linear system:

$$\sum_{j=1}^{N} A_{ij}\, \hat{F}_j \;=\; \Delta_i \tag{6.27}$$

in which the matrix $\mathbf{A}$ and the vector of $\boldsymbol{\Delta}$ will be now defined. Each off-diagonal element of the matrix A_{ij} has a contribution $-\varepsilon_{ij}^{-2}$ if $i \in \Omega_j$ and another identical one if $j \in \Omega_i$; instead, the diagonal terms are the negative of the sum of all the off-diagonal on that line $A_{ii} = \sum_{j\neq i} A_{ij}$; we notice that $\mathbf{A}$ is symmetric even if the connectivity matrix (derived from the NG) is not. As for the constant vector, each Δ_i is defined as $\Delta_i := -\left(\sum_{j|i\in\Omega_j} + \sum_{j\in\Omega_i}\right) \hat{\delta F}_{ji}/\varepsilon_{ji}^2$. Notice that, since the free energies enter equation (6.26) only as terms of differences, their value is determined except for an arbitrary additive constant. This implies that the linear system (6.27) is underdetermined and is reflected in the fact that matrix $\mathbf{A}$ is singular. However, bearing this in mind, the linear system can easily be solved using any standard linear algebra library (e.g. implementing the conjugate gradient method[97]), returning our free energy estimates $\{\hat{F}_i\}_i$.

Equation (6.27) defines a procedure that allows computing all the free energies simultaneously, by weighting in some uncertainty-dependent manner all the various paths' contributions. Notice that in one dimension and with two neighbours (one on the left and one on the right) this scheme is exactly equivalent to TI in its classical formulation[107]. Unlike standard TI techniques, this approach does not require the introduction of CVs nor the definition of a regular grid over which data about the free energy gradient are collected. This makes the approach suitable for free energy estimates in for high-dimension. We name this approach Binless Multidimensional Thermodynamic Integration (BMTI). BMTI has gives excellent results even in its simplified formulation presented in this section, as we will show in section 6.3.2.2.

6.3.2.1 The estimation of error for BMTI

Given a log-likelihood like the one in equation (6.25), the covariance matrix of the maximum-likelihood estimators that can be derived, once again, by taking the equal sign in the Cramér–Rao Bound inequality:

$$\mathbf{cov}[\hat{\mathbf{F}}]_{\mathrm{ij}} := \left\langle -\frac{\partial^2}{\partial \mathrm{F_i} \partial \mathrm{F_j}} \mathcal{L}(\mathbf{F} \mid \hat{\boldsymbol{\delta}\mathrm{F}}\,, \mathbf{C}) \right\rangle^{-1} . \tag{6.28}$$

The diagonal elements of this covariance matrix represent our uncertainty estimates on the MLEs $\{\hat{F}_i\}_i$. Importantly, in the case we are considering, in which the sparse covariance matrix $\mathbf{C}$ of the $\delta\hat{F}$s is replaced by its diagonal $\mathbf{D}$, the inverse of $\mathbf{cov}[\hat{\mathbf{F}}]$ corresponds exactly to the matrix $\mathbf{A}$ in equation (6.27), therefore:

$$\varepsilon_i^2 = \mathrm{var}[\hat{F}_i] = (\mathbf{A}^{-1})_{ii}. \tag{6.29}$$

Thus, estimating the error on the estimates amounts to inverting matrix $\mathbf{A}$. Since this matrix is singular, expression $\mathbf{A}^{-1}$ must be interpreted as the pseudoinverse[149] of $\mathbf{A}$.

Unfortunately, the error in equation (6.29) has proven to underestimate the real statistical error (see discussion in section 6.3.2.2 and Figure 6.9), regardless of the fact that the solution of equation (6.27) returns very accurate predictions for the free energies. We suspect this has to do with the redundancy of expression (6.26): despite depending on N_{spar} parameters $\{\hat{\delta F}_a\}_a$ and N_{spar} parameters $\{\varepsilon_a\}_a$, these are far from independent measures, since they are all computed starting from $N \ll N_{\mathrm{spar}}$ datapoints. Therefore, we expect the effective DOFs of the N_{spar}-terms sum constituting $\mathcal{L}(\mathbf{F} \mid \hat{\boldsymbol{\delta F}}\,, \mathbf{D})$ to be at most $N-1$; they are exactly $N-1$ in the case $\{\mathbf{x}_i\}_i$ are actually IID random variables.

One possible way to address this problem is dividing the BMTI log-likelihood by some factor accounting for the redundancy contained in its formulation. This idea was backed by looking at a simplified version of the problem in one dimension. In Appendix D.3.2 we briefly discuss this case. Generalising this result to our case was for us unfeasible without, once again, confronting the challenging problem of the estimation and invesion of the correlation matrix of the δFs, discussed in Appendix D.3.3. We attempted an empirical generalisation based on the one-dimensional results. Our intuition suggests that this divisor should be chosen of the order of the average neighbourhood size, $\mathcal{O}(\langle\hat{k}_i\rangle_i)$; this would increase the estimated variances of the $\hat{F}$s, in equation (6.29) of approximately the same order. Therefore we choose it, for each term (i,j) of the sum defining BMTI log-likelihood in equation (6.26), to be the geometric average of the selected neighbourhood sizes for points i and j; thus, the redundancy-corrected BMTI log-likelihood becomes:

$$\mathcal{L}_{\mathrm{r}}^{\mathrm{BMTI}}(\mathbf{F} \mid \hat{\boldsymbol{\delta F}}\,, \mathbf{D}) \;:=\; -\sum_{i=1}^{N}\sum_{j\in\Omega_i}(\hat{k}_i\,\hat{k}_j)^{-\frac{1}{2}}\,\frac{(F_j - F_i - \hat{\delta F}_{ij})^2}{2\varepsilon_{ij}^2}\;. \tag{6.30}$$

We discuss the efficacy of this empirical correction on the error estimates of BMTI in in paragraph 6.3.2.2.3 of the next section. Remarkably, the redundancy-corrected model in equation (6.30) produces the same results for the estimated $\hat{F}$s as the one in equation (6.26): on all tested datasets the free energies estimated in the two manners differ at most at the third significant digit; we take this as a clue that there is indeed some redundancy in the BMTI model and that the direction is the right one.

However, we anticipate that, while this empirical correction is exact for the simplified one-dimensional case considered in Appendix D.3.2, the quality of the error estimates immediately deteriorates in two or more dimensions, improving error estimates only partly, but not adequately.

6.3.2.2 Performance of BMTI

6.3.2.2.1 Accuracy of BMTI We assess the accuracy of the BMTI estimator by considering three statistical tests discussed in section 3.2.3.1 of Chapter 2: firstly, the correlation plots showing for every point i the estimated free energies $\hat{F}_i$ plotted against the ground truth values F_i; the scatterplot should lie on average on the line $\hat{F}_i = F_i$. Secondly, the absolute error of the estimator $\epsilon_i^{L_1} = |F_i - \hat{F}_i|$ introduced in equation (3.14) plotted as a function of the true free energy F_i in order to examine the behaviour of the estimator in all ranges of data density. Third, the mean absolute error from equation (3.15), expressed as: $\mathcal{E}^{L_1} = \frac{1}{N}\sum_{i=1}^{N}|F_i - \hat{F}_i|$, which quantifies the performance of $\hat{F}$ globally on a dataset. We compare the performance of BMTI to that of other nonparametric estimators both graphically and quantitatively. Such estimators are PAk, *BMTI* and also: standard kNN but with k selected as $k^{\mathrm{sel}} := \langle\hat{k}_i\rangle_i$ i.e. as the average optimal k found by PAk for the dataset; a fixed-bandwidth Gaussian KDE with h^{sel} selected according to Scott's rule of thumb[170, 194]; in the evaluation of $\mathcal{E}^{L_1}$, in Table 6.1, we consider also $\hat{k}$NN introduced in section 3.1.3 of Chapter 3, just below equation (3.7) and our adaptive version of the Gaussian KDE, PAkde, introduced in section 3.1.3 of Chapter 3 and explained in Appendix B.

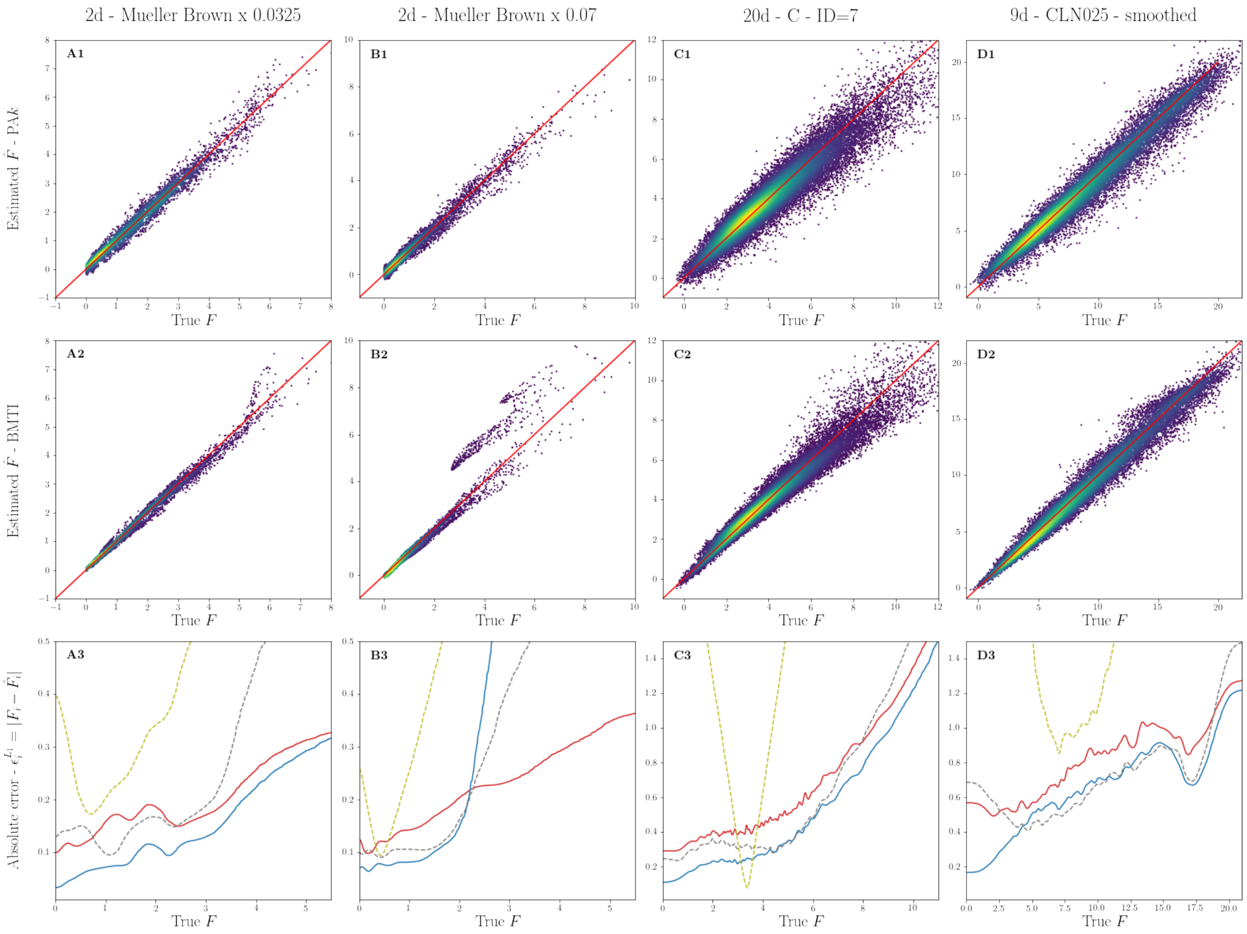


Figure 6.7: BMTI estimator performance tested on various systems and compared to other methods. The four systems, one for each column, are indicated in the column header; they are all described in Appendix A; their embedding dimensionalities go from 2 to 20, while their intrinsic dimensionalies go from 2 to 9. For the Mueller-Brown potentials in columns A and B the sample size is 5000 and the ID is $d = 2$; for the system in column C, a trajectory of 30000 points in $d = 7$ dimensions embedded in $D = 20$ is considered; in column D the considered sample has 20000 points and ID $d = 9$. Top and middle rows: correlation plots of estimated free energies $\hat{F}_i$ against true values F_i. Top: PAk estimator. Middle: BMTI estimator. In red the lines $\hat{F}_i = F_i$. Bottom row: absolute error $\epsilon_i^{L_1}$ as a function of the free energy F_i; a Gaussian filter has been applied to the data for readability. Various estimators are compared: in red PAk; in blue BMTI; in dashed grey kNN with k selected as $k^{\text{sel}} := \langle \hat{k}_i \rangle_i$; in dashed light green Gaussian KDE with smoothing parameter h^{sel} selected according to Scott's rule of thumb[170, 194].

Let us look at Figure 6.7. In the first two columns we consider two samples of 5000 points of the classical bidimensional Mueller-Brown potential, described in Appendix A.4; the one in column A is extracted at temperature which is double w.r.t. the sample in column B; this is equivalent to sampling at $\beta = k_B T^{-1} = 0.035$ and $\beta = 0.07$ the Mueller Brown potential in equation (A.4) for the systems in column A and in column B respectively. Thus, the free energy barrier to escape the global minimum of the potential is twice as big in the second column as it is in the first one.

In the first system we can see that BMTI correlation plot, in panel A2, is much sharper than that of PAk, in panel A1; this is especially true at low free energy values, while BMTI estimates become a bit noisier very high in the free energy range. In panel A3 we can see that BMTI, in

System	PAk	BMTI	kNN		GKDE h^{Scott}		$\hat{k}$NN	PAkde
	$\mathcal{E}^{L_1}$	$\mathcal{E}^{L_1}$	$\mathcal{E}^{L_1}$	k^{sel}	$\mathcal{E}^{L_1}$	h^{sel}	$\mathcal{E}^{L_1}$	$\mathcal{E}^{L_1}$
2d Gaussian	0.14	0.08	0.15	194	0.06	0.22	0.10	0.13
20d-A ($d = 2$)	0.12	0.07	0.12	313	0.79	0.68	0.08	2.71
2d-MB x 0.035	0.16	0.10	0.19	123	0.41	0.24	0.13	0.13
2d-MB x 0.07	0.11	0.25	0.20	155	0.41	0.24	0.13	0.14
6d potential	0.50	0.26	0.36	26	0.67	0.40	0.34	0.37
20d-C ($d = 7$)	0.52	0.36	0.46	22	1.69	0.65	0.43	1.67
9d CLN025 sm.	0.76	0.61	0.63	13	2.39	0.47	0.65	0.90

Table 6.1: Absolute error of PAk, BMTI and other nonparametric estimators.

blue, outperforms PAk, in red, in terms of absolute error across the whole range of free energies; the difference between the two estimators' performance is quite evident and narrows for low-density points. Standard kNN, in dashed grey, outperforms PAk only in a small range of intermediate values and gives very biased estimates at high F values. The Gaussian KDE with fixed smoothing parameter has an L_1 error comparable with the other estimators only in a small range of F values, before and after which it rapidly diverges.

In the second Mueller-Brown system, column B, instead, the high free energy barrier causes the regions of the two saddle points of the Mueller-Brown potential to be visited few times, if any at all, as we can see in Figure A.1. Consequently, the NG of the second system is disconnected or at least weakly connected and the solution of equation 6.27 produces three separate blocks, corresponding to the three basins. It is like we had performed thermodynamic integration on three separate systems, starting from three different arbitrary offset free energy values. The free energies within these clusters are all coherent among themselves, but the three offsets are not the same, producing the disconnected correlation that we see in panel B2. The correlation plot in panel B1 shows instead that PAk does not seem to be affected by this disconnection. By looking at the absolute error as a function of F we see that BMTI again outperform PAk in denser regions, corresponding to the basin where the global minimum is located, but then suddenly diverges as F increases. An analogous behaviour is observed for kNN. The Gaussian KDE again works well only in a limited range of Fs, in which it even beats PAk, but it is not reliable outside it.

In column C of Figure 6.7 we consider a sample of 30000 points extracted from the 20-dimensional system described in Appendix A.8.3 which has intrinsic dimension $d = 7$; in column D we consider a sample of 20000 points of the KDE-smoothed 9-dimensional potential described in Appendix A.7.1.1, which has intrinsic dimension $d = 9$. BMTI performs better than PAk on both systems as we can see from the sharper correlation plots in panels C2 and D2 compared to PAk's in panels C1 and D1 and from panels C3 and D3, where the error of BMTI is constantly under that of PAk.

Standard kNN in both cases has a range in which its error is lower than PAk's and is also for a while very similar to BMTI's; however, it has a tendency to be biased outside the range in which the selected k^{sel} is optimal. As for the Gaussian KDE, its behaviour is similar to that observed in panels A3 and B3.

Table 6.1 reports the mean absolute errors for various estimators in order to assess their global performance. The systems considered are the same as in Figure 6.7, with the addition of two more: a sample of 10000 points extracted from the bivariate Gaussian distribution in column A of Figure A.2 of Appendix A; 10000 points extracted sampled from the 6-dimensional potential considered many times so far (see Appendix A.6). Overall, we see that BMTI is always the best performing with two exceptions: first, the Mueller-Brown potential sampled at $\beta = 0.07$ (column B of Figure 6.7): in this case estimates are accurate as long as one is interested in the global minimum of the potential, but they become unreliable at higher free energy values; second, the 2-dimensional Gaussian, which has a convex harmonic potential with only one minimum, where the fixed-bandwidth Gaussian KDE curiously outperforms all other methods and also slightly BMTI, which displays the second best score; this is the only exploit of this method, which performs poorly on all other systems. As for the adaptive Gaussian KDE, PAkde, it does a good job up to dimensionality $D = 6$, where it fairly competes with the other methods, but its performance becomes worse when the embedding dimensionality is pushed higher, no matter the fact that the ID might be low; in fact, the Gaussian kernel is D-variate, so the information content of one sample point having coordinates $\mathbf{x}_i$ is diluted over all D dimensions, rather than being retained within the intrinsic manifold. Finally, standard kNN performs worse than PAk in low dimensionality, but it becomes globally more accurate in high dimensions: this is probably because, k^{sel} chosen as $\langle \hat{k}_i \rangle_i$ keeps a low neighbourhood size in higher density regions, where PAk mistakenly slightly overestimates the optimal neighbourhood size, also due to the curse of dimensionality; however, we still know by looking at the bottom row in Figure 6.7 that PAk always outperforms kNN, at high F values, where it can use its point-adaptiveness to reduce the number of neighbours to few units. Interestingly, on various systems $\hat{k}$NN performs better than PAk itself, from which it takes the selected $\hat{k}$; this is probably due to the fact that PAk's intrinsic roughness produces noisy estimates even at high-density, where instead standard kernel methods converge more rapidly to the ground truth values; however, despite not hereby illustrated, we have seen that at high free energy values PAk always surpasses $\hat{k}$NN, a sign that PAk likelihood correction is fundamental in rare data conditions to capture some curvature effects on regions where the the free energy cannot be approximated as constant.

Last but not least, it is important to stress that the only methods that have proven reliable across the whole set of test systems and samples, giving consistently fair estimates, are PAk, $\hat{k}$NN

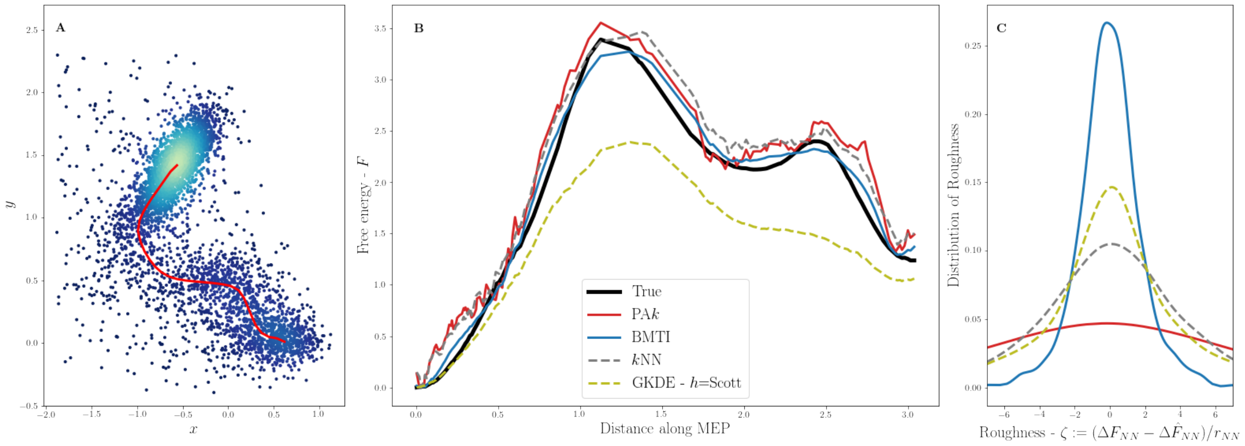


Figure 6.8: Smoothness of BMTI estimator compared to other methods. We consider the Mueller-Brown potential in equation (A.4) sampled 5000 times at an inverse thermodynamic temperature $\beta = 0.035$. **A** Scatter plot of the sample coloured according to the density. The red curve represents the NN-interpolated MEP discussed in Appendix A.4. **B** Vaious free energies along the NN-interpolated MEP. In black the analytic freee energy; in red PAk; in blue BMTI; in dashed grey kNN with k selected as $k^{\text{sel}} := \langle \hat{k}_i \rangle_i$; in dashed light green Gaussian KDE with smoothing parameter h^{sel} selected according to Scott's rule of thumb[170, 194]. **C** Distribution of roughness measured on dataset points for various estimators; the colour code is the same as in panel B.

and standard kNN. Despite BMTI impressive performance, there was one single case in which it did not excel; but in that case it stumbled heavily. We are. currently developing a strategy to improve the performance of BMTI in the case of weakly connected or disconnected NGs or at least to discriminate the conditions in which an alternative to BMTI should be used.

6.3.2.2.2 Smoothness of BMTI One of the main concerns that pushed us to develop BMTI as an improvement to the already well-performing PAk was addressing the problem of PAk's intrinsic roughness and, in general, the amplification of local sample fluctuations carried out by kNN-based methods. Indeed, the free energy is a smooth function, namely that free energy values at neighbouring points should be close to each other. In kNN methods every estimate F_i is derived independently from all estimates on other points $\{F_j\}_{j \neq i}$. In fact, the log-likelihoods that define kNN, $\hat{k}$NN and PAk do not depend on the value of F at other points. Thus, the global likelihood for a vector of free energies $\mathbf{F}$ will be the product of N likeihoods for the points i (see sections 3.1.2 and 3.2.2 of Chapter 3).

We now want to assess how good BMTI is in obtaining smooth free energy estimates. A multidimensional FES can be rugged and bumpy, but we are only interested in local free energy fluctuations artificially introduced by our estimators, rather than those which are simply features of the system considered. Based on this concept, we introduce an observable to quantify locally the deviation of free energy estimator gradient with respect to the behaviour of the true landscape. We call it *roughness* We want to have a measure of the spatial coherence of free energy estimates, comparing it to the spatial rate at which the true free energy varies. Thus, we define the local roughness as:

$$\zeta_i := \frac{\Delta F_{i,\mathrm{NN}_i} - \Delta \hat{F}_{i,\mathrm{NN}_i}}{r_{i,\mathrm{NN}_i}} \; , \tag{6.31}$$

where $\Delta F_{i,\mathrm{NN}_i}$ is the ground truth free energy difference between point i and its nearest neighbour, $\Delta \hat{F}_{i,\mathrm{NN}_i}$ is the estimated free energy difference between point i and its nearest neighbour and r_{i,NN_i} is the Cartesian distance between the two points. Basically, ζ_i is the difference between the Newton quotients between a point i and its nearest neighbour in the estimated FES and in the ground truth FES. The roughness quantifies the presence of local spikes in the estimated free energy surface which are not present in the real one. As we know from calculus, the limit for $r_{i,\mathrm{NN}_i} \to 0$ of the Newton quotient is the gradient projection along the direction of $\mathbf{r}_{i,\mathrm{NN}_i}$, thus:

$$\lim_{r_{i,\mathrm{NN}_i} \to 0} \zeta_i \; = \; \frac{\mathbf{r}_{i,\mathrm{NN}_i}}{r_{i,\mathrm{NN}_i}} \cdot \Big(\nabla_{\mathbf{x}} F(\mathbf{x}_i) - \nabla_{\mathbf{x}} \hat{F}(\mathbf{x}_i) \Big) \; . \tag{6.32}$$

In the case of an unbiased consistent estimator the gradient of the estimated free energy should of course converge to that of the true one; hence, in the limit of infinite statistics, where the limit in equation (6.32) is automatically taken, ζ_i should go to zero.

In probabilistic terms, the distribution of the RV defined in equation (6.31) should concentrate around zero for an unbiased estimator, with a variance that goes to zero increasing the statistics. Fixing the statistic and a test system, the spread of the distribution of ζ is higher if the derivatives of free energy estimates are noisier (namely, if the free energy estimates are rougher).

In order to visualise how BMTI roughness compares to that of other estimators we can look at Figure 6.8. In panel A, represented as a red curve, we observe the NN-interpolated Minimum Energy Path (MEP) between the two main minima of the potential. The way this path is constructed is reported in Appendix A.4; in brief, is the path formed only of sample points which is most similar to the real MEP connecting the two main minima of the potential.

In panel B we can look at the estimator performance along the one-dimensional CV given by the length of the NN-interpolated path from the global minimum to a point. We call this CV simply the *distance along the MEP*; it is zero in the global minimum and it has its maximum value when the path reaches the second-deepest minimum of the Mueller-Brown potential. In black we observe the true free energy values, in blue BMTI estimates, in red PAk estimates. While the general trend along the path and the free energy differences between couples of free energy extrema (local maxima or minima) are well captured by both estimators, we clearly see the spikyness of PAk as opposed to BMTI smoothness; this effect is more evident in the free energy minima, where the sample density is higher a pathological behaviour which does not preclude PAk's overall efficiency, but prevents

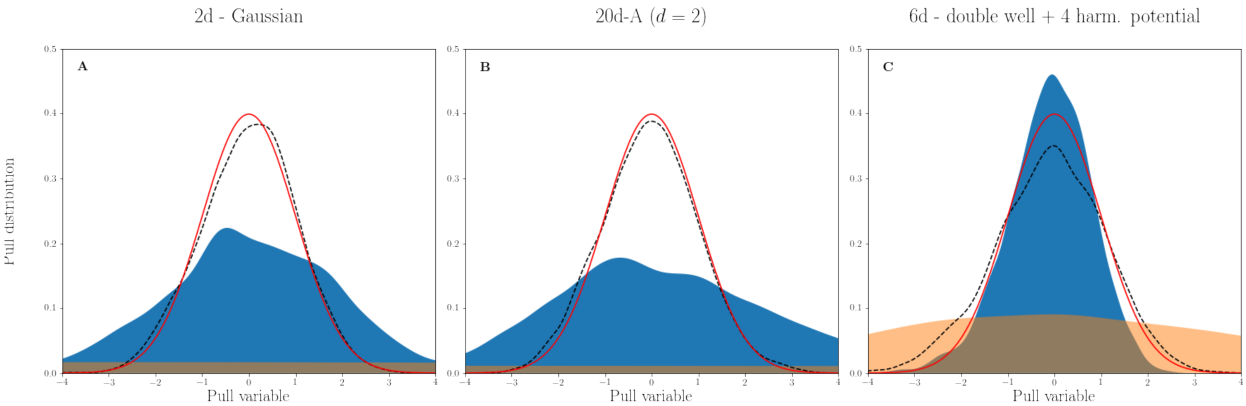


Figure 6.9: Testing the effect of the empirical redundancy factor in BMTI likelihood for various systems. Distribution of the observed pull $\hat{\chi}_i$ in equation (3.16) for vaious estimators and various systems. Each panel corresponds to a different system, indicated in the panel titles (cfr. Appendix A). In orange the pull distribution where pull variables are computed using the original BMTI error in equation (6.29); in blue the redundancy-corrected error is adopted; the dashed black line corresponds to the pull distribution of PAk estimates; in red the standard normal distribution $\mathcal{N}(0, 1)$.

it from always crushing the competition of other nonparametric methods, as seen in Table 6.1. Standard kNN performs quite similarly to the previous two approaches, but with a tendency to always overestimate the free energy, which is coherent with the higher mean absolute error in Table 6.1. The Gaussian KDE, instead, rapidly detaches from the ground truth curve and systematically underestimates the free energy.

Finally, in panel C we can inspect the distribution of the observed roughness for the various estimators. The image proves that BMTI is the smoothest of all considered estimators, followed by the Gaussian KDE, then kNN and finally by PAk, confirming in a more quantitative way the insight provided by panel B. Notice that, if one were surprised by the bad performance of the intuitively smooth Gaussian KDE estimator, it should be considered that this is partly an effect of the Gaussian KDE biasedness, well represented in panel B. In fact, while smoothness is not a sufficient condition for unbiasedness, unbiasedness is a necessary condition for scoring low in roughness as per our definition in equation 6.31. Nonetheless, even the adaptive Gaussian KDE, not reported in panel C for readability, has a broader roughness distribution than BMTI. This can be understood by thinking that, even by adding many smooth bumps one over the other, there is no guarantee that the obtained result will not be rugged, due to the finiteness of the statistic. BMTI is the only nonparametric method considered which enforces coherence among estimates at neighbouring points, so it is rightfully the smoothest among them.

6.3.2.2.3 Performance of BMTI uncertainty estimator In order to assess the performance of the uncertainty estimator for BMTI given by equation (6.29) corrected or not by the redundancy factor, we look at the pull distributions for BMTI and the redundancy-corrected BMTI and compare them to the pull distribution of our benchmark method, PAk. We consider three systems: in panel A a sample of 10000 points extracted from the bivariate Gaussian distribution in column A of Figure A.2 of Appendix A; in panel B a sample of 5000 obtained from the Mueller-Brown potential in equation (A.4) sampled at $\beta = 0.035$ (the sample scatter plot is also visible in panel A of Figure 6.8); in panel C 10000 points extracted sampled from our usual 6-dimensional (see Appendix A.6).

PAk error proves fairly accounted for on all tested systems, with the black dashed line following the red solid line (the theoretical prediction) quite closely; even in the 6-dimensional case the performance is satisfying, considering that the dimensionality is quite high. Instead, the pull distribution of the uncorrected BMTI estimator, in orange, is always much wider than the standard normal, a sign that the quantity in equation 6.29 highly underestimates the correct variance. Even in its corrected version, in blue, the results are ambiguous: in the two systems of ID $d = 2$, in panels A and B, the redundancy correction still leaves BMTI error underestimated, leaving the pull quite spread, too much for systems in which the COD does not play; in the 6-dimensional system in panel C, in contrast, the correction overcompensates and the pull results even narrower than a standard normal.

The behaviour of BMTI and of its error has been tested on all systems considered in this thesis work and presented in Appendix A; in most of the cases the correction leaves the error underestimated, like in panels A and B, but in some other cases it overcompensates like in panel C. We could identify a factor to help predict whether the corrected variance would turn out underestimated or overestimated, such as could for example be the dimensionality: in the case of the 20-dimensional system of ID $d = 7$ (cfr. Appendix A.8.3) the error is overcompensated like in panel C, but in the 9-dimensional case (cfr. Appendix A.7.1.1) it is underestimated. This is confusing and can only lead us to the conclusion that we still do not have a viable option for accurately estimating the uncertainty on BMTI estimates. On a positive note, however, the redundancy correction always seems to retrieve at least the correct order of magnitude for the estimator variance, so this error estimate can be used in context in which an approximate error estimate is sufficient.

6.3.3 Discussion

The BMTI free energy estimator has been conceived with the purpose of producing estimates that are close in value for neighbouring point, like for continuous and differentiable functions. Our motivation lies in the fact that the flagship free energy estimation method in our group, PAk, displays a noisy behaviour, which is visible especially in contexts of low dimensionality and high sample density, where other methods' fluctuations typically reduce.

The core of BMTI is an approach to nonparametrically estimate the free energy gradient and produce accurate estimates $\hat{\delta F}$ of the free energy differences δF between neighbouring points. The $\hat{\delta F}$s are considered as marginal random variables of a multivariate normal distribution whose diagonal covariance matrix has the estimated variances of the $\hat{\delta F}$s as entries. We can interpret such distribution as a likelihood for the error-affected observations $\hat{\boldsymbol{\delta F}}$ as a function of the free energies $\mathbf{F}$ seen as parameters. The maximisation of this log-likelihood model produce the BMTI estimates $\hat{\mathbf{F}}$, which have proven to perform very well in several contexts. Despite being a maximum-likelihood estimator, for the same motivations as in the case of PAk, we regard BMTI as a nonparametric estimator, since it is only based on very local estimates and does not make any assumption on the FES functional form.

It is worth mentioning that, as far as we are aware, no method other than ours in literature is able to provide accurate nonparametric free energy gradient estimates in high-dimensionality settings like the ones we have considered. The only approach to our knowledge that seems to efficiently compute the free enrgy (actually, log-density) gradient up to dimension $D = 7$ is a parametric method presented in reference [166]. Our approach, hovever, has the advantages of a great computational simplicity (it only involves computing small sample averages) and the fact that it can be pushed to very high embedding dimensionalities D as long as the order of the intrinsic dimensionality remains limited. Moreover, we believe that the efficacy of BMTI resides in the high accuracy of the estimators $\hat{\delta F}$, which can then be used as robust building blocks to estimate the whole FES.

In order to illustrate BMTI features we can start by going back to the meaning of the acronym, namely Binless Multidimensional Thermodynamic Integration. In particular, let us first focus on the TI part. Thermodynamic Integration[107] is a term which indicates a wealth of techniques to perform free energy calculations. What they have in common is that they proceed by integrating thermodynamic quantities to retrieve differences in thermodynamic potentials[88, 125, 126]. In the cases interesting for us, the integrated quantity is the gradient of the (canonical) free energy. We do not aim to discuss here TI and connected techniques extensively, but the evident connection of BMTI to TI methods imposes to at least give a brief overview on how in this class of approaches the reconstruction of the FES is carried out.

Typically[40], like in the original paper by Kirkwood[107] and in many others, TI is carried out by integrating the gradient along a one-dimensional reaction CV[16, 33, 47, 53, 72, 185], in order obtain the potential of mean force. As pointed out in reference [41], integrating the mean force in more than one dimension is regarded in literature as a very difficult task. Sometimes the FES is reconstructed by sampling the mean force on a dense grid[94]. However, this approach is computationally demanding and thus allows for the FES reconstruction only up to $d = 2, 3$[47, 56, 124]. Within this range of spatial dimensions, it is possible to reconstruct the free energy gradient in a more sophisticated way by solving the Poisson equation [7, 93, 116]. This technique is however quite delicate, since it involves the intrinsically noisy numerical estimation of the free energy Laplacian on a shifted mesh w.r.t. the one where samples of the free energy gradients are cumulated. This approach requires to deal only with conservative fields, which should integrate to zero on loop integrals. In other cases, with the help of enhanced sampling techniques[3, 54, 121, 163], the dimensionality of the CV space explored is much higher than one, even up to 10 dimensions[36]. In these cases one faces the choice to either compute free energy differences between important free energy landmarks integrating the mean force along a one-dimensional path or to reconstruct the whole FES.

However, to the best of our knowledge, the whole FES reconstruction is done in literature only up to a CV space of four dimensions, in the case of the widely adopted[2, 34] variational reconstruction method introduced in reference [122]. This approach adopts radial-basis functions for the representation of the free energy. The points at which the radial basis functions should be centred must however be carefully chosen, involving a great deal of technicalities. The number of chosen centres is always kept below few hundreds, setting an upper limit to the estimator level of detail.

Instead, BMTI has proven accurate and reliable on all tested systems with embedding dimension up to $D = 20$, hence the “M” for Multidimensional in the acronym. Unlike many FES reconstruction strategies[7, 47, 56, 93, 94, 116] it does not necessarily require a dense sampling of the configuration space and performs well even with small samples of few thousand points, as proven by the fact that all the samples presented in this thesis work have between 5000 and 30000 points. The points used as inputs for the estimator do not need to be constrained on a regular grid (hence the “B” of binless). Moreover, BMTI retains all the adaptive structure of PAk described in Chapter 3: the restriction to the low-dimensional intrinsic data manifold and the point-adaptive neighbourhood selection (on top of the intrinsic adaptiveness typical of kNN-based methods), which defines the directed neighbours graph (see sections 3.1.1 and 3.1.3). All these features, the “binlessness”, the restriction to the intrinsic manifold and the double-adaptivity, expose BMTI much less to the curse of dimensionality. In order to reconstruct the FES in a TI-like fashion, BMTI does not require to select a number of interesting points for which the relative free energy difference is computed: it rather considers all the possible paths connecting points in the NG and computes all the relative free energy differences along them simultaneously in form of the linear system (6.27). As already stressed, this crucially relies on the quality of the estimates produced for the $\{\delta F_{ij}\}_{ij}$, which enter BMTI as main ingredients, and of their errors $\{\varepsilon_{ij}\}_{ij}$, which establish the weights attributed to the paths in the simultaneous optimisation. In addition its ability to operate with few points, BMTI also works in the opposite regime. The points entering BMTI estimates do not even need to be selected or healed and there are no precautions to be taken, unlike in reference[122]: all points in the sample can be retained. Computationally, BMTI can handle up to $\mathcal{O}(10^5)$ points on a desktop computer. A Python package implementing this approach has been recently publicly released by our group on GitHub[79]. If one is interested in the error computation, however, numbers are limited by memory requirements, since the procedure described by equation 6.29 requires to compute the dense inverse of the sparse matrix $\mathbf{A}$, which has size $N \times N$. All these considerations lead us to believe that BMTI is a competitive option when the free energy gradient information is to be employed to reconstruct the FES, i.e. in the TI class of methods.

The only possible downside w.r.t. other TI approaches is that BMTI has been conceived to deal with equilibrium samples, i.e. it needs the sample points to be extracted by the unbiased underlying distribution. If this is not the case, the output of the gradient estimator $\hat{\mathbf{g}}$, defined in equation (6.8), on the biased sample does not represent the gradient of the ground truth FES. Nonetheless, there are no apparent inconveniences in employing BMTI likelihood even with free energy gradient data gathered in biased simulations, as long as these are are unbiased estimates of the free energy gradient; these gradients would enter the $\hat{\delta F}$ estimators in equation (6.15). Even if the NG determined from this sample would not probably reflect the structure of the ground truth FES, this would not cause any trouble in the FES reconstruction, as long as correct weights are attributed to the log-likelihood terms; again, the estimator performance would rely on the accuracy of the errors on the $\hat{\delta F}$s, a fact which requires a sensible uncertainty estimation on the free energy gradients or mean force[47, 160]. An alternative path, would be to apply to BMTI the same reasoning that, applied to PAk in Chapter 5, led to the punctual reweighting scheme in equation (5.6) for static biases and thus to bPAk. Since, like PAk, BMTI is a maximum-likelihood estimator, we have no reason to believe that BMTI does not perform a kind of fitting analogous to that of PAk. Also, again, recall both methods share the aforementioned double-adaptiveness and the restriction to the intrinsic data manifold. Therefore, it seems safe to apply the punctual reweighting in scheme also to BMTI.

We point out that neither of these two pathways for the generalisation of BMTI to biased samples have been explored so far. Yet, for all the above, but both of them look like promising future directions.

Concerning the smoothness, BMTI has proven to solve the issue that motivated us to develop it. The spurious noise introduced by redundant counts in kernel methods and even more severely by PAk (cfr. 3.2.5.3) is removed by correctly accounting for correlations among estimates at neighbouring points, as discussed in section 6.8. Together with its high accuracy, this will make BMTI our preferred method in many contexts. However, as all TI approaches, it fails to give consistent estimates in the case of disconnected or weakly connected NGs, as discussed commenting e.g. column B of Figure 6.7.

Evidently, BMTI does not have the same versatility as kNN-based methods: approaching a system without prior knowledge the only choice to play it safe appears to be resorting to these methods, which are the most robust. Nonetheless, it would not make sense to reject BMTI for this reason, since it has proven to outperform all other nonparametric methods in most conditions. Thus, we propose the following simple procedure that can be applied as a preliminary screening to assess, in case the ground truth free energy of a system is unknown, whether BMTI estimates on that system will be reliable or not. As we saw, again, in Figure 6.7, the correlation plot between the BMTI estimates and the true free energies becomes visibly disconnected when the NG is disconnected or weakly connected; PAk's correlation plot, instead, is concentrated in a single bulk in all conditions. Therefore, a simple test is looking at the correlation plot between PAk and BMTI estimates: if it is connected, BMTI estimator should be preferred, while PAk should be opted for otherwise.

Finally, the issue of BMTI error estimation is worth a mention. As discussed in section 6.3.2.1, the correct framework to estimate the error on $\{\hat{F}_i\}_i$ is that of the CRB. Due to the redundancy of the information carried by the $\hat{\delta F}$s in BMTI likelihood, the resulting error estimates are underestimated of some orders of magnitude. As discussed in section 6.3.2.2.3, the way we proposed to get around this issue was dividing each term in the BMTI likelihood by an empirical factor accounting for such redundancy. Unfortunately, the proposed factor is not able to consistently unbias BMTI error estimates. It however corrects at least the order of magnitude, maybe a hint that, tweaking the form of the redundancy count, the undertaken path could lead us to fair results. In applications where only the order of magnitude of the error is required, this redundancy-corrected BMTI error estimate can serve the cause.

As of yet, the only way to give an accurate estimate of the free energy errors in BMTI seems to be inverting the full correlation matrix of the $\hat{\delta F}$s, plugging this matrix into the log-likelihood (6.25) and computing the log-likelihood's Hessian as in equation (6.28). This is a path that we have taken only recently, when we have worked out expression (D.23) for the matrix $\mathbf{C}$, so it has not been seriously put to test yet.

Bibliography

[1] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

[2] C. F. Abrams and E. Vanden-Eijnden. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proceedings of the National Academy of Sciences*, 107(11):4961–4966, 2010.

[3] J. B. Abrams and M. E. Tuckerman. Efficient and Direct Generation of Multidimensional Free Energy Surfaces via Adiabatic Dynamics without Coordinate Transformations. *J. Phys. Chem. B*, 112:15742–15757, 2008.

[4] I. S. Abramson. Arbitrariness of the pilot estimator in adaptive kernel methods. *Journal of Multivariate analysis*, 12(4):562–567, 1982.

[5] I. S. Abramson. On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, pages 1217–1223, 1982.

[6] I. S. Abramson. Adaptive Density Flattening–A Metric Distortion Principle for Combating Bias in Nearest Neighbor Methods. *The Annals of Statistics*, 12(3):880 – 886, 1984.

[7] H. Alrachid and T. Lelièvre. Long-time convergence of an adaptive biasing force method: Variance reduction by Helmholtz projection. *SMAI J. Comput. Math.*, 1:55–82, 2015.

[8] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[9] J. An, M. Totrov, and R. Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & Cellular Proteomics*, 4(6):752–761, 2005.

[10] K. Anand, G. J. Palm, J. R. Mesters, S. G. Siddell, J. Ziebuhr, and R. Hilgenfeld. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. *EMBO Journal*, 21(13):3213–3224, 2002.

[11] S. Ausaf Ali, I. Hassan, A. Islam, F. Ahmad, et al. A review of methods available to estimate solvent-accessible surface areas of soluble proteins in the folded and unfolded states. *Current Protein and Peptide Science*, 15(5):456–476, 2014.

[12] U. Bacha, J. Barrila, A. Velazquez-Campoy, S. A. Leavitt, and E. Freire. Identification of Novel Inhibitors of the SARS Coronavirus Main Protease 3CLpro. *Biochemistry*, 43(17):4906–4912, 2004.

[13] F. Baftizadeh, F. Pietrucci, X. Biarnés, and A. Laio. Nucleation process of a fibril precursor in the c-terminal segment of amyloid-β. *Phys. Rev. Lett.*, 110:168103, Apr 2013.

[14] R. Balestriero, J. Pesenti, and Y. LeCun. Learning in high dimension always amounts to extrapolation, 2021.

[15] C. Bartels and M. Karplus. Multidimensional adaptive umbrella sampling: Applications to main chain and side chain peptide conformations. *Journal of Computational Chemistry*, 18(12):1450–1462, 1997.

[16] J. E. Basner and C. Jarzynski. Binless estimation of the potential of mean force. *The Journal of Physical Chemistry B*, 112(40):12722–12729, 2008.

[17] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798 – 1828, 2013.

[18] R. C. Bernardi, M. C. Melo, and K. Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1850(5):872–877, 2015.

[19] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is “nearest neighbor” meaningful? In *Database Theory - ICDT’99*, volume LNCS 1540, pages 217–235, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[20] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML ’06, page 97–104, New York, NY, USA, 2006. Association for Computing Machinery.

[21] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, 1 edition, 2006.

[22] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, et al. Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961–1972, 2009.

[23] M. Breden and C. Kuehn. Rigorous validation of stochastic transition paths. *Journal de Mathématiques Pures et Appliquées*, 131:88–129, 2019.

[24] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.

[25] M. Bzówka, K. Mitusińska, A. Raczyńska, A. Samol, J. A. Tuszyński, and A. Góra. Structural and evolutionary analysis indicate that the sars-COV-2 mpro is a challenging target for small-molecule inhibitor design. *International Journal of Molecular Sciences*, 21(9):3099, 2020.

[26] T. Cacoullos. Estimation of a Multivariate Density. In *Tech. report; No. 40.* University of Minnesota, Department of Statistics, 1964.

[27] F. Camastra and A. Staiano. Intrinsic dimension estimation: Advances and open problems. *Inf. Sci.*, 328:26–41, 2016.

[28] P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Math. Probl. Eng.*, 2015, 2015.

[29] M. Carli and A. Laio. Statistically unbiased free energy estimates from biased simulations. *Molecular Physics*, 119(19-20):e1899323, 2021.

[30] M. Carli, G. Sormani, A. Rodriguez, and A. Laio. Candidate Binding Sites for Allosteric Inhibition of the SARS-CoV-2 Main Protease from the Analysis of Large-Scale Molecular Dynamics Simulations. *The Journal of Physical Chemistry Letters*, 12:65–72, 2020.

[31] N. N. Cencov. Estimation of an unknown distribution density from observations. *Soviet Math.*, 3:1559–1566, 1962.

[32] M. Ceriotti, G. A. Tribello, and M. Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci.*, 108(32):13023–13028, 2011.

[33] L. Y. Chen. Thermodynamic Integration in 3n Dimensions Without Biases or Alchemy for Protein Interactions. *Front. Phys.*, 8(June):1–12, 2020.

[34] M. Chen, M. A. Cuendet, and M. E. Tuckerman. Heating and flooding: A unified approach for rapid generation of free energy surfaces. *J. Chem. Phys.*, 137(2), 2012.

[35] M. Chen, S. Mao, and Y. Liu. Big data: A survey. *Mobile Networks and Applications*, 19:171–209, 2014.

[36] M. Chen, T. Q. Yu, and M. E. Tuckerman. Locating landmarks on high-dimensional free energy surfaces. *Proc. Natl. Acad. Sci. U. S. A.*, 112(11):3235–3240, 2015.

[37] Y. C. Chen. A tutorial on kernel density estimation and recent advances. *Biostat. Epidemiol.*, 1(1):161–187, 2017.

[38] Y. W. Chen, C.-P. B. Yiu, and K.-Y. Wong. Prediction of the sars-cov-2 (2019-ncov) 3c-like protease (3cl pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research*, 9(129), 2020.

[39] Y. Cheng. Mean Shift, Mode Seeking, and Clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, 1995.

[40] C. Chipot and A. Pohorille. *Free Energy Calculations Theory and Applications in Chemistry and Biology.* Springer-Verlag, Berlin, Heidelberg, 1 edition, 2007.

[41] G. Ciccotti, R. Kapral, and E. Vanden-Eijnden. Blue Moon sampling, vectorial reaction coordinates, and unbiased constrained dynamics. *ChemPhysChem*, 6(9):1809–1814, 2005.

[42] F. Cocina, A. Vitalis, and A. Caflisch. Sapphire-Based Clustering. *Journal of Chemical Theory and Computation*, 16(10):6383–6396, 2020.

[43] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[44] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci.*, 102(21):7426–7431, 2005.

[45] P. M. Collins, A. Douangamath, R. Talon, A. Dias, J. Brandao-Neto, T. Krojer, and F. von Delft. Chapter eleven - achieving a good crystal system for crystallographic x-ray fragment screening. In C. A. Lesburg, editor, *Modern Approaches in Drug Discovery*, volume 610 of *Methods in Enzymology*, pages 251 – 264. Academic Press, 2018.

[46] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[47] J. Comer, J. C. Gumbart, J. Hénin, T. Lelievre, A. Pohorille, and C. Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B*, 119(3):1129–1151, 2015.

[48] P. J. Conn, A. Christopoulos, and C. W. Lindsley. Allosteric modulators of gpcrs: a novel approach for the treatment of cns disorders. *Nature reviews Drug discovery*, 8(1):41–54, 2009.

[49] A. Cooper and D. T. Dryden. Allostery without conformational change - A plausible model. *European Biophysics Journal*, 11(2):103–109, 1984.

[50] P. Cossio, A. Laio, and F. Pietrucci. Which similarity measure is better for analyzing protein structures in a molecular dynamics trajectory? *Physical Chemistry Chemical Physics*, 13(22):10421–10425, 2011.

[51] H. Cramér. *Mathematical Methods of Statistics*, volume 9 of *Princeton Mathematical Series*. Princeton University Press, 1946.

[52] D. E. Shaw Research. Molecular Dynamics Simulations Related to SARS-CoV-2, 2020.

[53] E. Darve and A. Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, 115(20):9169–9183, 2001.

[54] E. Darve, D. Rodríguez-Gómez, and A. Pohorille. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.*, 128(14), 2008.

[55] F. De Smet, A. Christopoulos, and P. Carmeliet. Allosteric targeting of receptor tyrosine kinases. *Nature biotechnology*, 32(11):1113–1120, 2014.

[56] F. Dehez, M. Tarek, and C. Chipot. Energetics of ion transport in a peptide nanotube. *The Journal of Physical Chemistry B*, 111(36):10633–10635, 2007. PMID: 17705530.

[57] L. Demortier and L. Lyons. Everything you always wanted to know about pulls. *CDF note*, 43, 2002.

[58] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B*, 39(1):1–22, 1977.

[59] U. n. s. Diamond Light Source. Main protease structure and xchem fragment screen, 2020.

[60] I. Dubanevics and T. C. McLeish. Computational analysis of dynamic allostery and control in the sars-cov-2 main protease. *bioRxiv*, 2020.

[61] M. d'Errico, E. Facco, A. Laio, and A. Rodriguez. Automatic topography of high-dimensional data sets by non-parametric density peak clustering. *Information Sciences*, 560:476–492, 2021.

[62] M. L. Eaton. *Multivariate statistics: a vector space approach.* John Wiley and Sons, New York, 1983.

[63] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[64] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2019.

[65] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, 14:153–158, 1969.

[66] V. Erba, M. Gherardi, and P. Rotondo. Intrinsic dimension estimation for locally undersampled data. *Sci. Rep.*, 9(1):1–9, 2019.

[67] S. Eyrisch and V. Helms. Transient pockets on protein surfaces involved in protein-protein interaction. *Journal of medicinal chemistry*, 50 15:3457–64, 2007.

[68] E. Facco, M. D'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.*, 7(1):1–11, 2017.

[69] A. W. Fenton. Allostery: an illustrated definition for the second secret of life. *Trends in biochemical sciences*, 33(9):420–425, 2008.

[70] G. Fiorin, M. L. Klein, and J. Hénin. Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, 111(22-23):3345–3362, 2013.

[71] E. Fix and J. Hodges. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *USAF School of Aviation Medicine, Randolph Field, Texas*, Report 4(Project Number 21-49-004), 1951.

[72] D. Frenkel and B. Smit. *Understanding Molecular Simulation (Computational Science Series, Vol 1).* Academic Press, 2 edition, 2002.

[73] J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.*, 1(1):55–77, 1997.

[74] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, San Diego, CA, United States, 1990.

[75] K. Fukunaga and L. D. Hostetler. The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition. *IEEE Trans. Inf. Theory*, 21(1):32–40, 1975.

[76] D. Gentile, V. Patamia, A. Scala, M. T. Sciortino, A. Piperno, and A. Rescifina. Putative inhibitors of sars-cov-2 main protease from a library of marine natural products: A virtual screening and molecular modeling study. *Marine drugs*, 18(4):225, 2020.

[77] M. Gentile, F. Courbin, and G. Meylan. Interpolating point spread function anisotropy. *Astron. Astrophys.*, 549, 2012.

[78] I. Gimondi, G. A. Tribello, and M. Salvalaglio. Building maps in collective variable space. *The Journal of Chemical Physics*, 149(10):104104, 2018.

[79] A. Glielmo, M. Carli, I. Macocco, D. Doimo, C. Zeni, A. Rodriguez, M. d'Errico, M. Uhrin, and A. Laio. Dadapy: Distance-based analysis of data-manifolds in python, 2021.

[80] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.*, 121(16):9722–9758, 2021.

[81] A. Glielmo, C. Zeni, B. Cheng, G. Csányi, and A. Laio. Ranking the information content of distance measures. *ArXiv*, abs/2104.15079:1–8, 2021.

[82] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of medicinal chemistry*, 28 7:849–57, 1985.

[83] D. Granata, C. Camilloni, M. Vendruscolo, and A. Laio. Characterization of the free-energy landscapes of proteins by nmr-guided metadynamics. *Proceedings of the National Academy of Sciences*, 110(17):6817–6822, 2013.

[84] D. Granata and V. Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Sci. Rep.*, 6:31377, 2016.

[85] A. Grottesi, N. Bešker, A. Emerson, C. Manelfi, A. R. Beccari, F. Frigerio, E. Lindahl, C. Cerchia, and C. Talarico. Computational Studies of SARS-CoV-2 3CLpro: Insights from MD Simulations. *International Journal of Molecular Sciences*, 21(15):5346, 2020.

[86] A. Gut. *An Intermediate Course in Probability.* Springer Publishing Company, Incorporated, 2nd edition, 2009.

[87] P. Hall and J. S. Marron. Choice of Kernel Order in Density Estimation. *The Annals of Statistics*, 16(1):161 – 173, 1988.

[88] J.-P. Hansen and L. Verlet. Phase transitions of the lennard-jones system. *Phys. Rev.*, 184:151–161, Aug 1969.

[89] B. S. Hanson, L. Dougan, S. A. Harris, and D. J. Read. A generalised mechano-kinetic model for use in multiscale simulation protocols. *bioRxiv*, 2020.

[90] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer, New York, 2 edition, 2009.

[91] A. S. Hauser, M. M. Attwood, M. Rask-Andersen, H. B. Schiöth, and D. E. Gloriam. Trends in gpcr drug discovery: new agents, targets and indications. *Nature reviews Drug discovery*, 16(12):829–842, 2017.

[92] N.-B. Heidenreich, A. Schindler, and S. Sperlich. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Adv. Stat. Anal.*, 97(4):403–433, 2013.

[93] J. Hénin. Fast and Accurate Multidimensional Free Energy Integration. *J. Chem. Theory Comput.*, 17(11):6789–6798, 2021.

[94] J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein. Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. *J. Chem. Theory Comput.*, 6(1):35–47, 2010.

[95] G. Henkelman and H. Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 113(22):9978–9985, 2000.

[96] G. Henkelman, B. P. Uberuaga, and H. Jónsson. Climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, 113(22):9901–9904, 2000.

[97] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49:409–436, 1952.

[98] S. Honda, T. Akiba, Y. S. Kato, Y. Sawada, M. Sekijima, M. Ishimura, A. Ooishi, H. Watanabe, T. Odahara, and K. Harata. Crystal structure of a ten-amino acid protein. *Journal of the American Chemical Society*, 130(46):15327–15331, 2008.

[99] W. Humphrey, A. Dalke, and K. Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 2 1996.

[100] H. A. Hussein, A. Borrel, C. Geneix, M. Petitjean, L. Regad, and A.-C. Camproux. Pockdrug-server: a new web server for predicting pocket druggability on holo and apo proteins. *Nucleic acids research*, 43(W1):W436–W442, 2015.

[101] A. J. Izenman. Recent Developments in Nonparametric Density Estimation. *J. Am. Stat. Assoc.*, 86(413):205, 1991.

[102] A. Jiménez-Alberto, R. M. Ribas-Aparicio, G. Aparicio-Ozores, and J. A. Castelán-Vega. Virtual screening of approved drugs as potential sars-cov-2 main protease inhibitors. *Computational biology and chemistry*, 88:107325, 2020.

[103] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, Y. Duan, J. Yu, L. Wang, K. Yang, F. Liu, R. Jiang, X. Yang, T. You, X. Liu, X. Yang, F. Bai, H. Liu, X. Liu, L. W. Guddat, W. Xu, G. Xiao, C. Qin, Z. Shi, H. Jiang, Z. Rao, and H. Yang. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293, 2020.

[104] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer Science & Business Media, New York, NY, Mar. 2013.

[105] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A*, 374(2065):20150202, 2016.

[106] M. Kandeel and M. Al-Nazawi. Virtual screening and repurposing of fda approved drugs against covid-19 main protease. *Life sciences*, page 117627, 2020.

[107] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3(5):300–313, 1935.

[108] M. N. Kobrak. Systematic and statistical error in histogram-based free energy calculations. *Journal of Computational Chemistry*, 24(12):1437–1446, 2003.

[109] D. Kokh, S. Richter, S. Henrich, P. Czodrowski, F. Rippmann, and R. Wade. Trapp: A tool for analysis of transient binding pockets in proteins. *Journal of chemical information and modeling*, 53 5:1235–52, 2013.

[110] F. Korn, B.-U. Pagel, and C. Faloutsos. On the "dimensionality curse" and the "self-similarity blessing". *IEEE Transactions on Knowledge and Data Engineering*, 13(1):96–111, 2001.

[111] M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.*, 37(2):233–243, 1991.

[112] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[113] V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1):1–11, 2009.

[114] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[115] B. Lee and F. Richards. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55(3):379–IN4, 1971.

[116] T. Lelièvre, M. Rousset, and G. Stoltz. *Free Energy Computations A Mathematical Perspective.* Imperial College Press, 2010.

[117] D. Levitt and L. Banaszak. Pocket: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of molecular graphics*, 10 4:229–34, 1992.

[118] C. P. Loftsgaarden, D. O.; Quesenberry. A Nonparametric Estimate of a Multivariate Density Function. *Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[119] B. Ma, C. J. Tsai, T. Haliloğlu, and R. Nussinov. Dynamic allostery: Linkers are not merely flexible. *Structure*, 19(7):907–917, 2011.

[120] Y. P. Mack and M. Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, 1979.

[121] L. Maragliano and E. Vanden-Eijnden. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.*, 426(1-3):168–175, 2006.

[122] L. Maragliano and E. Vanden-Eijnden. Single-sweep methods for free energy calculations. *J. Chem. Phys.*, 128(18):1–10, 2008.

[123] E. Marinari and G. Parisi. Trattatello di probabilit a-teoria delle probabilit a per fisici, computational scientists and computer scientists. 2000.

[124] V. Marinova and M. Salvalaglio. Time-independent free energies from metadynamics via mean force integration. *The Journal of Chemical Physics*, 151(16):164115, 2019.

[125] I. R. McDonald and K. Singer. Machine calculation of thermodynamic properties of a simple fluid at supercritical temperatures. *The Journal of Chemical Physics*, 47(11):4766–4772, 1967.

[126] H. Meirovitch, S. Cheluvaraja, and R. White. Methods for Calculating the Entropy and Free Energy and their Application to Problems Involving Protein Flexibility and Ligand Binding. *Curr. Protein Pept. Sci.*, 10(3):229–243, 2009.

[127] J. Meixner. Coldness and temperature. *Archive for Rational Mechanics and Analysis*, 57:281–290, 1975.

[128] L. Mittal, A. Kumari, M. Srivastava, M. Singh, and S. Asthana. Identification of potential molecules against covid-19 main protease through structure-guided virtual screening approach. *Journal of Biomolecular Structure and Dynamics*, 39(10):3662–3680, 2021. PMID: 32396769.

[129] J. Monod, J.-P. Changeux, and F. Jacob. Allosteric proteins and cellular control systems. *Journal of Molecular Biology*, 6(4):306 – 329, 1963.

[130] I. Müller. The coldness, a universal function in thermoelastic bodies. *Archive for Rational Mechanics and Analysis*, 41:319–332, 1971.

[131] K. Müller and L. D. Brown. Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theoretica chimica acta*, 53:75–93, 1979.

[132] T. Nagler and C. Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[133] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[134] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, 231:289–337, 1933.

[135] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.*, 71:361–390, 2020.

[136] R. Nussinov. Introduction to protein ensembles and allostery. *Chemical Reviews*, 116(11):6263–6266, 2016.

[137] R. Nussinov and C.-J. Tsai. Allostery in disease and in drug discovery. *Cell*, 153(2):293–305, 2013.

[138] C. Owen, P. Lukacik, C. Strain-Damerell, A. Douangamath, A. Powell, J. Fearon, D.and Brandao-Neto, A. Crawshaw, M. Aragao, D.and Williams, R. Flaig, D. Hall, K. McAauley, D. Stuart, F. von Delft, and M. Walsh. Rcsb pdb - 6y84: Sars-cov-2 main protease with unliganded active site (2019-ncov, coronavirus disease 2019, covid-19).

[139] A. Ozakin and A. Gray. Submanifold density estimation. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[140] E. S. Page and R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, Princeton, NJ, 1961.

[141] S. Pant, M. Singh, V. Ravichandiran, U. Murty, and H. K. Srivastava. Peptide-like and small-molecule inhibitors against covid-19. *Journal of Biomolecular Structure and Dynamics*, pages 1–10, 2020.

[142] T. Paramo, A. East, D. Garzón, M. B. Ulmschneider, and P. Bond. Efficient characterization of protein cavities within molecular simulation trajectories: trj_cavity. *Journal of chemical theory and computation*, 10 5:2151–64, 2014.

[143] C. Pargellis, L. Tong, L. Churchill, P. F. Cirillo, T. Gilmore, A. G. Graham, P. M. Grob, E. R. Hickey, N. Moss, S. Pav, et al. Inhibition of p38 map kinase by utilizing a novel allosteric binding site. *Nature structural biology*, 9(4):268–272, 2002.

[144] B. U. Park and J. S. Marron. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association*, 85(409):66–72, 1990.

[145] E. Parzen. On the Estimation of Probability Density Functions and Mode. *Ann. Math. Statist*, 33:1065–1076, 1962.

[146] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2(7-12):559–572, 1901.

[147] L. Peliti and S. Pigolotti. *Stochastic Thermodynamics: An Introduction.* Princeton University Press, 2021.

[148] B. Pelletier. Kernel density estimation on riemannian manifolds. *Statistics & Probability Letters*, 73:297–304, 2005.

[149] R. Penrose. On best approximate solutions of linear matrix equations. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(1):17–19, 1956.

[150] B. Peters. *Reaction Rate Theory and Rare Events.* Elsevier, 2017.

[151] S. Piana and A. Laio. Advillin folding takes place on a hypersurface of small dimensionality. *Phys. Rev. Lett.*, 101(20):1–4, 2008.

[152] F. Pietrucci. Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Reviews in Physics*, 2:32–45, 2017.

[153] T. Pillaiyar, M. Manickam, V. Namasivayam, Y. Hayashi, and S. H. Jung. An overview of severe acute respiratory syndrome-coronavirus (SARS-CoV) 3CL protease inhibitors: Peptidomimetics and small molecule chemotherapy. *Journal of Medicinal Chemistry*, 59(14):6595–6628, 2016.

[154] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, 134(17):174105, 2011.

[155] C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

[156] G. N. Ramachandran and V. Sasisekharan. Conformation of Polypeptides and Proteins. *Advances in Protein Chemistry*, 23(C):283–437, 1968.

[157] A. Rodriguez, M. D'Errico, E. Facco, and A. Laio. Computing the Free Energy without Collective Variables. *J. Chem. Theory Comput.*, 14(3):1206–1215, 2018.

[158] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[159] A. Rodriguez, P. Mokoema, F. Corcho, K. Bisetty, and J. J. Perez. Computational study of the free energy landscape of the miniprotein CLN025 in explicit and implicit solvent. *Journal of Physical Chemistry B*, 115(6):1440–1449, 2011.

[160] D. Rodriguez-Gomez, E. Darve, and A. Pohorille. Assessing the efficiency of free energy calculation methods. *J. Chem. Phys.*, 120(8):3563–3578, 2004.

[161] M. Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837, 1956.

[162] M. Rosenblatt. Global measures of deviation for kernel and nearest neighbor density estimates. In T. Gasser and M. Rosenblatt, editors, *Smoothing Techniques for Curve Estimation*, pages 181–190, Berlin, Heidelberg, 1979. Springer Berlin Heidelberg.

[163] L. Rosso, P. Mináry, Z. Zhu, and M. E. Tuckerman. On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *The Journal of Chemical Physics*, 116(11):4389–4402, 2002.

[164] T. O. S. Hovmöller, T. Zhou. Conformations of amino acids in proteins research papers. *Biological Crystallography*, D58:768–776, 2002.

[165] G. L. Sala, S. Decherchi, M. D. Vivo, and W. Rocchia. Allosteric communication networks in proteins revealed through pocket crosstalk analysis. *ACS Central Science*, 3:949 – 960, 2017.

[166] H. Sasaki, Y.-K. Noh, and G. Niu. Direct Density Derivative Estimation. *Neural Comput.*, 28(6):1101–1140, 2016.

[167] J. Schmidhuber. Deep Learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[168] P. Schmidtke, A. Bidon-Chanal, F. J. Luque, and X. Barril. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*, 27(23):3276–3285, 10 2011.

[169] B. Scholkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, 1998.

[170] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey, USA, 2 edition, 2015.

[171] U. Seifert. Stochastic thermodynamics: From principles to the cost of precision. *Physica A-statistical Mechanics and Its Applications*, 504:176–191, 2017.

[172] B. D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. Mcleavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Communications of the ACM: Volume 51 Issue 7. *Commun. ACM*, pages 91–97, 2008.

[173] D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L. S. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y. H. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. Ben Schafer, N. Siddique,

C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, and C. Young. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. *Int. Conf. High Perform. Comput. Networking, Storage Anal. SC*, 2015-Janua(January):41–53, 2014.

[174] M. R. Shirts and J. D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129(12):124105, 2008.

[175] M. R. Shirts and V. S. Pande. Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration. *The Journal of Chemical Physics*, 122(14):144107, 2005.

[176] A. Shrake and J. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351–371, 1973.

[177] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[178] B. W. Silverman and M. C. Jones. E. fix and j.l. hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). *International Statistical Review*, 57:233, 1989.

[179] G. Sormani, A. Rodriguez, and A. Laio. Explicit Characterization of the Free-Energy Landscape of a Protein in the Space of All Its Cα Carbons. *Journal of Chemical Theory and Computation*, 16(1):80–87, 2020.

[180] C. J. Stone. An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates. *The Annals of Statistics*, 12(4):1285 – 1297, 1984.

[181] T. Strutz. *Data fitting and uncertainty: A practical introduction to weighted least squares and beyond*. Springer, 2011.

[182] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

[183] T. Sztain, R. Amaro, and J. A. McCammon. Elucidation of cryptic and allosteric pockets within the SARS-CoV-2 protease. *bioRxiv*, 2020.

[184] Z. Tan, E. Gallicchio, M. Lapelosa, and R. M. Levy. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *The Journal of Chemical Physics*, 136(14):144102, 2012.

[185] Z. Tan, E. Gallicchio, M. Lapelosa, and R. M. Levy. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *The Journal of Chemical Physics*, 136(14):144102, 2012.

[186] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[187] B. Y. G. R. Terrell and D. W. Scott. Variable Kernel Density Estimation. *Ann. Stat.*, 20(3):1236–1265, 1992.

[188] W. S. Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, Jan. 1952.

[189] J. Torrie, G. M Valleau. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.*, 23:187, 1977.

[190] M. Tuckerman. *Statistical mechanics: theory and molecular simulation.* Oxford university press, 2010.

[191] B. A. Turlach. Bandwidth selection in kernel density estimation: A review. In *CORE and Institut de Statistique*, 1993.

[192] M. T. ul Qamar, S. M. Alqahtani, M. A. Alamri, and L.-L. Chen. Structural basis of sars-cov-2 3clpro and anti-covid-19 drug discovery from medicinal plants. *Journal of pharmaceutical analysis*, 10(4):313–319, 2020.

[193] G. Upton and I. Cook. *A Dictionary of Statistics.* Oxford University Press, 2008.

[194] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[195] O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.*, pages 1–12, 2020.

[196] J. Wagner, J. Sorensen, N. Hensley, C. Wong, C. Zhu, T. Perison, and R. Amaro. Povme 3.0: Software for mapping binding pocket flexibility. *Journal of chemical theory and computation*, 13 9:4584–4592, 2017.

[197] J. R. Wagner, C. T. Lee, J. D. Durrant, R. D. Malmstrom, V. A. Feher, and R. E. Amaro. Emerging computational methods for the rational discovery of allosteric drugs. *Chemical reviews*, 116(11):6370–6390, 2016.

[198] T. R. Weikl and F. Paul. Conformational selection in protein binding and function. *Protein Science*, 23(11):1508–1518, 2014.

[199] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

[200] P. Wu, M. H. Clausen, and T. E. Nielsen. Allosteric small-molecule kinase inhibitors. *Pharmacology & therapeutics*, 156:59–68, 2015.

[201] H. Yang, M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, et al. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences*, 100(23):13190–13195, 2003.

[202] G. Young and A. S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, Mar. 1938.

[203] R. Yousefzadeh. Deep learning generalization and the convex hull of training sets, 2021.

[204] C. Zeni, A. Anelli, A. Glielmo, and K. Rossi. Machine learning potentials always extrapolate, it does not matter, 2021.

[205] G. H. Zerze, C. M. Miller, D. Granata, and J. Mittal. Free energy surface of an intrinsically disordered protein: comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics. *Journal of chemical theory and computation*, 11 6:2776–82, 2015.

[206] B. W. Zhang, S. Arasteh, and R. M. Levy. the uwham and swham software package. *Scientific reports*, 9(1):1–9, 2019.

[207] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, and R. Hilgenfeld. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved a-ketoamide inhibitors. *Science*, 368(6489):409–412, 2020.

[208] P. Zhou, X. L. Yang, X. G. Wang, B. Hu, L. Zhang, W. Zhang, H. R. Si, Y. Zhu, B. Li, C. L. Huang, H. D. Chen, J. Chen, Y. Luo, H. Guo, R. D. Jiang, M. Q. Liu, Y. Chen, X. R. Shen,

X. Wang, X. S. Zheng, K. Zhao, Q. J. Chen, F. Deng, L. L. Liu, B. Yan, F. X. Zhan, Y. Y. Wang, G. F. Xiao, and Z. L. Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273, 2020.

[209] M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, et al. Sars-cov-2 simulations go exascale to capture spike opening and reveal cryptic pockets across the proteome. *bioRxiv*, 2020.

Appendix A

Test systems

A.1 1-dim. double well potential

This potential, adopted in Figure 3.2.5.3, has the form:

$$U(\mathbf{x}) = -\log\left[\frac{1}{2}\mathcal{N}(-0.5, 1) \; + \; \frac{1}{2}\mathcal{N}(1.6, 0.5)\right] \tag{A.1}$$

A.2 2-dim. double well potential

The PDF of this potential, represented in panel A of Figure 5.1, has the form:

$$U_{2d} \; := \; \left(2\, e^{-(x-1.5)^2 - (y-2.5)^2} + 3\, e^{-2\,x^2 - 0.25\,y^2}\right)^3 \tag{A.2}$$

A.3 2-dim. modified Mueller-Brown potential on a glassy background

The PDF corresponding to this potential is obtained by superimposing on a box $[-4, 4] \times [-2, 2]$ with periodic boundary conditions the following distributions:

- a bivariate multi-peak PDF of form:

$$\begin{aligned} f_{\mathrm{mp}} \quad := \quad & 0.11\,[3.4\, e^{-6.5\,(x+1)^2 + 11\,(x+1)(y-0.5) - 6.5\,(y-0.5)^2} \\ & \qquad +2\, e^{-(x+0.5)^2 - 10\,(y+0.5)^2} + 4\, e^{-(x-0.5)^2 - 10\,(y+1)^2}] \end{aligned} \tag{A.3}$$

- 90 rescaled bivariate Gaussians $0.005 \times \mathcal{N}\left(\mathbf{c}, 0.04 \cdot \mathbb{1}_2\right)$ with the centres $\mathbf{c}$ randomly sampled in the rectangle $[-3.6, 3.6] \times [-1.8, 1.8]$; their integral is 0.45;

- a uniform background such that added to U_{mp} they integrate to 0.55;

Adding these three contributions to we obtain a PDF that, as in Chapter 5, we call $p(\mathbf{x})$. Then we can also consider $p(\mathbf{x})^{10}$ re-normalised to 1, which for further reference we simply indicate by p^{10}:

$$p^{10}(\mathbf{x}) := \frac{(p(\mathbf{x}))^{10}}{\int (p(\mathbf{x}))^{10}\,\mathrm{d}\mathbf{x}} \;.$$

The two systems display metastability between the two main basins. By construction, the potential barrier in p^{10} is exactly 10 times as high as the one in p.

A.4 2-dim. Mueller-Brown potential

This is the classical bivariate Mueller-Brown potential[131], whose expression is:

$$\begin{aligned} U_{\mathrm{MB}} \;\; := \;\; & 15\, e^{0.7\,(x+1)^2 + 0.6\,(x+1)(y-1) + 0.7\,(y-1)^2} \; - \; 200\, e^{-(x-1)^2 - 10\, y^2} \\ & - 100\, e^{-x^2 - 10\,(y-0.5)^2} \; - \; 170\, e^{-6.5\,(x+0.5)^2 + 11\,(x+0.5)(y-1.5) - 6.5\,(y-1.5)^2} \end{aligned} \tag{A.4}$$

and whose contour plot can be seen in panel A of Figure A.4. From this potential, we generated various samples at various temperatures. In particular, we focused on a range of temperatures around which all three basins of the potential were visited even extracing only 5000 points. We found that for the inverse value of coldness $\beta = 0.035$ the saddle points are also fairly populated, while halving the temperature, at $\beta = 0.07$, the points outside the minima are rarer, as can be seen in panels B1 and B2 of Figure A.4. The two systems display free energy barriers from the global minimum to the neighbouring basin which are around $\sim 3.7k_BT$ and $\sim 3.7k_BT$ high respectively, as visible in panel C of Figure A.4. We use these two settings in order to test our free energy estimators in different conditions of connectivity of the neighbours graph between sample points.

In order to compute the minimum energy path (MEP) connecting the two main minima we use the Nudged Elastic Band (NEB) algorithm[96] in its improved tangent formulation [95] with 32 images. For the exact location of the two minima we use the values in reference [23]. We call the MEP for this system the polygonal chain that linearly interpolates between the 32 images. Next, we want to find a path as close as possible to the MEP but which only connects points in the sample. We sample our MEP homogeneously 20 times for each image, so that we extract a set of 621 points along the MEP. For each of these MEP points, we look for its nearest neighbour in the data sample we are considering. If the distance between the MEP point and the NN in the dataset is below a given threshold we keep the point, otherwise we reject it. The collection of all these sample points forms what we call the NN-interpolated MEP. For both data samples of 5000 points extracted from

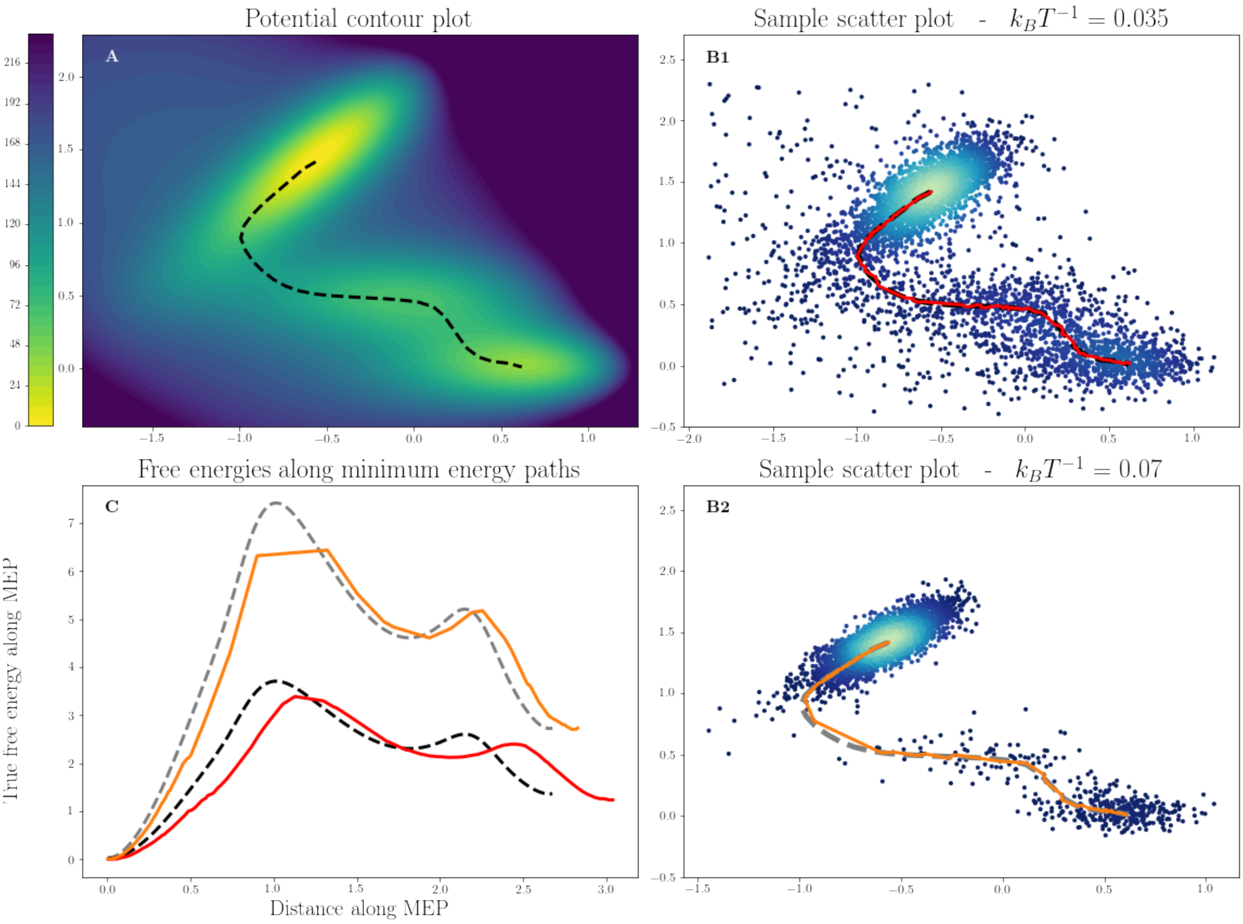


Figure A.1: Illustration of the Mueller-Brown potential used as test system. **A** Contour plot of the Mueller-Brown potential in equation (A.4). For the reader's convenience, the minimum of the potential has been shifted to 0. Also, for a better readability, the colour map has been cutoffed to 230, otherwise it would be saturated by the diverging behaviour in the top right corner. The black dashed curve represents the MEP connecting the two minima computed via the NEB algorithm. **B1-2** Scatter plots of two sets of 5000 points sampled from the Mueller-Brown potential. B1: The thermodynamic beta is $\beta = 0.035$; in dashed black the MEP (same as panel A), in red the NN-interpolated MEP, the path connecting points within a distance 10^{-2} from the real MEP. B2: The thermodynamic beta is is $\beta = 0.07$; in dashed grey the MEP (same as panel A), in yellow the NN-interpolated MEP, the path connecting points within a distance 10^{-2} from the real MEP. **C** Analytic free energies for the two systems along the MEPs. In solid red and dashed black the free energies at $\beta = 0.035$ along the the MEP and the NN-interpolated MEP respectively; in solid yellow and dashed grey the free energies at $\beta = 0.07$ along the the MEP and the NN-interpolated MEP respectively.

the Mueller-Brown potential at $\beta = 0.035$ and $\beta = 0.07$ we consider a NN interpolation threshold of 2×10^{-2}. In the first sample the NN-interpolated MEP contains 460 points; in the second sample the number of points satisfying the threshold requirement is reduced to 280. The MEP is clearly visible as the dashed curve in panel A of Figure A.4. The NN-interpolated MEPs are represented in coloured solid lines panels B1 and B2. The ground truth free energies along the various paths are visible in panel C.

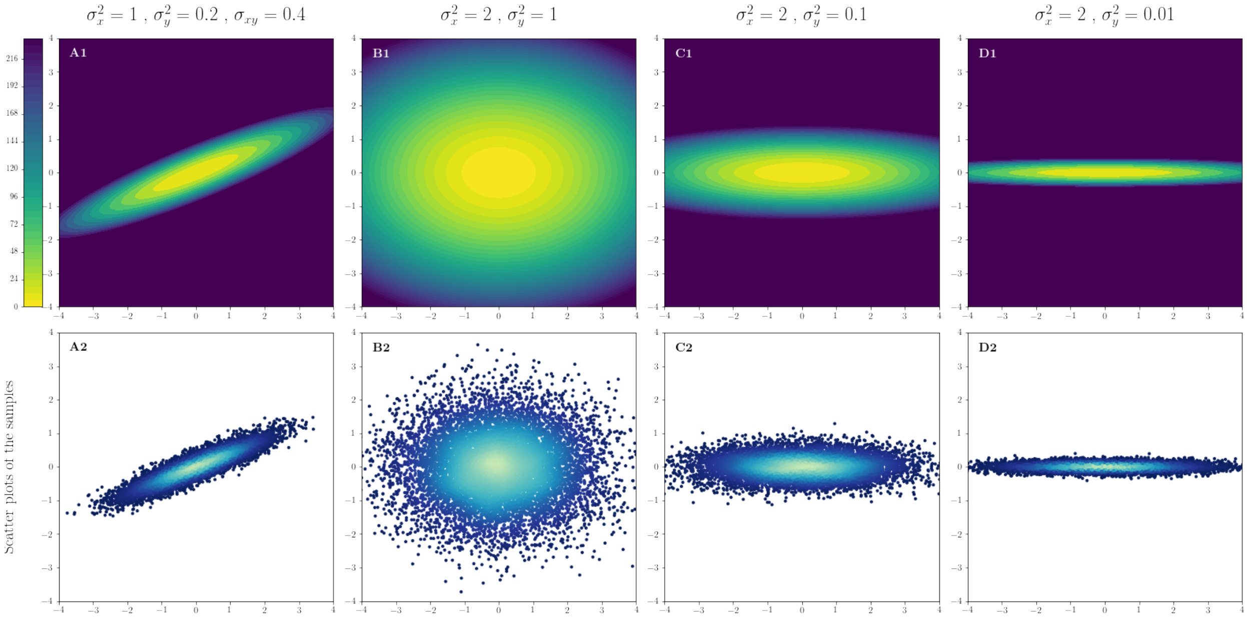


Figure A.2: Bivariate potentials from four Gaussian distributions used as test system. Each column represents a different system. All Gaussians are centered at the origin. The parameters of each system's covariance matrix are indicated in the header of each column. Top: contour plots of the potential surfaces. Bottom: four samples of 10000 points from the above potentials.

A.5 2-dim. Gaussian distibutions

These four systems have bivariate normal distributions $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. We name the two elements on the diagonal of this matrix σ_x^2 and σ_y^2, while the two identical off-diagonal terms are σ_{xy}. The first Gaussian, whose corresponding potential is represented in panel A1 of Figure A.2, has $\sigma_x^2 = 1$, $\sigma_y^2 = 0.2$ and $\sigma_{xy} = 0.4$. The second system, in panel B1, has $\sigma_x^2 = 2$, $\sigma_y^2 = 1$ and $\sigma_{xy} = 0$. The third system, in panel C1, has $\sigma_x^2 = 2$, $\sigma_y^2 = 0.1$ and $\sigma_{xy} = 0$. The fourth system, in panel D1, has $\sigma_x^2 = 2$, $\sigma_y^2 = 0.01$ and $\sigma_{xy} = 0$. In the bottom row of the figure scatter plots of samples of 10000 points from the above potentials are shown.

A.6 6-dim. potential: 2-dim. double well potential plus 4-dim. harmonic directions

This potential in the first two dimensions is exactly the same as the one in section A.2. The additional 4 dimensions feel a (convex) harmonic potential centered at the origin and with unitary curvature.

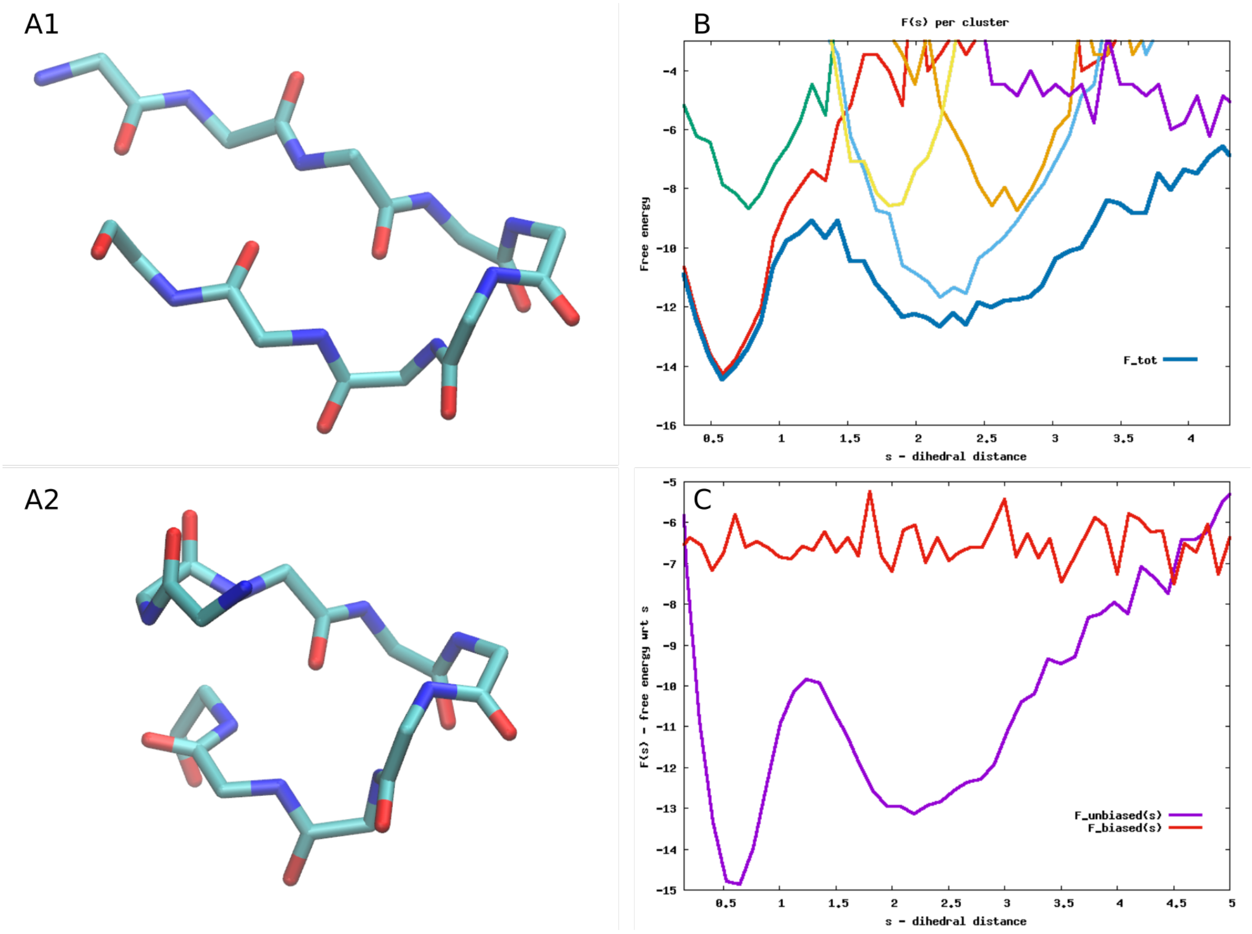


Figure A.3: The CLN025 β-hairpin used as test system. **A1** and **A2** Visualisation of two metastable states of the peptide, found by clustering[61] after estimating the FES with PAk. **A1** A configuration in the global minimum of the free energy, corresponding to the crystal structure. **A2** A configuration in the second most populated cluster, corresponding to the second minimum. **B** Free energies along the one-dimensional ψ-dihedral distance collective variable s defined in equation (5.7) and computed on a histogram for a dataset of 4000 points. In blue the free energy profile of the whole dataset. In other colours the individual contributions of the 6 main clusters to the one dimensional free energy $F(s)$. The clusters are found as described in section A.7.1. **C** Free energy profiles $F(s)$ computed on a histogram fo two datasets of 9500 points (one unbiased and one biased).

A.7 CLN025 decapeptide

As a realistic system we consider a β-hairpin called CLN025[98]. This molecule is a small protein of 10 residues and 166 atoms and is one of the smallest peptides that display a stable secondary structure, in this case a β-sheet. Thanks to the relatively small size of the molecule we are able to produce both a long unbiased MD trajectory and a biased one. We simulate the protein in Gromacs[1] in explicit solvent. Since we are not interested in the precise phisical chemistry of the system, we use quite a small box, resulting in a total of 2959 atoms, 166 of CLN025, the rest from the 931 water molecules. To enhance the sampling of configuration space. We run a Replica Exchange MD[182] simulation with 16 replicas using equally spaced temperatures from 340K to 470K as done previously in reference [159]. In panels A1 and A2 Figure A.3 the visualisation of two metastable states of the peptide. The one in Figure A1 corresponds to the crystal structure.

A.7.1 9-dim. ψ dihedrals metric

The first feature space chosen is the 9-dimensional ψ-backbone-dihedra space. This choice implies of course a drastic dimensional reduction on the over-400-dimensional original atomic configuration space; still, even after this huge projection the system will show complex features and a reasonably high dimensionality, so that we are entitled to consider it a realistic case. Thus $D = 9$ is the embedding space dimension. The distance between two configurations $\mathbf{X}^a$ and $\mathbf{X}^b$ in this space is:

$$\theta(\mathbf{X}^a, \mathbf{X}^b) = \sqrt{\sum_n ((\psi_n^a - \psi_n^b))^2} \tag{A.5}$$

where $((\bullet))$ stands for 2π-periodicity within the brackets.

In order to generate a biased trajectory we apply the procedure described in section 5.3.2 of Chapter 5. The free energies along the one-dimensional ψ-dihedral distance collective variable s defined in equation (5.7) and computed on a histogram for two datasets of 9500 points (one unbiased and one biased) can be seen in panel C of Figure A.3.

To find the relevant clusters of the system we extract 4000 points from the trajectory. We then compute the ID of the system as $d = 7$. We estimate the free energy estimate in this space with PAk. Then we analyse the FES by a density peak clustering algorithm[61] and we find 6 main clusters. Individual contributions of these 6 clusters to the free energy profile $F(s)$ are visible in panel C of A.3. Configurations from the two main clusters are visualised in panels A1 and A2.

A.7.1.1 9-dim. ψ dihedrals metric, analytic potential via Gaussian KDE smoothing

We generate a sample of 38000 points. We analyse it in the space of the ψ-backbone-dihedrals. The estimated ID[68] is $d = 7$. With such ID we generate a 9-dimensional smooth potential using our adaptive Gaussian KDE presented in Appendix B of which we know the analytic value everywhere.

A.7.2 45-dim. alpha carbon distances metric

The second featurisation chosen is that of the distances between all alpha carbons (C_α). Since we have 10 residues there are 10 alpha carbons, thus $10 \times 9/2 = 45$ pairwise distances between them; therefore the embedding dimension of the chosen space is now $D = 45$. The metric on this space is the RMSD:

$$d_{RMS}(\mathbf{X}^a, \mathbf{X}^b) = \sqrt{\frac{2}{N(N-1)} \sum_{i>j} (d^a_{ij} - d^b_{ij})^2} \tag{A.6}$$

where $\mathbf{X}^a$ and $\mathbf{X}^b$ are two configurations in the 45-dimensional space and d^q_{ij} is the distance between the i-th and the j-th C_α in configuration $\mathbf{X}^q$.

A.8 20-dim. embeddings

We considered three systems used for the validation of PAk estimator in reference [157]and briefly drescribed in the following three subsections. They are trajectories of respectively 2,4 and 7 CVs of which the ground truth free energy is known. All the systems were treated in the same way. Initially, the FES is resampled in the space of the collective variables with a probability proportional to the exponential of negative of the free energy value. Then, the data points are twisted on a Swiss-roll by splitting the first of its coordinates in two by means of the transformation x 1 = x cos(x) and x 2 = x sin(x). Finally a rotation around a random vector in $D = 20$ is performed. In this manner each point sampled from the original distribution is embedded in a 20-dimensional space. See

A.8.1 2-dim. system before embedding

The original system before embedding is the projection on two collective variables of the nucleation of the C-terminal of amyloid-β[13].

A.8.2 4-dim. system before embedding

The original system before embedding is the projection on four collective variables of the folding of the third IgG-binding domain of protein G from streptococcal bacteria (GB3)[83].

A.8.3 7-dim. system before embedding

The original system before embedding is the projection on seven collective variables of the conformational space of the intrinsically disordered protein human islet amyloid polypeptide (hIAPP)[205].

Appendix B

Adaptive neighbourhood size selection in kernel density estimation: PAkde

The framework that allows to adaptively select the neighbourhood size for points in a sample and the restriction to the intrinsic data manifold, introduced in sections 3.1.3 and 3.1.1 of Chapter 3 can be employed also as a method for the bandwidth selection of kernel density estimators. The optimal values $\hat{k}$ in equation (3.7) give a quantitative measure of the maximum number of neighbours that can be included from the sample in a hypersphere centred on a given point i maintaining the PDF value within that hypervolume compatible with a constant according to some tolerance D_{thr}. Fixing a set of N neighbourhood sizes $\{\hat{k}_i\}_i$ returns as a side product also a set of lengthscales $\{r_i\}_i$: the smallest a r_i, the faster will be the variations of the PDF around point i and vice-versa.

This fact can be exploited also in the selection of the bandwidth h of any kernel density estimator of the form in equation (2.17), namely: $\hat{\rho}\left(\mathbf{x}\right) = \frac{1}{N}\sum_j^N \kappa_h\left(\mathbf{x}-\mathbf{x}_j\right)$. Based on the functional form of the kernel κ_h, its shape and tail properties, the $\{r_i\}_i$ scaled by some factor λ can be employed as local smoothing parameters, so that at any point i the selected bandwidth becomes $\hat{h}_i := \lambda r_i$.

We have tested this method in case of Gaussian KDEs on various systems and it has proven to considerably improve the estimates of fixed-bandwidth Gaussian KDEs. In a range of embedding dimensions ranging from $D=2$ to $D=9$ most optimal values for λ were between $\frac{1}{2}$ and $\frac{1}{3}$. We choose the factor $\lambda = \frac{1}{2}$ and define our point-adaptive Gaussian KDE:

$$\hat{\rho}\left(\mathbf{x}\right) = \frac{1}{N}\sum_j^N \kappa_{\hat{h}_j}\left(\mathbf{x}-\mathbf{x}_j\right) \ , \tag{B.1}$$

with $\hat{h}_i := \frac{1}{2} r_i$ for each point i. We call the estimator in equation (B.1) the Point-Adaptive (Gaussian) kernel density estimator, or PAkde. Its performance on several systems across a wide range of embedding and intrinsic dimensionalities can be examined in Chapter 6 in Table 6.1 and is discussed in section 6.3.2.2.1. A more refined bandwidth selection criterion should include a dependency on the dimensionality such as e.g. in Scott's or Silverman's rules of thumb[170, 177]. We will further investigate in this direction.

Appendix C

Structural description of the metastable states

We here present a description of all 18 metastable states in terms of their local contact structure and backbone arrangement and of the two observables SASA and PDA. Some parts of the description might be redundant with the main text of the manuscript.

From the analysis of the maximum residence time it is clear that states 1 and 2 of both m1 and m2 are among the longest-lived metastable states. All four are in fact very similar to the crystallographic structure (PDB 6Y84): they all have the left flap and the linker loop in contact between each other (cont. Met^{49}-Gln^{189}); the left flap is closed (cont. Glu^{47}-Leu^{57} broken, cont. Thr^{25}-Cys^{44} formed) and the linker loop stretched towards it (cont. Leu^{167}-Arg^{188} broken), covering the lower part of the binding pocket. The contact and backbone structures of states m2:1 and m2:2 are almost identical and even a visual inspection with the software VMD confirms the two states can be considered in practice as the same metastable state (even the SASA and PDVA have compatible values within errorbars); the difference between states m2:1,m2:2 and m1:1,m1:2 is the fact that the latter two have the F140-C145 loop (we call it *upper flap*) tilted downwards (contacts 28 vs 143-144 and 118 vs 142 not formed, dihedral 144 in β instead of α configuration), which hides the catalytic Cys^{145}, resulting in a slightly lower SASA and PDVA. The differences between m1:1 and m1:2, instead, are mostly in the linker loop, which in m1:2 is wider in proximity of the pocket (cont. 185-186 vs 192 not formed) and narrower towards the end (contacts 132 vs 196 and 197-198 vs 238 formed, 131 vs 199 not formed).

Two other states which are similar to each other in terms of contact structure are m2:6 and m2:7. The upper flap is not bent downwards (dihedral 144 in α configuration, as most of the states in m2), leaving some SASA for the catalytic Cys^{145}. In m2:7 the left flap is more stretched towards the linker loop, and the linker loop is open wider, granting slightly lower PDVA and SASA. In both cases, however, the catalytic dyad is quite accessible.

Then there are states m1:9 and m1:10 which are very similar in their contact and backbone structure, with the exception of the left flap, which is much more open in state m1:10. States m1:9 and m1:10 (especially the former) are then both structurally similar to m1:7: the only difference among the contacts is 132 vs 196, which is formed in m1:7 and not formed in m1:9 and m1:10, allowing the lower loop to be more flexible. In all three states the upper flap is tilted downwards; surprisingly, despite the fact that the left flap is wide open, two out of these three states are detected as closed by our observables. In m1:9 the side-chains of the residues in the loops surrounding the binding pocket are oriented towards the catalytic dyad, causing such state to rank among the lowest in SASA; moreover, cont. 285 vs 285* in this state is not completely formed (*n* configuration). State m1:7 ranks among the lowest in PDVA and as the lowest in SASA; the reason lies in the sidechains of the lower and left flaps, in particular of Thr^{45} and Gln^{189}, which form a contact and effectively close the access to the reactive site.

Another couple of similar states is that of m1:4 and m1:11: they characterised by a very open left flap (cont. 47 vs 57 formed) and the upper flap still tilted downwards. They rank among the most open in PDVA but not very high in SASA, due to the upper flap and to sidechains orientation (especially in m1:11). State m1:4 is among the only three states in which the contact of the dimer interface (cont. 285 vs 285*) is a little looser than in the others.

The remaining states do not present close similarities to others in terms of contact structure; we describe them in approximate order of decreasing openness of the catalytic pocket. The most open state according to both PDVA and SASA is m2:4; its upper flap is not tilted downwards and is retracted from the pocket, distancing from the β-sheet M162-G170 loop (we call it *right loop*), leaving cont. 138 vs 172 not formed; the left flap is very open (although the dihedrals of this loop are quite variable among the configurations of such state); the linker loop is slightly contracted and wide (cont. 131 vs 199 and 132 vs 196 not formed), not stretching towards the left flap as in other closed or partly-closed states; all of the above play to leave the catalytic dyad well exposed.

State m1:8 also ranks very high in PDVA and in SASA, despite the upper flap tilted downwards. The left flap is very open, although dihedrals 43-46 are not all in α configuration; their particular arrangement ($\alpha\beta\alpha c$), however, grants that the biggest sidechains of the left flap are not oriented towards the binding pocket. The linker loop is not strerched towards the left flap, but rather down, towards the interface with the solvent; it is quite open (dihedral 189 in c instead of β configuration) in proximity of the pocket and all its sidechains do not obstruct the access to the cavity (in particular those of Arg^{188} and Gln^{189}, responsible for a low SASA in other states).

State m1:5 is characterised by an having the left flap open (although less than e.g. state m1:4

and m1:11), with cont. 47 vs 57 formed, and the upper loop not tilted. The right loop leans slightly towards the tip of linker loop (Arg^{188}), causing cont. 138 vs 172 to be broken and cont. 167 vs 188 to be formed between the sidechain of Leu^{167} and the backbone of Arg^{188}. All other contacts far from the pocket are formed. The linker loop leans towards the left flap rather than down.

In state m1:3 the position of the upper flap and of the right loop are approximately the same as in m1:5. The linker loop stretches a bit more toward the left flap, causing contacts 132 vs 196 and 197-198 vs 238 to be broken. The left flap is closed, forming contact 49 vs 189 with the linker loop. The lower part of the pocket results closed, but the catalytic dyad is left quite exposed from above, which yields a central position in both SASA and PDVA ranks.

Also state m1:6 leaves the pocket quite accessible from the top and covered from the bottom. The linker loop is quite open, while the left flap is closed and stretched towards it. The peculiar shape of the left flap brings the α-carbons of Ser^{46} and Arg^{188} very clpse together, which results in a very low PDVA (second lowest in the ranking).

State m2:3 ranks as the third lowest in both SASA and PDVA. Cys^{145} is not well covered, but on the other hand His^{41} is less accessible than in most other states. As most m2 states, m2:3 has the upper flap flat and cont. 138 vs 172 not formed. The linker loop is not stretched, leaving the contacts with residue Arg^{131} unformed or partly unformed. The left flap is really closed and stretched towards the linker loop and its dihedrals are arranged in such a way that cont. 49 vs 189 is not formed; however, these two most mobile loops have a contact between Glu^{47} and Gln^{189}.

Finally, state m2:5 is the one with the lowest PDVA and is among the lowest-ranked in SASA. Its conformation is quite peculiar: the linker loop is all retracted and coiled (it is the only state of m2 forming cont. 167 vs 188). The left flap is all stretched towards the linker loop (cont. 49 vs 189 formed), which, with the contribution of the sidechains, almost completely covers the catalytic His^{41}. The upper flap, rather than being flat or tilted down, is oriented upwards, causing a deformation in the II domain which allows cont. 138 vs 172 to be formed. Remarkably, m2:5 is one of the three states with cont. 285 vs 285* not tightly formed.

Appendix D

Gradient of the free energy

D.1 Useful n-dimensional integrals

D.1.1 Vector identities

Vector scaled by scalar product of two vectors:

$$\mathbf{a}(\mathbf{b}\cdot\mathbf{c}) = (\mathbf{a}\cdot\mathbf{b})\mathbf{c} \tag{D.1}$$

Outer and scalar products:

$$(\mathbf{a}\otimes\mathbf{b})\cdot\mathbf{c} = (\mathbf{a}\cdot\mathbf{c})\mathbf{b} \overset{(D.1)}{=} (\mathbf{c}\cdot\mathbf{b})\mathbf{a} \tag{D.2}$$

D.1.2 Hyperspherical coordinates and the volume of the unit-radius n-sphere

In order to perform integrals on n-dimensional spheres it is useful to project cartesian coordinates on hyperspherical ones:

$$\begin{aligned}
x_1 &= r\cos(\varphi_1)\\
x_2 &= r\sin(\varphi_1)\cos(\varphi_2)\\
x_3 &= r\sin(\varphi_1)\sin(\varphi_2)\cos(\varphi_3)\\
&\vdots\\
x_{n-1} &= r\sin(\varphi_1)\cdots\sin(\varphi_{n-2})\cos(\varphi_{n-1})\\
x_n &= r\sin(\varphi_1)\cdots\sin(\varphi_{n-2})\sin(\varphi_{n-1})\ .
\end{aligned}$$

With this change of coordinates the infinitesimal volume element $\mathrm{d}^n\mathbf{x} = \mathrm{d}x_1\cdots\mathrm{d}x_n$ becomes:

$$\mathrm{d}V_n = \left|\det \frac{\partial(x_i)}{\partial\,(r, \varphi_j)}\right| \mathrm{d}r\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2 \cdots \mathrm{d}\varphi_{n-1} = r^{n-1}\sin^{n-2}(\varphi_1)\sin^{n-3}(\varphi_2)\cdots\sin(\varphi_{n-2})\,\mathrm{d}r\,\mathrm{d}\varphi_1\,\mathrm{d}\varphi_2\cdots\mathrm{d}\varphi_{n-1}$$

We express the volume of an n-dimensional sphere $B^n(R)$ as $V_n = \omega_n R^n$. Therefore, its surface is $S_n = \partial_R V_n = n\omega_n R^{n-1}$. The quantity ω_n is the volume of the n-sphere of unitary radius, whose expression can be derived by computing $\omega_n := \int_{B^n(R)} 1\,\mathrm{d}\mathbf{x}$, which gives:

$$\omega_n = \frac{2}{n}\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \tag{D.3}$$

D.1.3 Mean square radius

Let us then go ahead and compute the mean square displacement in a n-dimensional ball $B^n(R)$ of volume V_n and radius R:

$$\begin{aligned} V_n\,\langle\mathbf{x}^2\rangle_{B^n} &= \int_{B^n} \mathbf{x}^2\,\mathrm{d}^n\mathbf{x} = \int_{B^n} r^2\,\mathrm{d}V_n = \int_0^R r^2\,S_n(r)\,\mathrm{d}r = \int_0^R r^2\,n\,\omega_n\,r^{n-1}\,\mathrm{d}r = n\,\omega_n\,\frac{R^{n+2}}{n+2} \\ &= R^2\,\frac{n}{n+2}\,V_n \end{aligned} \tag{D.4}$$

D.1.4 Mean outer product of the displacement

It is required in many of the above equations to be able to compute the average outer product of the displacement in a n-dimensional ball $B^n(R)$ of volume V_n and radius R:

$$\begin{aligned} V_n\,\langle\mathbf{x}\mathbf{x}^{\mathrm{T}}\rangle_{B^n} &= \int_{B^n} \mathbf{x}\mathbf{x}^{\mathrm{T}}\,\mathrm{d}^n\mathbf{x} \\ &= \int_{B^n} \sum_{i,j}(x_i x_j \hat{\mathbf{e}}_i\hat{\mathbf{e}}_j^{\mathrm{T}})\,\mathrm{d}^n\mathbf{x} \quad \text{where } \{\hat{\mathbf{e}}_i\}_i \text{ are standard basis elements} \\ &= \int_{B^n} \sum_{i}(x_i^2 \hat{\mathbf{e}}_i\hat{\mathbf{e}}_i^{\mathrm{T}})\,\mathrm{d}^n\mathbf{x} \quad \text{all off-diagonal terms vanish for integration of odd function on even domain} \\ &= \mathbb{1}_n\,I_n(R)\,V_n \end{aligned}$$

since all directions are equal for symmetry reasons, the diagonal matrix in the second-last line must be proportional to the n-dimensional identity matrix $\mathbb{1}_n$ and we call the scaling factor $I_n(R)$. Now notice that Eq. (D.4) is the trace of Eq. eq. (D.5) and so:

$$R^2\,\frac{n}{n+2} = \langle\mathbf{x}^2\rangle_{B^n} = \langle\mathrm{Tr}(\mathbf{x}\mathbf{x}^{\mathrm{T}})\rangle_{B^n} = \mathrm{Tr}(\langle\mathbf{x}\mathbf{x}^{\mathrm{T}}\rangle_{B^n}) = \mathrm{Tr}(\mathbb{1}_n)\,I_n(R) = n\,I_n(R) \quad \Rightarrow \quad I_n(R) = \frac{R^2}{n+2}$$

and therefore:

$$\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \rangle_{B^n} = \mathbb{1}_n \, \frac{R^2}{n+2} \tag{D.5}$$

D.2 Estimating free energy derivatives

D.2.1 Mean shift

D.2.1.1 Analytical expression for the mean shift

First of all, let us consider the Taylor expansion of a density $\rho(\mathbf{x})$ around point $\mathbf{x}_i$:

$$\rho(\mathbf{x}) = \rho(\mathbf{x}) = \rho(\mathbf{x}_i) \; + \; \nabla_{\mathbf{x}}^{\mathrm{T}} \rho(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i) \; + \; \frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^{\mathrm{T}} \, \nabla_{\mathbf{x}}^2 \, \rho(\mathbf{x}_i)(\mathbf{x} - \mathbf{x}_i) + \mathcal{O}\left((\mathbf{x} - \mathbf{x}_i)^3\right) \; . \tag{D.6}$$

The mean shift around point $\mathbf{x}_i$ within region $\Omega_i := B^d(r_i, \mathbf{x}_i)$ of volume V_d is defined as:

$$\langle (\mathbf{x} - \mathbf{x}_i) \rangle_{\Omega_i, \rho} := \frac{\int_{\Omega_i} \rho(\mathbf{x})(\mathbf{x} - \mathbf{x}_i) \, \mathrm{d}\mathbf{x}}{\int_{\Omega_i} \rho(\mathbf{x}) \, \mathrm{d}\mathbf{x}} \; . \tag{D.7}$$

For a lighter notation we choose the specific case $\mathbf{x}_i = \mathbf{0}$, but the derivation remains valid also in the more general case. Inserting the expansion (6.3) into equation (6.4) and taking into account the results (D.3), (D.4) and (D.5):

$$\begin{aligned}
\langle (\mathbf{x} - \mathbf{x}_i) \rangle_{\Omega_i, \rho} &= \frac{\int_{\Omega_i} \rho(\mathbf{x}) \, \mathbf{x} \, \mathrm{d}\mathbf{x}}{\int_{\Omega_i} \rho(\mathbf{x}) \, \mathrm{d}\mathbf{x}} \\
&= \frac{\rho(\mathbf{x}_i) \cancelto{0}{\int_{\Omega_i} \mathbf{x} \, \mathrm{d}\mathbf{x}} + \; \nabla_{\mathbf{x}}^{\mathrm{T}} \rho(\mathbf{x}_i) \int_{\Omega_i} \mathbf{x} \, \mathbf{x}^{\mathrm{T}} \, \mathrm{d}\mathbf{x} \; + \; \frac{1}{2} \nabla_{\mathbf{x}}^2 \, \rho(\mathbf{x}_i) \cancelto{0}{\int_{\Omega_i} \mathbf{x} \, \mathbf{x}^{\mathrm{T}} \mathbf{x} \, \mathrm{d}\mathbf{x}}}{\rho(\mathbf{x}_i) \int_{\Omega_i} 1 \, \mathrm{d}\mathbf{x} \; + \; \nabla_{\mathbf{x}}^{\mathrm{T}} \rho(\mathbf{x}_i) \cancelto{0}{\int_{\Omega_i} \mathbf{x} \, \mathrm{d}\mathbf{x}} + \; \frac{1}{2} \; \mathrm{Tr} \left[\nabla_{\mathbf{x}}^2 \, \rho(\mathbf{x}_i) \int_{\Omega_i} \mathbf{x} \, \mathbf{x}^{\mathrm{T}} \mathrm{d}\mathbf{x} \right]} \; + \; \mathcal{O}(V_d \, r_i^4) \\
&= \frac{\nabla_{\mathbf{x}} \rho(\mathbf{x}_i) \cancel{V_d} \, \frac{r_i^2}{d+2}}{\rho(\mathbf{x}_i) \cancel{V_d} \; + \; \frac{1}{2} \; \mathrm{Tr} \, \nabla_{\mathbf{x}}^2 \, \rho(\mathbf{x}_i) \cancel{V_d} \, \frac{r_i^2}{d+2}} \; + \; \mathcal{O}(\cancel{V_d} \, r_i^4) \\
&= \frac{\nabla_{\mathbf{x}} \rho(\mathbf{x}_i) \, \frac{r_i^2}{d+2}}{\rho(\mathbf{x}_i) \left(1 \; + \; \frac{\mathrm{Tr} \, \nabla_{\mathbf{x}}^2 \, \rho(\mathbf{x}_i)}{2 \, \rho(\mathbf{x}_i)} \, \frac{r_i^2}{d+2} \right)} \; + \; \mathcal{O}(r_i^4) \\
&= \frac{r_i^2}{d+2} \frac{\nabla_{\mathbf{x}} \rho(\mathbf{x}_i)}{\rho(\mathbf{x}_i)} \left(1 - \frac{\mathrm{Tr} \nabla_{\mathbf{x}}^2 \rho(\mathbf{x}_{\mathrm{i}})}{2 \rho(\mathbf{x}_i)} \frac{r_i^2}{d+2} \right) \; + \; \mathcal{O}(r_i^4)
\end{aligned} \tag{D.8}$$

where the neglected integrals vanish for integration of an odd function on a symmetric domain.

D.2.1.2 Operational definition of sample mean shift

Now, let us focus on the estimation of the mean shift on the left-hand side of the equation from a sample. As discussed in section 6.1, we compute it as the sample average of the shift observable $(\mathbf{x}-\mathbf{x}_i)$ over the fist $\hat{k}_i - 1$ NNs of $\mathbf{x}_i$; mathematically, we use the sample density estimator $\hat{\rho}_s$ in equation (2.15) combined with a restriction on the region $\Omega_i = B^d(r_{\hat{k}_i,i}, \mathbf{x})$, so that:

$$\begin{aligned}
\langle(\mathbf{x}-\mathbf{x}_i)\rangle_{B^d(r_{\hat{k}_i},\mathbf{x}_i),\hat{\rho}_s} &= \frac{\int_{B^d(r_{\hat{k}_i},\mathbf{x}_i)} \hat{\rho}_s(\mathbf{x})(\mathbf{x}-\mathbf{x}_i)\,\mathrm{d}\mathbf{x}}{\int_{B^d(r_{\hat{k}_i},\mathbf{x}_i)} \hat{\rho}_s(\mathbf{x})\,\mathrm{d}\mathbf{x}} \\
&= \frac{\frac{1}{N}\sum_{j=1}^{N}\int_{B^d(r_{\hat{k}_i},\mathbf{x}_i)} \delta(\mathbf{x}_j-\mathbf{x})\,(\mathbf{x}-\mathbf{x}_i)\,\mathrm{d}\mathbf{x}}{\hat{k}_i/N} \\
&= \frac{1}{\hat{k}_i}\sum_{j=1}^{N}\int I_{B^d(r_{\hat{k}_i},\mathbf{x}_i)}\,\delta(\mathbf{x}_j-\mathbf{x})\,(\mathbf{x}-\mathbf{x}_i)\,\mathrm{d}\mathbf{x} \\
&= \frac{1}{\hat{k}_i}\sum_{j=1}^{N} I_{B^d(r_{\hat{k}_i},\mathbf{x}_i)}\,(\mathbf{x}_j-\mathbf{x}_i) \\
&= \frac{1}{\hat{k}_i}\sum_{j=1}^{\hat{k}_i-1}(\mathbf{x}-\mathbf{x}_i)
\end{aligned} \tag{D.9}$$

where $I_{B^d(r_{\hat{k}_i},\mathbf{x}_i)}$ is the indicator function of the selected neighbourhood of point i and can be rewritten in terms of product of Heaviside theta distributions and of boxcar functions as:

$$I_{B^d(r_{\hat{k}_i},\mathbf{x}_i)} = \prod_{\alpha=1}^{d} \mathrm{Box}_{[-r_{\hat{k}_i},r_{\hat{k}_i}]}(x_\alpha - x_{i,\alpha}) = \prod_{\alpha=1}^{d}\left[\Theta(x_\alpha - x_{i,\alpha} + r_{\hat{k}_i}) - \Theta(x_\alpha - x_{i,\alpha} - r_{\hat{k}_i})\right] \ . \tag{D.10}$$

The above derivation justifies equation (6.7); it corresponds to using as density estimator for $\rho(\mathbf{x})$ over the whole region Ω_i the value $\hat{\rho}_i$ estimated with $\hat{k}$NN.

D.2.1.3 Proof that the mean shift estimates the average free energy gradient

Let us consider the fact that $\nabla_{\mathbf{x}}F(\mathbf{x}_i) = -\frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}_i)}{\rho(\mathbf{x}_i)}$, having dropped the Boltzmann factor and, starting from expression (6.3) let us give an expression for the density gradient:

$$\nabla_{\mathbf{x}}\rho(\mathbf{x}_i) = \nabla_{\mathbf{x}}\rho(\mathbf{x}_i) + \frac{1}{2}\nabla^2_{\mathbf{x}}\,\rho(\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i) + \frac{1}{6}(\mathbf{x}-\mathbf{x}_i)^{\mathrm{T}}\,\nabla^3_{\mathbf{x}}\,\rho(\mathbf{x}_i)(\mathbf{x}-\mathbf{x}_i) + \mathcal{O}\left((\mathbf{x}-\mathbf{x}_i)^3\right) \ . \tag{D.11}$$

Therefore we can write the mean value of the free energy gradient over Ω_i, considering again without loss of generality the specific case $\mathbf{x}_i = \mathbf{0}$:

$$
\begin{aligned}
\langle \nabla_{\mathbf{x}} F(\mathbf{x}) \rangle_{\Omega_i} &= \frac{\int_{\Omega_i} \rho(\mathbf{x})\, \nabla_{\mathbf{x}} F(\mathbf{x})\, \mathrm{d}\mathbf{x}}{\int_{\Omega_i} \rho(\mathbf{x})\, \mathrm{d}\mathbf{x}} \\
&= \frac{\int_{\Omega_i} \cancel{\rho(\mathbf{x})}\, \nabla_{\mathbf{x}} \rho(\mathbf{x}) / \cancel{\rho(\mathbf{x})}\, \mathrm{d}\mathbf{x}}{\int_{\Omega_i} \rho(\mathbf{x})\, \mathrm{d}\mathbf{x}} \\
&= \frac{\nabla_{\mathbf{x}} \rho(\mathbf{x}_i) \int_{\Omega_i} 1\, \mathrm{d}\mathbf{x} \;+\; \frac{1}{2}\, \nabla^2_{\mathbf{x}}\, \rho(\mathbf{x}_i) \cancelto{0}{\int_{\Omega_i} \mathbf{x}\, \mathrm{d}\mathbf{x}} +\; \nabla^3_{\mathbf{x}}\, \rho(\mathbf{x}_i) \int_{\Omega_i} \mathbf{x}\, \mathbf{x}^{\mathrm{T}} \mathrm{d}\mathbf{x}}{\rho(\mathbf{x}_i) \int_{\Omega_i} 1\, \mathrm{d}\mathbf{x} \;+\; \nabla^{\mathrm{T}}_{\mathbf{x}} \rho(\mathbf{x}_i) \cancelto{0}{\int_{\Omega_i} \mathbf{x}\, \mathrm{d}\mathbf{x}} +\; \frac{1}{2}\, \mathrm{Tr} \left[\nabla^2_{\mathbf{x}}\, \rho(\mathbf{x}_i) \int_{\Omega_i} \mathbf{x}\, \mathbf{x}^{\mathrm{T}} \mathrm{d}\mathbf{x} \right]} \;+\; \mathcal{O}(V_d\, r_i^4) \\
&= \frac{\nabla_{\mathbf{x}} \rho(\mathbf{x}_i) \cancel{V_d} \;+\; \frac{1}{6}\, \nabla^3_{\mathbf{x}}\, \rho(\mathbf{x}_i) \cdot \mathbb{1}_D \cancel{V_d}\, \frac{r_i^2}{d+2}}{\rho(\mathbf{x}_i) \cancel{V_d} \;+\; \frac{1}{2}\, \mathrm{Tr}\, \nabla^2_{\mathbf{x}}\, \rho(\mathbf{x}_i) \cancel{V_d}\, \frac{r_i^2}{d+2}} \;+\; \mathcal{O}(\cancel{V_d}\, r_i^4) \\
&= \frac{\nabla_{\mathbf{x}} \rho(\mathbf{x}_i) \;+\; \frac{1}{6}\, \nabla^3_{\mathbf{x}}\, \rho(\mathbf{x}_i) \cdot \mathbb{1}_D\, \frac{r_i^2}{d+2}}{\rho(\mathbf{x}_i) \left(1 \;+\; \frac{\mathrm{Tr}\, \nabla^2_{\mathbf{x}}\, \rho(\mathbf{x}_i)}{2\, \rho(\mathbf{x}_i)}\, \frac{r_i^2}{d+2} \right)} \;+\; \mathcal{O}(r_i^4) \\
&= \frac{\nabla_{\mathbf{x}} \rho(\mathbf{x}_i)}{\rho(\mathbf{x}_i)} \left(1 - \frac{\mathrm{Tr} \nabla^2_{\mathbf{x}} \rho(\mathbf{x}_\mathrm{i})}{2\rho(\mathbf{x}_i)} \frac{r_i^2}{d+2} \right) \;+\; \frac{1}{6}\, \nabla^3_{\mathbf{x}}\, \rho(\mathbf{x}_i) \cdot \mathbb{1}_D\, \frac{r_i^2}{d+2} \;+\; \mathcal{O}(r_i^4) \\
&= \hat{\mathbf{g}}_i \;+\; \mathcal{O} \left(\frac{1}{6}\, \nabla^3_{\mathbf{x}}\, \rho(\mathbf{x}_i) \cdot \mathbb{1}_D\, \frac{r_i^2}{d+2} \right)
\end{aligned}
\tag{D.12}
$$

where $\hat{\mathbf{g}}_i$ is the estimator defined in equation (6.8). Therefore $\hat{\mathbf{g}}_i$ is an estimator of the average quantity $\langle \nabla_{\mathbf{x}} F(\mathbf{x}_i) \rangle_{\Omega_i}$ over the region Ω_i, up to cubic order in the Taylor expansion of $\rho(\mathbf{x})$.

D.2.2 Correlation structure of the free energy gradients

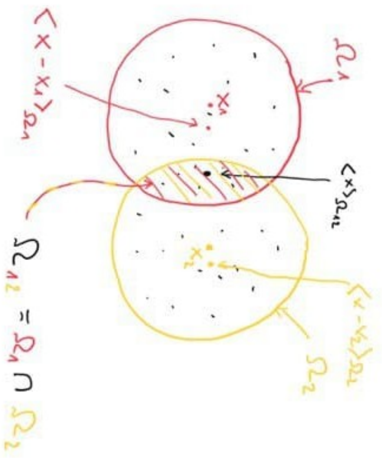

Ω_1
$\langle x \rangle_{\Omega_{12}}$
$\langle x - x_2 \rangle_{\Omega_2}$
Ω_2
x_1
x_2
$\langle x - x_1 \rangle_{\Omega_1}$
$\Omega_{12} = \Omega_1 \cap \Omega_2$

Figure D.1: Representation of neighbourhoods defining local estimates of the gradients and their correlation structures.

Let us consider two points $\mathbf{x}_1$ and $\mathbf{x}_2$ for which we have selected the optimal number of neighbours

k_1 and k_2 which define the hyperspherical neighbourhoods $\Omega_1 = B^d(r_{k_1}, \mathbf{x}_1)$ and $\Omega_2 = B^d(r_{k_2}, \mathbf{x}_2)$. Ignoring prefactors, the true free energy gradients $\mathbf{g}_1$ and $\mathbf{g}_2$ in the two points are estimated, according to equation (6.8), as:

$$\begin{aligned}
\hat{\mathbf{g}}_1 &= \frac{1}{k_1}\sum_{j=1}^{k_1}(\mathbf{x}_j - \mathbf{x}_1) = \frac{1}{k_1}\sum_{j=1}^{N} I_{\Omega_1}(\mathbf{x}_j - \mathbf{x}_1) \\
\hat{\mathbf{g}}_2 &= \frac{1}{k_2}\sum_{j=1}^{k_2}(\mathbf{x}_j - \mathbf{x}_2) = \frac{1}{k_1}\sum_{j=1}^{N} I_{\Omega_2}(\mathbf{x}_j - \mathbf{x}_2)
\end{aligned} \tag{D.13}$$

where the rightmost sums run over all N points thanks to the use of the indicator functions; when $\mathbf{x}_1$ and $\mathbf{x}_2$ are sample points then the leftmost sums run from $j = 0$ to $k_1 - 1$ and $k_2 - 1$ respectively. Of course, since we use unbiased estimators, we have $\langle \hat{\mathbf{g}}_1 \rangle = \mathbf{g}_1 = \langle (\mathbf{x} - \mathbf{x}_1) \rangle_{\Omega_1}$ and $\langle \hat{\mathbf{g}}_2 \rangle = \mathbf{g}_2 = \langle (\mathbf{x} - \mathbf{x}_2) \rangle_{\Omega_2}$.

D.2.2.1 Variance-covariance matrix of the free energy gradients

Let us first consider the auto-covariance matrix of the gradient estimator $\hat{\mathbf{g}}_1$ at a point $\mathbf{x}_1$:

$$\begin{aligned}
\boldsymbol{\sigma}_1^2 = \mathbf{cov}[\hat{\mathbf{g}}_1, \hat{\mathbf{g}}_1] :&= \langle \hat{\mathbf{g}}_1 \hat{\mathbf{g}}_1^{\mathrm{T}} \rangle - \langle \hat{\mathbf{g}}_1 \rangle \langle \hat{\mathbf{g}}_1^{\mathrm{T}} \rangle \\
&= \left\langle \left[\frac{1}{k_1}\sum_{i=1}^{N} I_{\Omega_1}(\mathbf{x}_i - \mathbf{x}_1)\right]\left[\frac{1}{k_1}\sum_{j=1}^{N} I_{\Omega_1}(\mathbf{x}_j - \mathbf{x}_1)^{\mathrm{T}}\right] \right\rangle - \mathbf{g}_1\mathbf{g}_1^{\mathrm{T}} \\
&= \left\langle \left[\frac{1}{k_1}\sum_{i=1}^{k_1}(\mathbf{x}_i - \mathbf{x}_1)\right]\left[\frac{1}{k_1}\sum_{j=1}^{k_1}(\mathbf{x}_j - \mathbf{x}_1)^{\mathrm{T}}\right] \right\rangle_{\Omega_1} - \mathbf{g}_1\mathbf{g}_1^{\mathrm{T}} \\
&= \frac{1}{k_1^2}\sum_{i,j}^{k_1} \left\langle (\mathbf{x}_i - \mathbf{x}_1)(\mathbf{x}_j - \mathbf{x}_1)^{\mathrm{T}} \right\rangle_{\Omega_1} - \mathbf{g}_1\mathbf{g}_1^{\mathrm{T}} \\
&= \frac{1}{k_1^2}\left[\sum_{i}^{k_1} \left\langle (\mathbf{x}_i - \mathbf{x}_1)(\mathbf{x}_i - \mathbf{x}_1)^{\mathrm{T}} \right\rangle_{\Omega_1} + \sum_{i \neq j} \left\langle (\mathbf{x}_i - \mathbf{x}_1)(\mathbf{x}_j - \mathbf{x}_1)^{\mathrm{T}} \right\rangle_{\Omega_1}\right] - \mathbf{g}_1\mathbf{g}_1^{\mathrm{T}} \\
&= \frac{1}{k_1^{\cancel{2}}}\left[\cancel{k_1}\langle (\mathbf{x} - \mathbf{x}_1)(\mathbf{x} - \mathbf{x}_1)^{\mathrm{T}} \rangle_{\Omega_1} + \cancel{k_1}(k_1 - 1)\langle (\mathbf{x} - \mathbf{x}_1) \rangle_{\Omega_1} \langle (\mathbf{x} - \mathbf{x}_1)^{\mathrm{T}} \rangle_{\Omega_1}\right] - \mathbf{g}_1\mathbf{g}_1^{\mathrm{T}} \\
&= \frac{1}{k_1}\langle (\mathbf{x} - \mathbf{x}_1)(\mathbf{x} - \mathbf{x}_1)^{\mathrm{T}} \rangle_{\Omega_1} + \frac{k_1 - 1}{k_1}\langle \hat{\mathbf{g}}_1 \rangle \langle \hat{\mathbf{g}}_1^{\mathrm{T}} \rangle - \mathbf{g}_1\mathbf{g}_1^{\mathrm{T}} \\
&= \frac{1}{k_1}\left[\langle (\mathbf{x} - \mathbf{x}_1)(\mathbf{x} - \mathbf{x}_1)^{\mathrm{T}} \rangle_{\Omega_1} - \mathbf{g}_1\mathbf{g}_1^{\mathrm{T}}\right] \\
&= \frac{1}{k_1}\left[\langle (\mathbf{x} - \mathbf{x}_1)(\mathbf{x} - \mathbf{x}_1)^{\mathrm{T}} \rangle_{\Omega_1} - \langle (\mathbf{x} - \mathbf{x}_1) \rangle_{\Omega_1} \langle (\mathbf{x} - \mathbf{x}_1)^{\mathrm{T}} \rangle_{\Omega_1}\right] \\
&= \frac{1}{k_1}\mathbf{cov}[(\mathbf{x} - \mathbf{x}_1), (\mathbf{x} - \mathbf{x}_1)]_{\Omega_1}
\end{aligned} \tag{D.14}$$

Notice that, regarding the $\{\hat{\mathbf{g}}_i\}_i$ as RVs, we observe only a single realisation for each i in one sample; for a generic RV this would make it impossible to estimate their variances unless assumptions on their distribution were made

However sum of RVs –> teo lim centr

Operatively, $\boldsymbol{\sigma}_1^2$ can be estimated from the sample substituting the mean values in the equations with sample averages over the k_1 points in Ω_1. In particular, $\langle(\mathbf{x}-\mathbf{x}_1)\rangle_{\Omega_1}$ is estimated by $\hat{\mathbf{g}}_1$ in equation (6.8), while $\langle(\mathbf{x}-\mathbf{x}_1)(\mathbf{x}-\mathbf{x}_1)^{\mathrm{T}}\rangle_{\Omega_1}$ is estimated form neighbours $\mathbf{x}_j$ of $\mathbf{x}_1$ as $\frac{1}{k_1}\sum_{j=1}^{k_1}(\mathbf{x}_j - \mathbf{x}_1)(\mathbf{x}_j-\mathbf{x}_1)^{\mathrm{T}}$. Thus, for a given point $\mathbf{x}_i$ the variance-covariance matrix $\boldsymbol{\sigma}_i^2$ of the free energy gradient is computed from the sample (reintroducing prefactors and notation so far neglected) as:

$$\hat{\boldsymbol{\sigma}}_i^2 := \left(k_B T\,\frac{d+2}{r_i^2}\right)^2 \frac{1}{\hat{k}_i - 1}\left[\sum_{j=1}^{\hat{k}_i}\frac{1}{\hat{k}_i}(\mathbf{x}_j-\mathbf{x}_i)(\mathbf{x}_j-\mathbf{x}_i)^{\mathrm{T}} - \hat{\mathbf{g}}_i\hat{\mathbf{g}}_i^{\mathrm{T}}\right] \tag{D.15}$$

where, again, the prefactor $1/(\hat{k}_i-1)$ comes from the Bessel's correction for the unbiased sample variance estimator[193].

D.2.2.2 Cross-covariance matrix of the free energy gradients

We can write

$$\begin{aligned}\langle\hat{g}_1\hat{g}_2\rangle &= \frac{1}{k_1k_2}\sum_{i,j}^{N,N}\langle H_1(\mathbf{x}_i-\mathbf{x}_1)H_2(\mathbf{x}_j-\mathbf{x}_2)\rangle \\ &= \frac{1}{k_1k_2}\sum_{i}^{N}\langle H_1(\mathbf{x}_i-\mathbf{x}_1)H_2(\mathbf{x}_i-\mathbf{x}_2)\rangle \\ &= \frac{k_{12}}{k_1k_2}\langle(\mathbf{x}-\mathbf{x}_1)(\mathbf{x}-\mathbf{x}_2)\rangle_{\Omega_{12}}\end{aligned} \tag{D.16}$$

where from the first to the second line we have used the fact that the $\mathbf{x}_i$ are *independent* and from the second to the third we have used the fact that they are *identically distributed* along with the properties of the cutoff functions.

$$\begin{aligned}\langle\hat{g}_1\hat{g}_2\rangle &= \frac{1}{k_1k_2}\sum_{i,j}^{N,N}\langle H_1(\mathbf{x}_i-\mathbf{x}_1)H_2(\mathbf{x}_j-\mathbf{x}_2)\rangle \\ &= \frac{1}{k_1k_2}\left[\sum_{i}^{N}\langle H_1(\mathbf{x}_i-\mathbf{x}_1)H_2(\mathbf{x}_i-\mathbf{x}_2)\rangle + \sum_{i\neq j}^{N(N-1)}\langle H_1(\mathbf{x}_i-\mathbf{x}_1)\rangle\langle H_2(\mathbf{x}_j-\mathbf{x}_2)\rangle\right] \\ &= \frac{1}{k_1k_2}\left[k_{12}\langle(\mathbf{x}-\mathbf{x}_1)(\mathbf{x}-\mathbf{x}_2)\rangle_{\Omega_{12}} + (k_1k_2-k_{12})\langle(\mathbf{x}-\mathbf{x}_1)\rangle_{\Omega_1}\langle(\mathbf{x}-\mathbf{x}_2)\rangle_{\Omega_2}\right]\end{aligned} \tag{D.17}$$

D.2.2.3 Possible estimators for the Pearson correlation coefficient entering ε_{ij}

D.2.2.3.1 Intersection over union we assume it to be well approximated by the ratio between the volume of their intersection $\Omega_i \cap \Omega_j$ and their union $\Omega_i \cup \Omega_j$. Moreover, since $\hat{k}_i$ is proportional to the d-volume $V_i = \omega_d r_i^d$ of Ω_i via the relation $\hat{k}_i = V_i\, \hat{\rho}_i^{\hat{k}\mathrm{NN}}$, we define a proxy for p_{ij} as:

$$\hat{p}_{ij} = \frac{\hat{k}_{ij}}{\hat{k}_i + \hat{k}_j - \hat{k}_{ij}} \;, \tag{D.18}$$

D.2.2.3.2 Consider also $\hat{p}_{ij} = \frac{\hat{k}_{ij}}{\sqrt{\hat{k}_i\, \hat{k}_j}}$

D.2.2.3.3 Consider also $\hat{p}_{ij} = \frac{\hat{k}_{ij}}{\hat{k}_i\, \hat{k}_j}$, which goes from 0 to $1/\max\{k_1, k_2\}$

D.2.3 An expression for $\hat{\delta F}^{\,i}_{ij}$ in terms of distances between points

By using expression (6.8) for the free energy gradient sample estimator $\hat{\mathbf{g}}_i$ and the definition of the vector difference between two points i and j, $\mathbf{r}_{ij} := \mathbf{x}_j - \mathbf{x}_i$, we can give an expression of $\hat{\delta F}^{\,i}_{ij}$ in equation (6.14) solely in terms of distances between points:

$$\hat{\delta F}^{\,i}_{ij} \;=\; -k_B T\, \frac{d+2}{r^2}\, \frac{1}{\hat{k}_i} \sum_{l=1}^{\hat{k}_i} \frac{\mathbf{r}_{ij}^2 + \mathbf{r}_{il}^2 - \mathbf{r}_{jl}^2}{2} \tag{D.19}$$

D.3 Support to the discussion about BMTI

D.3.1 Solution of the BMTI for uncorrelated $\hat{\delta F}$s

Let us consider the log-likelihood model defining BMTI in equation (6.26), that we repeat here for the reader's convenience:

$$\mathcal{L}(\mathbf{F} \mid \hat{\boldsymbol{\delta F}}\,, \mathbf{D}) \;:=\; -\sum_{i=1}^{N} \sum_{j \in \Omega_i} \frac{(F_j - F_i - \hat{\delta F}_{ij})^2}{2\varepsilon_{ij}^2} \;.$$

We maximise it analytically with respect to the vector $\mathbf{F}$ by setting its gradient to zero:

$$\begin{aligned} 0 = \frac{\partial}{\partial F_i} \mathcal{L}(\mathbf{F} \mid \hat{\boldsymbol{\delta F}}\,, \mathbf{D}) &= -\sum_{k} \sum_{j \in \Omega_k} \frac{1}{\sigma_{kj}} (F_j - F_k - \delta F_{kj})(\delta_{ji} - \delta_{ki}) \\ &= -\sum_{j \mid i \in \Omega_j} \frac{1}{\sigma_{ji}} (F_i - F_j - \delta F_{ji}) + \sum_{j \in \Omega_i} \frac{1}{\sigma_{ij}} (F_j - F_i - \delta F_{ij}) \;, \end{aligned}$$

where the notation $j \mid i \in \Omega_j$ indicates the set of points j which include the point i in their neighbourhood Ω_j. Working out equation the above calculation further, we can bring all the maximisation parameters $\{F_i\}_i$ on the left-hand side of the equal sign:

$$\sum_{j|i\in\Omega j}\frac{1}{\sigma_{ji}^2}\delta F_{ji}-\sum_{j\in\Omega_i}\frac{1}{\sigma_{ij}^2}\delta F_{ij}=\sum_{j|i\in\Omega j}\frac{1}{\sigma_{ji}^2}(F_i-F_j)-\sum_{j\in\Omega_i}\frac{1}{\sigma_{ij}^2}(F_j-F_i)\ .$$

Distinguishing on the right-hand side of the expression the sums in which F's are summed over from those those who can be factored out and defining

$$\Delta F_i := \sum_{j|i\in\Omega j}\frac{1}{\sigma_{ji}^2}\delta F_{ji}-\sum_{j\in\Omega_i}\frac{1}{\sigma_{ij}^2}\delta F_{ij} \tag{D.20}$$

we obtain:

$$\Delta F_i=\left(\sum_{j|i\in\Omega j}\frac{1}{\sigma_{ji}^2}+\sum_{j\in\Omega i}\frac{1}{\sigma_{ij}^2}\right)F_i-\sum_{j|i\in\Omega j}\frac{1}{\sigma_{ji}^2}F_j-\sum_{j\in\Omega i}\frac{1}{\sigma_{ij}^2}F_j\ . \tag{D.21}$$

This linear system can be written in vectorial form as follows. First, for notational convenience, we rename the two sums in parenthesis as $S_{\to i}=\sum_{j|i\in\Omega j}\frac{1}{\sigma_{ji}^2}$ and $S_{i\to}=\sum_{j\in\Omega i}\frac{1}{\sigma_{ij}^2}$ with the two arrow symbols respectively indicating the points for which i is a neighbour ($\to i$) and the points in the neighbourhood of point i ($i\to$). Then we rewrite the last line using indicator functions for the set of points:

$$\begin{aligned}\Delta F_i &= (S_{\to i}+S_{i\to})\,F_i-\sum_{j|i\in\Omega j}\frac{1}{\sigma_{ji}^2}F_j-\sum_{j\in\Omega i}\frac{1}{\sigma_{ij}^2}F_j\\ &=\sum_j\left[(S_{\to i}+S_{i\to})\,\delta_{ji}-\frac{1}{\sigma_{ji}^2}I_{\{i\in\Omega_j\}}-\frac{1}{\sigma_{ij}^2}I_{\{j\in\partial_i\}}\right]F_j\\ &=:\sum_j A_{ij}F_j\end{aligned}$$

and equation (6.27) is recovered, with the definition of matrix $\mathbf{A}$ given in square brackets.

D.3.2 Empirical correction of redundancy in likelihood

D.3.2.1 The 1d uncorrelated gradients case

Consider the model

$$F_i-F_j\sim\Delta F_{ij}=(g_i+\eta_i)x_{ij} \tag{D.22}$$

where g_i is the exact gradient in i, x_{ij} is the vector between i and j and η_i is an uncorrelated gaussian noise of variance ε^2. This is assumed to be valid for $j\in NN_i$ the set of the first k neighbors of i.

This implies

$$\frac{F_i - F_j - \Delta F_{ij}}{\varepsilon x_{ij}} \sim \mathcal{N}(0,1)$$

The combination of variables defined above is equal to the same stochastic variable η_i/ε for all j. Therefore, we also have

$$\sum_{j \in NN_i} \frac{F_i - F_j - \Delta F_{ij}}{k \varepsilon x_{ij}} \sim \mathcal{N}(0,1)$$

Therefore the free energies F_i can be found by minimizing the quadratic likelihood

$$\mathcal{L}(F) = \frac{1}{2} \sum_i \left(\sum_{j \in NN_i} \frac{1}{k \varepsilon x_{ij}} \left(F_i - F_j - \Delta F_{ij} \right) \right)^2 = \frac{1}{2\varepsilon^2} \sum_i \left(\sum_{j \in NN_i} \frac{F_i - F_j}{k x_{ij}} - \tilde{g}_i \right)^2$$

The

$$\frac{\partial \mathcal{L}(F)}{\partial F_i} = \gamma_i \sum_j \frac{1}{x_{ij}} + \sum_j \gamma_j \frac{1}{x_{ij}}$$

D.3.3 Solution of the full BMTI model

D.3.4 Correlation structure of the $\widehat{\delta F}$'s

For any couple of neighboring points labelled by a single index $a := (i,j)$, the random variable $\widehat{\delta F}_a$ of mean value $\mu_a := \langle \widehat{\delta F}_a \rangle = F_j - F_i$ can be seen as marginal variable of a multivariate normal: $\widehat{\delta F}_a \sim \mathcal{N}(\mu_a \ , \ \varepsilon_a^2)$. Thus, the vector containing the $\widehat{\delta F}_a$'s for all couples of neighbours $\widehat{\boldsymbol{\delta F}} = \{\widehat{\delta F}_a\}_a$ is $\widehat{\boldsymbol{\delta F}} \sim \mathcal{N}(\boldsymbol{\mu} \ , \ \mathbf{C})$. The diagonal of this covariance matrix $\mathbf{C}$ is $C_{aa} = \varepsilon_a^2$ for all couples labelled by a. However $\mathbf{C}$ is generally not diagonal due to correlations between couples δF_a , δF_b estimated on at-least-partially overlapping regions of configurational space. Again, we refer the reader to

The exact form of $\mathbf{C}$ would again require the knowledge of $\mathbf{cov}[\hat{\mathbf{g}}_\mathrm{i}, \hat{\mathbf{g}}_\mathrm{j}]$, but, thanks to equation (6.19) its correlation structure can be captured quite well:

$$\begin{aligned} \langle \delta F_{ij} \delta F_{lm} \rangle &= \frac{\mathbf{r}_{ij}^{\mathrm{T}} \mathbf{r}_{lm}}{4} \langle (\hat{\mathbf{g}}_i + \hat{\mathbf{g}}_j)(\hat{\mathbf{g}}_l + \hat{\mathbf{g}}_m) \rangle \\ &= \frac{\mathbf{r}_{ij}^{\mathrm{T}} \mathbf{r}_{lm}}{4} \left[\langle \hat{\mathbf{g}}_i \hat{\mathbf{g}}_l \rangle + \langle \hat{\mathbf{g}}_i \hat{\mathbf{g}}_m \rangle + \langle \hat{\mathbf{g}}_j \hat{\mathbf{g}}_l \rangle + \langle \hat{\mathbf{g}}_j \hat{\mathbf{g}}_m \rangle \right] \\ &= \frac{1}{4} \left[p_{il} \, \varepsilon_{il}^{i} \, \varepsilon_{il}^{l} + p_{im} \, \varepsilon_{im}^{i} \, \varepsilon_{im}^{m} + p_{jl} \, \varepsilon_{jl}^{j} \, \varepsilon_{il}^{l} + p_{jm} \, \varepsilon_{jm}^{j} \, \varepsilon_{jm}^{m} \right] \ . \end{aligned} \tag{D.23}$$

One can decide whether to define $\mathbf{C}$ considering (i,j) and (j,i) different couples, thus preserving the directional structure of the neighbors graph, in which case its size is $N_{\text{spar}} \times N_{\text{spar}}$ with $N_{\text{spar}} = \sum_{i=1}^{N} (\hat{k}_i - 1) \approx N(< \hat{k} > -1)$; or one can consider them equal, in which case the size of $\mathbf{C}$ is even smaller. Either way, it looks like it could be possibly inverted for some thousand points, especially with a cutoff on $\hat{k}$. N_{spar} is generally much smaller than the total number of possible ordered couples

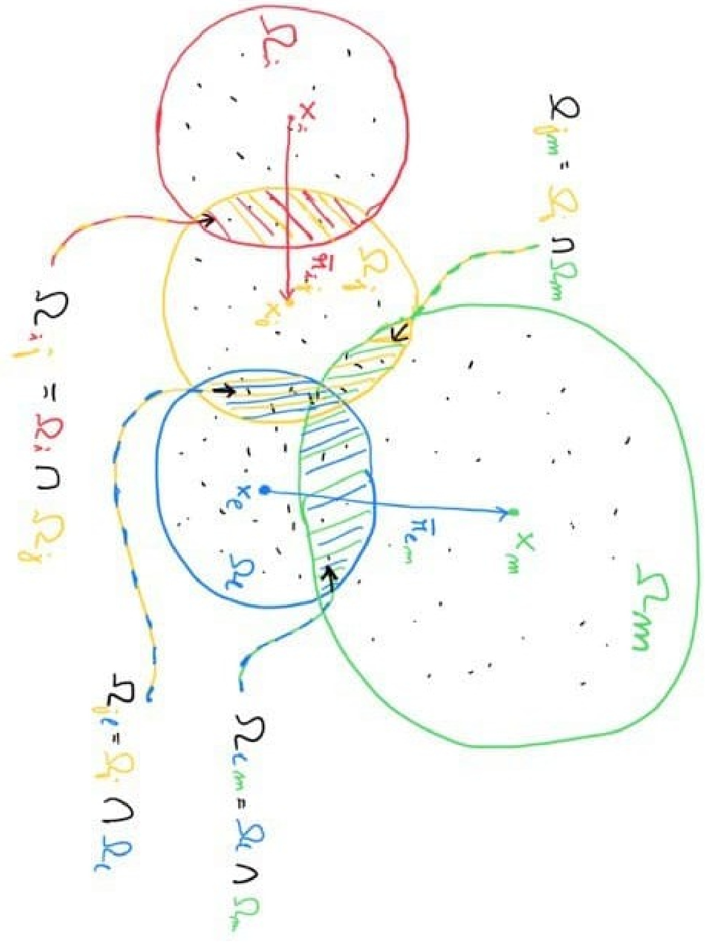


Figure D.2: Representation of neighbourhoods defining local estimates of the $\delta\hat{F}$'s and their correlation structures.

of indices, which is $N(N-1)$.