



PAPER • OPEN ACCESS

Density estimation via binless multidimensional integration

To cite this article: Matteo Carli *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 025052

View the [article online](#) for updates and enhancements.

You may also like

- [Explainable autoencoder for neutron star dense matter parameter estimation](#)
Francesco Di Clemente, Matteo Scialpi and Micha Bejger
- [Efficient prediction of aerodynamic forces in rarefied flow using convolutional neural network based multi-process method](#)
Haifeng Huang, Guobiao Cai, Chuanfeng Wei et al.
- [Data-driven thrust prediction in applied-field magnetoplasmadynamic thrusters for space missions using artificial intelligence-based models](#)
Tarik Pinaffo Almeida, Shahin Alipour Bonab and Mohammad Yazdani-Asrami



PAPER

OPEN ACCESS

RECEIVED
12 June 2024REVISED
9 April 2025ACCEPTED FOR PUBLICATION
1 May 2025PUBLISHED
4 June 2025

Original Content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



Density estimation via binless multidimensional integration

Matteo Carli^{1,2,*} , Alex Rodriguez^{3,4} , Alessandro Laio^{1,4,*} and Aldo Glielmo^{1,5,*} ¹ SISSA, Trieste, Italy² Harvard University, Cambridge, MA, United States of America³ University of Trieste, Trieste, Italy⁴ ICTP, Trieste, Italy⁵ Banca d'Italia, Rome, Italy

* Authors to whom any correspondence should be addressed.

E-mail: mcarli@sisssa.it, laio@sisssa.it and aldo.glielmo@bancaditalia.it**Keywords:** density estimation, free energy estimation, intrinsic manifold, unsupervised learning, nonparametric methods, mean shift, high-dimensional data analysisSupplementary material for this article is available [online](#)

Abstract

We introduce the binless multidimensional thermodynamic integration (BMTI) method for nonparametric, robust, and data-efficient density estimation. BMTI estimates the logarithm of the density by initially computing log-density differences between neighbouring data points. Subsequently, such differences are integrated, weighted by their associated uncertainties, using a maximum-likelihood formulation. This procedure can be seen as an extension to a multidimensional setting of the *thermodynamic integration*, a technique developed in statistical physics. The method leverages the manifold hypothesis, estimating quantities within the intrinsic data manifold without defining an explicit coordinate map. It does not rely on any binning or space partitioning, but rather on the construction of a neighbourhood graph based on an adaptive bandwidth selection procedure. BMTI mitigates the limitations commonly associated with traditional nonparametric density estimators, effectively reconstructing smooth profiles even in high-dimensional embedding spaces. The method is tested on a variety of complex synthetic high-dimensional datasets, where it is shown to outperform traditional estimators, and is benchmarked on realistic datasets from the chemical physics literature.

1. Introduction

Estimating a probability density function (PDF) from a finite set of samples is a fundamental challenge in statistics and machine learning, arising in a variety of practical applications [1–3]. In pursuing this task, parametric methods assume a functional form for the PDF and try to optimise a few parameters to best fit the observations [4–8]. They return smooth and robust PDF estimates even with little data, but badly specified parametric models can introduce systematic errors that are not healed by statistics [9]. Nonparametric density estimators, instead, seek to estimate a PDF without prior assumptions on the distribution [10]. This makes them more data-hungry, but also preferable to parametric ones when the shape and topography of the distribution peaks are not known *a priori* [11]. A popular example of a non-parametric density estimator is the kernel density estimator (KDE) [12], which reconstructs the PDF as a mixture of local copies of kernel functions—often Gaussians—around each data point. The k nearest neighbour (k NN) estimator [13] is another classic example: it estimates the density around each point proportionally to the inverse volume occupied in embedding space by the k nearest neighbours of that point. It can be thought of as a special type of KDE with a step-function kernel and an adaptive bandwidth selection. Gaussian KDEs perform remarkably well in very low dimensions, where they are typically preferred to k NN estimators as they provide much smoother density estimates. In fact, the kernel of k NN is non-differentiable. Practically, this can translate into noisier and less accurate estimates [1, 10]. However, the performance of fixed-bandwidth KDEs drastically deteriorates with the increase of the embedding space dimension. Already beyond 2 or 3

dimensions, these estimates become biased [2] and k NN outperforms standard KDEs due to its point-adaptive nature. In fact, high-dimensional data face the so-called *curse of dimensionality* [9, 14, 15].

The selection of the smoothing parameter confronts a *bias-variance tradeoff*: opting for a higher value enhances statistical stability while reducing noise, yet it may introduce bias when the underlying PDF exhibits substantial variations across the spanned region. Achieving a delicate balance becomes even more crucial when dealing with sparse data, as often encountered in high-dimensional spaces. To address this problem the bandwidth should be carefully selected globally or, preferably, locally (adaptively), a problem to which a great amount of research effort has been devoted [16–18]. In this work, we address the described shortcomings of traditional estimators and propose a nonparametric method able to provide accurate and smooth estimates of density in high dimensional spaces. Our method is based on a reconstruction of the density starting from measurements of the difference of the logarithm of PDF (log-density hereafter) among pairs of neighbouring points. The smoothness of the resulting density surface in our approach is guaranteed by the fact that the densities are estimated consistently on neighbouring points from the log-density difference between all the pairs of data within a neighbourhood. The log-density differences are estimated by a modified mean shift algorithm [19], extended through a restriction to the intrinsic data manifold and an adaptive selection of local bandwidth. The accuracy of our approach in high dimensions arises both from the error-correcting nature of the pair-measurements reconstruction and by the fact that all relevant quantities can be estimated directly on the intrinsic data-manifold, and not in the embedding space. The method we propose, named binless multidimensional thermodynamic integration (BMTI) is also remarkably data-efficient.

Our contributions are as follows:

- We propose an extension of the mean-shift algorithm which allows estimating log-density gradients directly on the intrinsic data manifold using a local bandwidth selection. The resulting estimates are considerably more robust against the curse of dimensionality.
- We propose a method to reconstruct—via a ‘binless multidimensional integration’ – a log-density profile from measurements of log-density differences between neighbouring points. We demonstrate that the approach gives rise to a robust, data-efficient, and smooth density estimator and show that it outperforms state-of-the-art methods at least up to 20 dimensions.
- We provide an open source code Python implementation of the BMTI log-density estimator and of the improved mean-shift gradient estimator⁶

2. Related work

2.1. The manifold hypothesis and ID estimation

According to the manifold hypothesis [6], the distribution of any real-world high-dimensional dataset has support on a manifold, called *intrinsic* or *latent data manifold*, whose *intrinsic dimension* (ID) [21] is typically much lower than the dimension of the embedding space [22, 23]. In recent times the manifold hypothesis has found strong empirical evidence within several scientific fields, from physics and biology [24, 25] to medicine [26] and even epidemiology [27] or deep learning [21, 28]. This led to a renovated interest in the development of accurate ID estimators [20, 24, 29–33] all of which, importantly, need to be able to cope with manifolds which are, in general, curved, twisted and topologically complex [6]. The idea of estimating the PDF on the intrinsic manifold to alleviate the curse of dimensionality was first proposed in [34] for KDEs. In [35], the authors use the concept of ID to improve traditional k NN-based estimators by restricting the computation of the volumes to the intrinsic manifold. In [36, 37], the authors propose deep learning approaches for density estimation directly on the intrinsic manifold. In the first approach, they employ deep generative neural networks, while in the second approach, they extend the normalising flow (NF) scheme introduced in [38].

2.2. Nonparametric density estimation

Nonparametric methods try to infer the underlying PDF from a dataset without assuming its specific functional form. They are fully data-driven and only look at local properties of the sample [39]. They usually depend on at least one parameter, which typically controls this definition of locality, generally named *smoothing parameter* or *bandwidth*. The optimal choice of this parameter is a long-standing open problem in statistics [40]. The simplest and oldest nonparametric methods are histograms [41], which divide the embedding space into fixed partitions. KDEs, instead, are among the most popular and versatile. They first appeared in [13], where the uniform KDE and the k NN estimator were introduced [42]. They were soon

⁶ Our implementation is available within DADapy [20] at <https://github.com/sissa-data-science/DADapy>.

generalised to a wider class of functions, among which smooth kernels, yield a continuous PDF. Convergence and asymptotic properties were studied [12, 43] and the approach was extended to the multidimensional case [44]. k NN stands out as a brilliant but very simple way to adaptively select the bandwidth of a KDE. Instead of fixing explicitly the bandwidth, it sets the number of neighbours to be considered for each estimate, which makes the local level of smoothing depend on the local density. An extension of the k NN idea to other kernels is the variable kernel estimator [45]. Other proposed methods for bandwidth selection include global methods [1, 11, 46, 47] and adaptive ones [4, 48–52]. Another class worth mentioning is composed of Voronoi density estimators [53]. They do not introduce a geometric bias even in the choice of the kernel. However, the computation of the Voronoi tessellation of the embedding space can be very computationally demanding and the estimates are typically discontinuous at the tile boundaries [54]. While these discontinuities have been recently treated [55], the resulting PDF is still very spiky with low statistics. Finally, in recent years, neural network approaches have gained traction in nonparametric density estimation due to their great flexibility [56–60]. However, these approaches are quite data-hungry and computationally intensive, requiring large datasets and complex training schemes.

2.3. Unnormalised log-density, energy, and free energy

In the task of inferring a PDF from a finite sample, computing the correct normalization can be very challenging if not prohibitive [61]. Fortunately, in many practical applications, one is only interested in the derivatives of the PDF [19] or in the relative value of the PDF at different points [62, 63]. It is the case of several pattern-recognition and image processing algorithms [39, 64–70], but also of physics, where the log-density over a projection of configuration space is typically proportional to a thermodynamic potential called *free energy* [71–74]. In the ML community, the unnormalised negative log-density quantity is referred to as *energy* [75], like in the case of *energy-based models* (EBMs) [76, 77]. Such quantity is the one we are concerned with in this work.

2.4. Thermodynamic integration (TI)

The term *TI* indicates a technique first developed in the field of chemical physics to reconstruct the free energy of a system through the knowledge of its derivatives. Traditionally, TI is carried out by integrating a directional derivative along a carefully chosen, physically meaningful, one-dimensional path [74, 78]. The approach has been extended to many dimensions [79], but integrating the gradient in more than one dimension is considered a difficult task [80]. Often the gradient is sampled on a dense grid [81], a data-intensive approach allowing for free energy reconstruction only up to dimension 2 or 3 [79, 82, 83]. In these cases, the most accurate integration method is the solution of the Poisson equation on a shifted mesh [84–86]. The most common integration method in literature [87, 88], introduced in [89] makes use of radial basis functions to reconstruct the free energy from the gradient, which has however been employed in spaces of no more than 4 dimensions due to its high computational demand. Finally, a recently proposed method [90] uses gradient information to define a Monte Carlo protocol to smartly sample the embedding space and populate a multidimensional histogram. This approach has been pushed, quite astonishingly, up to dimension 6 and it represents, to our knowledge, the highest embedding dimension at which TI has been previously employed.

3. Binless Multidimensional Thermodynamic Integration (BMTI)

Let $\rho(\mathbf{x})$ be the PDF of the multidimensional random variable $\mathbf{x}$. We assume that all the $\mathbf{x}_i$ are harvested independently from ρ . The goal of our approach is estimating the negative of the logarithm of the density, which from now we refer to as the negative log-density (NLD), and denote $F(\mathbf{x}) = -\ln \rho(\mathbf{x})$. The minus sign is introduced for consistency with the physics and physical chemistry literature, where such quantity is interpreted as a free energy. Due to this convention the maxima of $\rho(\mathbf{x})$ correspond to the minima of $F(\mathbf{x})$. We denote the NLD at a given datapoint $\mathbf{x}_i$ by $F_i = F(\mathbf{x}_i)$.

Suppose that each point $\mathbf{x}_i$ has a number k_i of neighbours that can be assumed to have a similar density. Importantly, the value of k_i is point-dependent. This neighbourhood will be defined more precisely in the following. This induces a—typically-sparse—directed *neighbourhood graph* (NG) (see figure 1) in which i is connected to j only if j is a neighbour of i . Further suppose that for any connected pair (i, j) in the NG we have an estimate of the NLD difference $\delta F_{ij} = F_j - F_i$. Then, in the spirit of the TI the NLD difference ΔF_{il} between any two points i and l , can be computed by choosing any of the multiple paths that connect i to l on the NG

$$\Delta F_{il} \simeq \delta F_{i,j_1} + \delta F_{j_1,j_2} + \dots + \delta F_{j_n,l} \quad (1)$$

where $j_1, \dots, j_n$ define one specific path from i to l . For each pair of endpoints i and l there are many possible choices of paths, but of course ΔF_{il} must be the same in all cases. This fact calls for a procedure that estimates

NLD differences among points considering contributions from different paths connecting them. In the rest of this section, we will introduce the *BMTI* estimator, which implements this idea. BMTI estimates negative log-densities at each point of the dataset simultaneously as the solution of a linear system obtained by maximizing a likelihood that incorporates the paths' weights. BMTI is *Binless* as it performs *TI* on a graph rather than using some kind of configuration space partitioning, and is *Multidimensional* since it performs remarkably well for high-dimensional data, for which, we will show, it typically outperforms other non-parametric methods.

3.1. Derivation of the BMTI estimator

For a lighter notation, let us label a pair of points linked by an edge on the NG by a single index $a = (i, j)$, so that $\{a, b, \dots\}$ is the set of N_e directed edges of the NG $\{(i, j), (l, m), \dots\}$, with $N_e = \sum_{i=1}^N (k_i - 1) = N(\langle k \rangle - 1)$. We assume that for each pair of neighbouring points a we have an unbiased estimate $\hat{\delta F}_a$ of the true NLD difference δF_a . We further assume that our vector of estimates $(\hat{\delta F})_a := \hat{\delta F}_a$ is normally distributed around the true values $(\delta F)_a := \delta F_a$. This implies that our estimator $\hat{\delta F}$ follows a multivariate normal distribution centred on the true vector δF with covariance matrix $\mathbf{C} := \text{cov}[\hat{\delta F}, \hat{\delta F}]$. In short, we assume

$$\hat{\delta F} \sim \mathcal{N}(\delta F, \mathbf{C}) . \quad (2)$$

In section 3.2 we introduce a procedure to effectively estimate NLD differences on the NG and their covariance. Now, by focusing on the $\hat{\delta F}$'s distribution

$$\mathcal{N}(\delta F, \mathbf{C}) \propto \exp \left[-\frac{1}{2} \sum_{a,b} \left(\delta F_a - \hat{\delta F}_a \right)^T C_{a,b}^{-1} \left(\delta F_b - \hat{\delta F}_b \right) \right] , \quad (3)$$

we note that the argument of the exponential, in square brackets, can be recast into a quadratic form for the F 's. By calling $(\mathbf{F})_i := F_i$ the vector of all the negative log-densities at sample points, this quadratic form reads $\mathbf{F}^T \mathbf{A} \mathbf{F} + \mathbf{b}^T \mathbf{F} + c$, where the $N \times N$ matrix $\mathbf{A}$, the N -vector $\mathbf{b}$ and the scalar c are explicit functions of the estimated NLD differences $\hat{\delta F}$ and on their covariance matrix $\mathbf{C}$. Therefore, we can interpret the logarithm of this N_e -variate Gaussian as a log-likelihood for the parameters $\mathbf{F}$ given the error-affected observations $\hat{\delta F}$

$$\mathcal{L}(\mathbf{F} | \hat{\delta F}, \mathbf{C}) \propto \ln \mathcal{N}(\delta F, \mathbf{C}) . \quad (4)$$

By maximizing this log-likelihood over the parameters $\mathbf{F}$, one obtains a linear system

$$\hat{\mathbf{F}} := \underset{\mathbf{F}}{\text{argmax}} \mathcal{L}(\mathbf{F} | \hat{\delta F}, \mathbf{C}) \Rightarrow \mathbf{A} \hat{\mathbf{F}} = \mathbf{b} \quad (5)$$

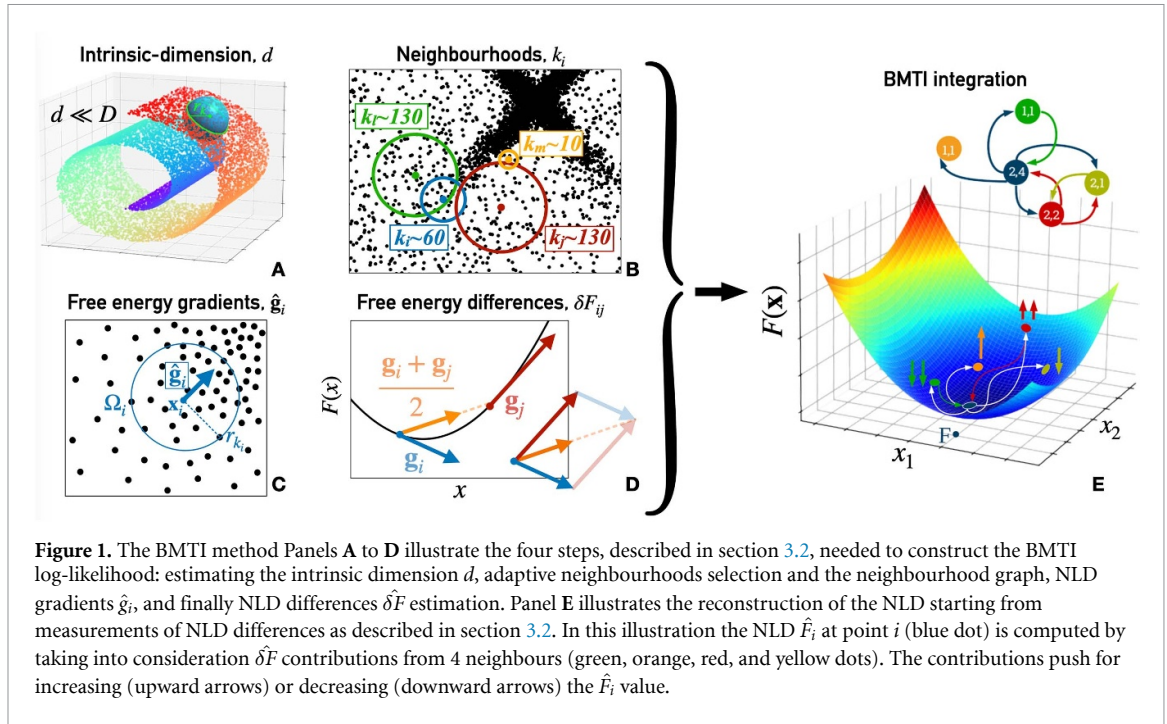
whose solution defines the BMTI NLD estimator

$$\hat{\mathbf{F}} = \mathbf{A}^{-1} \mathbf{b} . \quad (6)$$

The maximum-likelihood estimates $\hat{\mathbf{F}}$ are thus found simultaneously and coherently for all points in the dataset through a mechanism that is illustrated in figure 1(E). Locally, the estimated $\hat{F}_i$ receives additive contributions from its neighbours. These contributions are positive or negative depending on the sign of $\hat{\delta F}_{ij}$ and inversely proportional to the uncertainties on the $\hat{\delta F}$'s involving i or j . This framework also allows to estimate Cramér–Rao uncertainties from the Hessian of the log-likelihood (see section C.1 of the SM)

$$\sigma^2 [\hat{F}_i] := A_{ii}^{-1} . \quad (7)$$

Since the NLD enter the log-likelihood only in terms of differences, the values of the NLDs are determined up to an arbitrary additive constant. This makes the linear system in equation (5) underdetermined and the symmetric matrix $\mathbf{A}$ singular. In practice, the linear system can be regularised without loss of accuracy and solved using any standard linear algebra library implementing, e.g. the conjugate gradient method [91], while the matrix $\mathbf{A}^{-1}$ appearing in equations (6) and (7) can be interpreted as a Moore–Penrose pseudoinverse of $\mathbf{A}$ [92].



3.2. Estimation of the BMTI log-likelihood

In order to compute the log-likelihood in equation (4) we need to evaluate the $\delta\hat{F}$'s and their covariances. This is done in four steps: (i) the estimation of the ID, (ii) an adaptive neighbourhood selection and the construction of an NG, (iii) the estimation of local NLD gradients, and finally (iv) the estimation of NLD differences and their correlations. These steps are summarised in figures 1(A)–(D). The first two are based on previous work and are described in section 3.2.1. Step (iii) is described in section 3.2.2, while step (iv) is covered in section 3.2.3 and in the SM.

3.2.1. Restriction to the intrinsic manifold and adaptive neighbourhood selection

In order to restrict density estimates to the intrinsic data manifold, the first step is to reliably estimate its ID, d . In this work we adopt the *TwoNN* estimator [30] as implemented in [20]. The next assumption is that the data manifold $\mathcal{M}$ is Riemannian [34]. In this setting, on a small enough scale, geodesic distances on $\mathcal{M}$ are equal to Euclidean distances in the d -dimensional tangent hyperplane to $\mathcal{M}$; these are in turn equivalent to Euclidean distances in the embedding space of dimension D , since the PDF vanishes outside $\mathcal{M}$ [35, 93] (see figure 1(A)). Therefore, small local hyperspherical volumes in $\mathcal{M}$ can be computed as the volume of an Euclidean d -ball. We remark that an accurate estimate of d is crucial to estimate these volumes. BMTI also relies on the definition of the number k_i of nearest neighbours for each point i of the dataset. This is needed for two reasons: it selects a local region of space Ω_i for the estimation of NLD gradients, making the gradient estimates, and thus BMTI, point-adaptive and more robust in high dimensions; it allows for the construction of a directed NG in which i is connected to j when j is in Ω_i , which is essential for BMTI integration, as depicted at the top-right of figure 1(E). In our specific case, for each point i , we select the largest number of neighbours k_i over which the density does not vary significantly. For this purpose, we use the likelihood ratio test proposed in [35], where k_i gradually increases until the density of point i and its k_i th nearest neighbour are statistically different according to a certain tolerance. We refer the reader to section E.1 of the SM for an intuitive explanation and visualisation, and to the original paper for an extended explanation of this technique. Figure 1(B) illustrates the adaptive selection of k_i and the construction of the NG. Note that alternative NG selection methods can also be adopted. While BMTI performance will depend on the adaptivity and ability of the NG to define appropriately-sized neighbourhoods, the BMTI algorithm is defined for any choice of $\{k_i\}_i$. In section E.2 of the SM, we test the performance of BMTI with different choices of adaptive and fixed neighbourhood sizes. Overall, BMTI displays robustness with respect to the NG selection and consistently outperforms k NN estimators.

3.2.2. Estimation of log-density gradients

For notation convenience, let us refer to the gradient of the NLD at point i as

$\mathbf{g}_i := \nabla_{\mathbf{x}} F(\mathbf{x}_i) = \nabla_{\mathbf{x}} \rho(\mathbf{x}_i) / \rho(\mathbf{x}_i)$. The mean shift $\mathbf{m}_i$ around point i within the region Ω_i of radius r_i is defined

as the centred expectation $\mathbf{m}_i := \langle \mathbf{x} - \mathbf{x}_i \rangle_{\Omega_i}$. More explicitly

$$\mathbf{m}_i := \frac{\int_{\Omega_i} \rho(\mathbf{x}) (\mathbf{x} - \mathbf{x}_i) d\mathbf{x}}{\int_{\Omega_i} \rho(\mathbf{x}) d\mathbf{x}}. \quad (8)$$

By expanding the density $\rho(\mathbf{x})$ around $\mathbf{x}_i$ as

$$\rho(\mathbf{x}) \approx \rho(\mathbf{x}_i) + \nabla_{\mathbf{x}}^T \rho(\mathbf{x}_i) (\mathbf{x} - \mathbf{x}_i), \quad (9)$$

and by evaluating the integral in equation (8) with the expression in equation (9), we obtain the following approximate relationship between NLD gradient and mean shift, for which an intuition is provided by figure 1(C)

$$\mathbf{g}_i \approx -\frac{d+2}{r_i^2} \mathbf{m}_i. \quad (10)$$

By substituting the expectation $\mathbf{m}_i$ in the above equation with a sample average $\hat{\mathbf{m}}_i$, and by using the point-adaptively optimised neighbourhood size k_i , defined as described in section 3.2.1, to identify the region Ω_i we estimate the NLD gradient (see again figure 1(C)) as

$$\hat{\mathbf{g}}_i = -\frac{d+2}{r_{k_i}^2} \hat{\mathbf{m}}_i \quad (11)$$

$$\hat{\mathbf{m}}_i = \frac{1}{\tilde{k}_i} \sum_{j \in \Omega_i} (\mathbf{x}_j - \mathbf{x}_i) \quad (12)$$

where we defined $\tilde{k}_i = k_i - 1$ for notation convenience, and r_{k_i} is the distance between point i and its k_i th nearest neighbour. In sections A.1.2 and A.1.3 of the SM we report a rigorous derivation of the two equations above, which generalise the procedure in [19]. It is worth stressing that $\hat{\mathbf{g}}_i$ in our approach is doubly adaptive in the same sense discussed in section 3.2.1, since it restricts to the intrinsic manifold of dimension $d \ll D$ and operates an adaptive bandwidth selection; this is crucial to enhance the estimator's performance and robustness against the curse of dimensionality with respect to the original mean-shift gradient estimator [19]. Since the $\hat{\mathbf{g}}_i$ estimator is proportional to the arithmetic average of $k-1$ shift random variables (RVs), i.e. to the sample mean-shift estimator $\hat{\mathbf{m}}_i$ defined in equation (12), its auto-covariance $\text{var}[\hat{\mathbf{g}}_i]$ is proportional to the auto-covariance of the mean shift. The sample gradient autocovariance estimator reads

$$\text{var}[\hat{\mathbf{g}}_i] = \left(\frac{d+2}{r_{k_i}^2} \right)^2 \text{var}[\hat{\mathbf{m}}_i]. \quad (13)$$

Equation (13) allows to estimate uncertainties on $\hat{\mathbf{g}}_i$ and its components, as discussed more in depth in section A.1.5.

Note that within our framework other radially-symmetric kernels can be employed [66, 68, 94]. We choose the extension of Mean Shift described above because, as we will show, it provides a reliable estimate of the NLD differences and of the error of their error even in high dimensions. Importantly, [19] proves asymptotic unbiasedness, consistency, and uniform consistency of the estimator in equation (12) for all well-behaved kernel shapes [44], thus these are guaranteed also in our case. The performance of $\hat{\mathbf{g}}$ and its error estimator are assessed in section A.2 of the SM.

3.2.3. Estimation of the δF 's

Using the NLD gradient estimator $\hat{\mathbf{g}}$ in equation (12), we estimate the NLD differences δF_{ij} between neighbouring points $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$\delta F_{ij} := \frac{\hat{\mathbf{g}}_i + \hat{\mathbf{g}}_j}{2} \cdot \mathbf{r}_{ij}, \quad (14)$$

where $\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i$. Contracting $\mathbf{r}_{ij}$ with the average of the two gradient estimates is more accurate than using any of the two singularly. The intuition behind this choice is illustrated in figure 1(D). Formally, it is simple to show that by taking the half-sum of $\hat{\mathbf{g}}_i$ and $\hat{\mathbf{g}}_j$ the quadratic error terms of the two estimators cancel out, leading to errors of order $\mathcal{O}(\|\mathbf{r}_{ij}\|^3)$ in the final estimator for δF_{ij} . A derivation of equation (14) can be found in section B.1 of the SM.

The uncertainty on the estimate δF_{ij} can be quantified by its variance $\varepsilon_{ij}^2 := \text{var}[\delta F_{ij}]$, which can be derived from equation (14) to be

$$\begin{aligned}\varepsilon_{ij}^2 &= \mathbf{r}_{ij}^T \mathbf{var} \left[\frac{\hat{\mathbf{g}}_i + \hat{\mathbf{g}}_j}{2} \right] \mathbf{r}_{ij} \\ &= \frac{1}{4} \mathbf{r}_{ij}^T (\mathbf{var}[\hat{\mathbf{g}}_i] + \mathbf{var}[\hat{\mathbf{g}}_j] + 2 \mathbf{cov}[\hat{\mathbf{g}}_i, \hat{\mathbf{g}}_j]) \mathbf{r}_{ij}.\end{aligned}\quad (15)$$

Note that the ε_{ij}^2 are indeed the diagonal elements of the covariance matrix $\mathbf{C}$ in equation (2) as $C_{a,a} = \text{cov}[\delta F_a, \delta F_a] = \varepsilon_a^2$.

By defining the single-gradient estimates of the NLD as $\delta F_{ij}^i = \hat{\mathbf{g}}_i \cdot \mathbf{r}_{ij}$, with the upper bold index indicating whether the gradient on i or j is used, we can equivalently write the NLD estimates of equation (14) as $\delta F_{ij} = \frac{1}{2}(\delta F_{ij}^i + \delta F_{ij}^j)$ and their variance as $\varepsilon_{ij}^2 := \mathbf{r}_{ij}^T \cdot \mathbf{var}[\hat{\mathbf{g}}_i] \cdot \mathbf{r}_{ij}$. Finally, this allows us to rewrite equation (15) only in terms of scalar quantities

$$\varepsilon_{ij}^2 = \frac{1}{4} \left(\varepsilon_{ij}^{i^2} + \varepsilon_{ij}^{j^2} + 2 p^{ij} \varepsilon_{ij}^i \varepsilon_{ij}^j \right), \quad (16)$$

where p^{ij} is the Pearson correlation coefficient between δF_{ij}^i and δF_{ij}^j and is defined by

$$p^{ij} := \frac{\text{cov}[\delta F_{ij}^i, \delta F_{ij}^j]}{\varepsilon_{ij}^i \varepsilon_{ij}^j}. \quad (17)$$

The formulation in equations (16) and (17) is at the basis of our approach for estimating the full covariance structure of the δF 's, as will become clear in the next section.

3.2.4. Estimation of $\mathbf{C}$: the δF 's covariance

In order to accurately reconstruct the NLD according to equation (4), it is crucial to estimate the covariance matrix $\mathbf{C}$ of the δF 's, which—evidently from equation (14)—depends on the gradient covariance structure. The auto-covariance matrices of the gradient estimators appearing in equation (15), $\mathbf{var}[\hat{\mathbf{g}}_i]$ and $\mathbf{var}[\hat{\mathbf{g}}_j]$, can be estimated from the neighbourhoods Ω_i and Ω_j by means of equation (13). The cross-covariance between $\hat{\mathbf{g}}_i$ and $\hat{\mathbf{g}}_j$, instead, depends on a statistic computed on overlapping region between the two neighbourhoods, i.e. $\Omega_{i,j} := \Omega_i \cap \Omega_j$. In fact, as we show in section A.1.5 of the SM, equations (S.35) and (S.36), its expression reads

$$\begin{aligned}\mathbf{cov}[\hat{\mathbf{g}}_i, \hat{\mathbf{g}}_j] &= \frac{(d+2)^2}{r_{k_i}^2 r_{k_j}^2} \mathbf{cov}[\hat{\mathbf{m}}_i, \hat{\mathbf{m}}_j] \\ \mathbf{cov}[\hat{\mathbf{m}}_i, \hat{\mathbf{m}}_j] &= \frac{k_{i,j}}{\bar{k}_i \bar{k}_j} \left[\langle (\mathbf{x} - \mathbf{x}_j) (\mathbf{x} - \mathbf{x}_j)^T \rangle_{\Omega_{i,j}} - \mathbf{m}_i \mathbf{m}_j^T \right],\end{aligned}\quad (18)$$

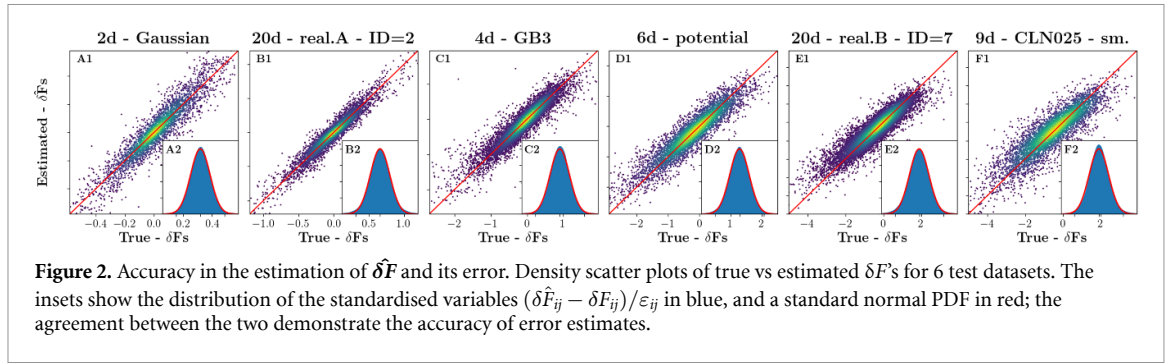
where the expectation in intersection region Ω_{ij} could theoretically be estimated using the $k_{i,j}$ points in the region. Unfortunately, it is common for i - j pairs to have only a small number of common neighbours $k_{i,j}$, and this would make the estimator very noisy. We here exploit our geometrical intuition on the NG (see figure 1, panels B and E) to give an empirical estimate of the Pearson correlation coefficient p^{ij} in equations (16) and (17). By calling $\chi^{ij} = |p^{ij}|$ and $s_{ij} = \text{sgn}(p^{ij})$, so that $p^{ij} = s^{ij} \chi^{ij}$, we define the following empirical estimators

$$\hat{p}^{ij} = \hat{s}^{ij} \hat{\chi}^{ij} \quad (19)$$

$$\hat{s}^{ij} = \text{sgn} \left(\delta F_{ij}^i \delta F_{ij}^j \right) \quad (20)$$

$$\hat{\chi}^{ij} = \frac{k_{i,j}}{k_i + k_j - k_{i,j}} \quad (21)$$

Thus, the absolute value of p^{ij} , namely χ^{ij} , is estimated by the *Jaccard index* [95] of the two neighbourhoods Ω_i and Ω_j ; it is bound to be between 0, when $\hat{\mathbf{g}}_i$ and $\hat{\mathbf{g}}_j$ are not correlated, i.e. $\Omega_i \cap \Omega_j = \emptyset$, and 1, when $i = j$. Its sign, instead, is approximated by equation (20): this approximation, as we will see, works well in practice. In section B.2.1 and B.2.3 of the SM we motivate the expressions in equations (19)–(21). By using them in equation (16) to estimate p^{ij} , we obtain an estimator for the variances of the δF 's in equation (15), i.e. for the diagonal elements of the covariance matrix $\mathbf{C}$. The great accuracy of these estimators is empirically assessed



in figure 2, together with the Gaussianity and unbiasedness of the $\delta\hat{F}$ estimators in equation (14), required for BMTI in equation (2).

One last step is represented by the estimation of generic elements of the δF 's covariance matrix, including off-diagonal terms, namely $C_{ij,lm} := \text{cov}[\delta\hat{F}_{ij}, \delta\hat{F}_{lm}]$, which read

$$C_{ij,lm} = \frac{1}{4} \mathbf{r}_{ij}^T (\text{cov}[\hat{\mathbf{g}}_i, \hat{\mathbf{g}}_l] + \text{cov}[\hat{\mathbf{g}}_i, \hat{\mathbf{g}}_m] + \text{cov}[\hat{\mathbf{g}}_j, \hat{\mathbf{g}}_l] + \text{cov}[\hat{\mathbf{g}}_j, \hat{\mathbf{g}}_m]) \mathbf{r}_{lm}. \quad (22)$$

Equation (22) is a general version of equation (15). Again, to obtain the general version of equation (16), one can contract the vector differences $\mathbf{r}$ with the gradient cross-covariance matrices and define Pearson correlation coefficients. We report here the final expression that we propose to estimate $C_{ij,lm}$ in equation (22)

$$\hat{C}_{ij,lm} = \frac{1}{4} (\hat{p}_{ij,lm}^{il} \hat{\epsilon}_{ij}^i \hat{\epsilon}_{lm}^l + \hat{p}_{ij,lm}^{im} \hat{\epsilon}_{ij}^i \hat{\epsilon}_{lm}^m + \hat{p}_{ij,lm}^{jl} \hat{\epsilon}_{ij}^j \hat{\epsilon}_{lm}^l + \hat{p}_{ij,lm}^{jm} \hat{\epsilon}_{ij}^j \hat{\epsilon}_{lm}^m), \quad (23)$$

where the correlation coefficients are estimated as

$$\hat{p}_{ij,lm}^{il} = \text{sgn}(\delta\hat{F}_{ij}^i \delta\hat{F}_{lm}^l) \hat{\chi}^{il}. \quad (24)$$

A more thorough derivation and justification of equations (22)–(24) are contained in sections B.2.2, B.2.3, and B.2.4 of the SM. Equations (23) and (24) provide a formula to estimate any element of the covariance matrix $\mathbf{C}$ defining the BMTI log-likelihood in equation (4). Note that the expression in equation (24) is coherent with equation (19) by identifying $p^{ij} \equiv p_{ij,ij}^{ij}$. Therefore, equation (16) is recovered from equation (23) by setting $ij = lm$.

3.3. Approximate BMTI likelihood

It is not the covariance matrix $\mathbf{C}$ itself that enters the definition of BMTI likelihood in equation (4), but rather its inverse, the precision matrix $\mathbf{C}^{-1}$. Since the N_e elements of $\delta\hat{\mathbf{F}}$ are estimated based only on N independent RVs (the sample points), the matrix $\mathbf{C}$ can in general be singular. Therefore, $\mathbf{C}^{-1}$ could be taken as a Moore–Penrose pseudoinverse. Since estimating the exact pseudoinverse becomes rapidly numerically costly as the sample size increases, in practice we approximate $\mathbf{C}^{-1}$ by a diagonal matrix $\mathbf{D}$ whose non-zero elements are D_{ij} . With it the BMTI likelihood reads

$$\mathcal{L}(\mathbf{F} | \delta\hat{\mathbf{F}}, \mathbf{D}) := - \sum_{i=1}^N \sum_{j \in \Omega_i} \frac{D_{ij}}{2} (\delta F_{ij} - \delta\hat{F}_{ij})^2. \quad (25)$$

In section C.2 of the SM we discuss various possible approaches to define the approximate precision $\mathbf{D}$ —all of which yield accurate predictions $\hat{\mathbf{F}}$ for all tested datasets and in all sampling regimes—and the cases in which they also allow recovering accurate uncertainty estimations. Here, in the main text, we recall only the simplest of these possible choices for $\mathbf{D}$, namely $D_{ij} := 1/C_{ij,ij} = 1/\epsilon_{ij}^2$, with which (25) becomes

$$\mathcal{L}(\mathbf{F} | \delta\hat{\mathbf{F}}, \mathbf{D}) := - \sum_{i=1}^N \sum_{j \in \Omega_i} \frac{(F_j - F_i - \delta\hat{F}_{ij})^2}{2\epsilon_{ij}^2}. \quad (26)$$

All the numerical experiments presented in section 4 are conducted using this setting.

The complexity of BMTI is dominated by the solution of the linear system in equation (6). The direct solution of a dense linear system generally requires $\mathcal{O}(N^2)$ in memory and $\mathcal{O}(N^3)$ in time [96]. However,

since the matrix $\mathbf{A}$ is typically very sparse, we can efficiently use sparse solvers. In all cases considered, we found that an iterative sparse solver, namely *conjugate gradient* [91], drastically reduces the memory and time complexity to linear and almost-linear respectively. A time-scaling assessment of this method for increasing dataset sizes is presented in figure 5(A). In section C.5 of the SM we discuss the scaling and implementation details.

3.4. Regularisation of the BMTI likelihood for disconnected NGs

BMTI integrates NLD differences on the NG and estimates the NLD up to a constant offset due to the PDF normalisation. An important shortcoming appears in cases in which, due to low sampling or to the PDF morphology, the NG on which the NLD differences are estimated becomes disconnected. In such a scenario, the BMTI formulation provided in section 3.1, fails to correctly reconstruct the relative density of the connected subgraphs. Fortunately, this problem can be healed by combining the log-likelihood of a strictly-local NLD estimator with the BMTI log-likelihood via a mixing hyperparameter α as a regularisation term, so that the total log-likelihood becomes

$$\mathcal{L}^{\text{tot}}(\mathbf{F}|\hat{\delta\mathbf{F}}, \mathbf{C}, \boldsymbol{\theta}) = \alpha \mathcal{L}^{\text{BMTI}}(\mathbf{F}|\hat{\delta\mathbf{F}}, \mathbf{C}) + (1 - \alpha) \mathcal{L}^{\text{reg}}(\mathbf{F}|\boldsymbol{\theta}) \quad (27)$$

where $\boldsymbol{\theta}$ indicates the parameters that $\mathcal{L}^{\text{reg}}$ depends on. As long as the regularising $\mathcal{L}^{\text{reg}}$ produces normalised NLD estimates, the NLD coming from $\mathcal{L}^{\text{tot}}$ will also be correctly normalised, also in the case of a disconnected NG. Our choice for $\mathcal{L}^{\text{reg}}$ falls on a series of k NN-based estimators which can be formulated in terms of maximum-likelihood estimators [35, 97]. Implementation details are discussed in the SM, Sec.C.4.

4. Numerical experiments

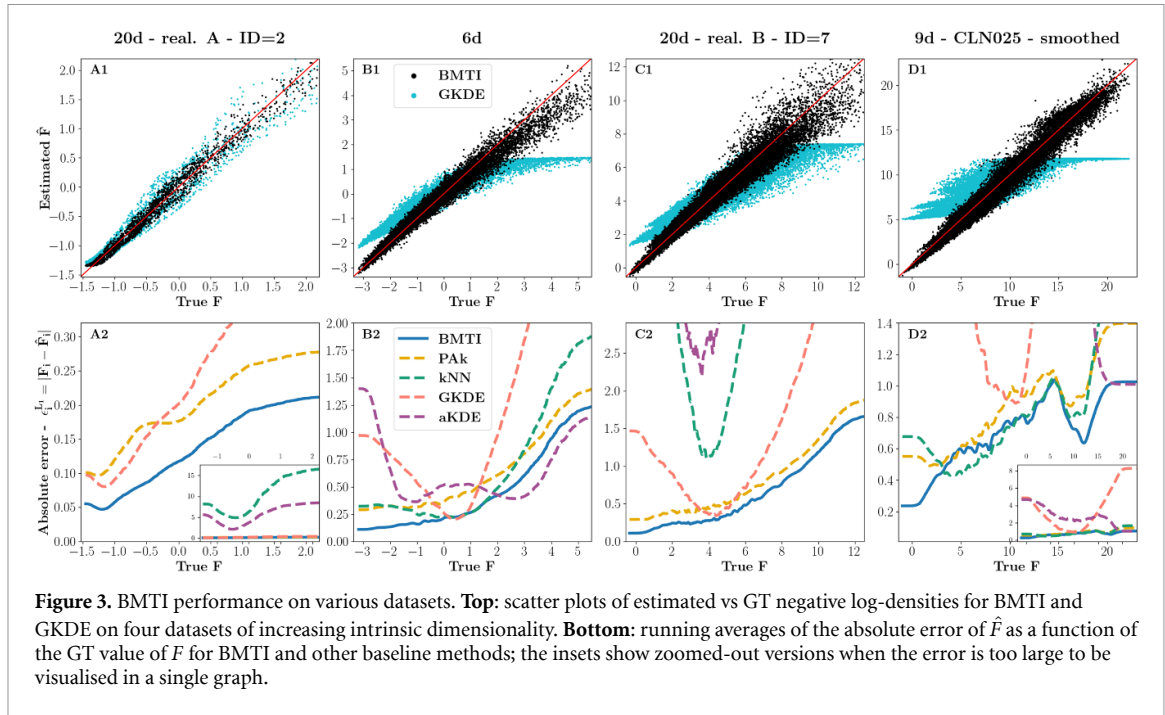
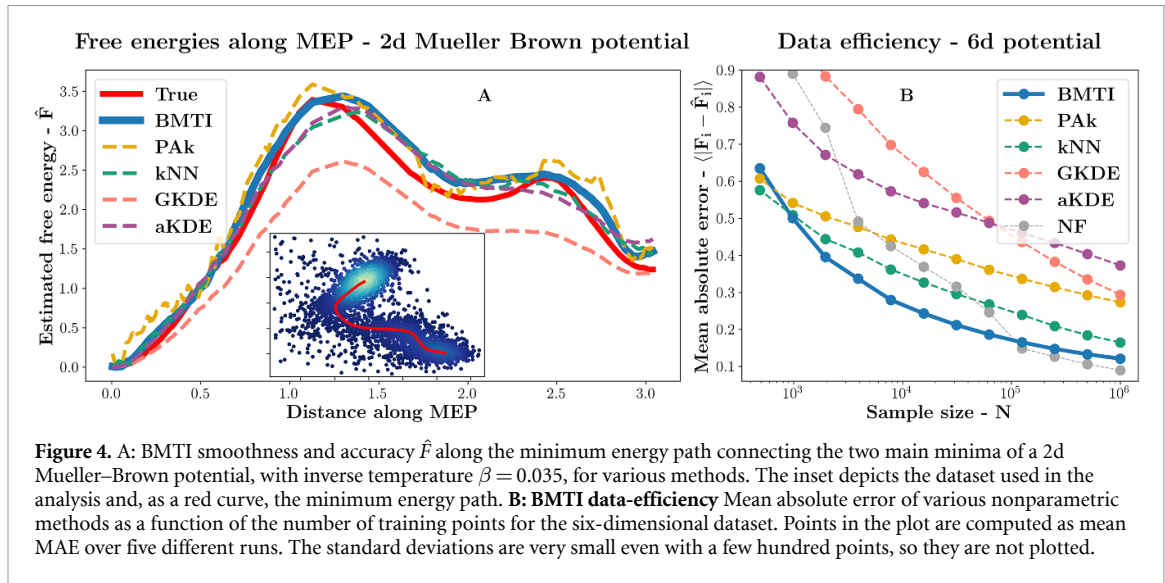
Test datasets We test the approach on various synthetic and realistic datasets, all of which are described in detail in Sec D of the SM. The synthetic datasets are sampled from known PDFs. We consider four of them, having embedding dimensions D from 2 to 9 coincident with their ID. Two of them are known analytically but display to some extent a realistic behaviour. One is the Mueller-Brown potential [98], an analytic 2-d potential designed to display the typical behaviour of metastable systems (the transitions occur through a non-trivial curved path); one is a 9-d dataset sampled from an analytic PDF obtained via Gaussian-KDE smoothing of a realistic physical chemistry dataset, of which it retains all the complexity (in fact, we regard it as ‘the hardest’ benchmark). The realistic datasets are all cases in which the true NLD is known only on the sample points, but the underlying PDF is not known analytically. They are sampled from rugged and complex landscapes taken from the molecular simulations literature [99–101]. One is a 4 dimensional dataset with ID 4. The other two have IDs 2 and 7 but they are both embedded in 20 dimensions.

Competing estimators In the case of the NLD estimators, we compare the performance of BMTI against that of other well-established nonparametric methods. For the Gaussian KDE (GKDE) class, we take as baseline the standard GKDE with Silverman’s smoothing parameter [1] and as state-of-the-art the *adaptive kernel estimator* (aKDE) [48, 49], both as implemented in [102]. For the k NN-based methods we take standard k NN with optimal global k selected via Abramson’s rule of thumb ($N^{D/D+4}$) [103] as baseline and the *point-adaptive k NN* (PAk) as state-of-the-art, both as implemented in [20].

Evaluation metrics We assess the accuracy of the estimators $\hat{\mathbf{g}}$, $\hat{\delta\mathbf{F}}$ and $\hat{\mathbf{F}}$ by looking at various metrics here described. In cases where we do not know the normalized ground truth (GT) PDF, but only the unnormalised NLD $\mathbf{F}$, we align the mean of the predictions $\hat{\mathbf{F}}$ to the mean of the GT. The first metric is the *absolute error*, or L_1 error, which quantifies the discrepancy of the estimator from the GT value as the L_1 norm. In the case of the NLD estimator at point i it reads $\epsilon_i = |F_i - \hat{F}_i|$. It is studied either as a function of the GT NLD or averaged on the whole dataset, the *mean absolute error* (MAE) $\langle \epsilon_i \rangle$. Another more qualitative but very insightful way to inspect an estimator performance is to plot the estimated estimator vs the GT value for all estimates, the *parity plot*. Additionally, we check the joint performance of estimators $\hat{y}$ of mean value y and of their uncertainty estimators $\hat{\sigma}_y$ by looking at the distribution of the standardised scores $(\hat{y} - y)/\hat{\sigma}_y$, also called the *pull distribution* [104], which is expected to be a standard Gaussian $\mathcal{N}(0, 1)$. Finally, in section F of the SM we compute and compare the Kullback–Leibler (KL) divergence from the GT to the estimated densities for the various methods.

4.1. Performance assessment and discussion

The performance of the $\hat{\delta\mathbf{F}}$ estimator for neighbouring points and its uncertainty is assessed in figure 2 on a variety of benchmark distributions. In all cases, the parity plot and the pull distribution are in very good



agreement with the predictions. This suggests that, with the pipeline described in section 3.2, and the estimators in equation (14) provide an accurate estimate of the NLD difference and, remarkably, of the associated error even in cases in which the ID is high. The exceptional quality of these estimates is a necessary condition to infer the NLD globally. Figure 2 proves empirically that, with the choice for the $\delta\hat{F}$ and $\mathbf{C}$ estimators presented in sections 3.2.3 and 3.2.4, the Gaussianity condition in equation (2), required for the BMTI algorithm, is satisfied.

The performance of the BMTI estimator is benchmarked in figures 3, 4 and in table 1. Table 1 summarises the global performance of the estimators on various datasets. BMTI is consistently the best-performing nonparametric estimator on all datasets except in the nine-dimensional case, in which it is comparable to kNN.

In the top row of figure 3, we compare the parity plot of BMTI and GKDE, which is probably the most popular density estimation method. BMTI's parity plots are slimmer and more symmetric than GKDE's, a sign of both lower variance and absence of bias. Instead, GKDE is observed to develop systematic errors as the ID of the dataset increases, underestimating NLD differences. The bottom row shows that, while other methods can be particularly accurate at specific values of F for specific datasets, BMTI consistently outperforms all of them across the whole F spectrum except occasionally at very high values.

Table 1. Performance of $\hat{F}$: MAE of various methods on six datasets. The two best performances for each dataset are highlighted in bold.

Dataset	BMTI	PAk	kNN	GKDE	aKDE
2d Gaussian	0.11	0.16	0.24	0.22	0.14
20d-A($d=2$)	0.10	0.16	7.75	0.17	4.32
2d-MBx0.035	0.12	0.16	0.30	0.36	0.25
6d potential	0.26	0.44	0.37	0.72	0.53
20d-B($d=7$)	0.36	0.52	2.87	0.80	4.34
9d smooth.	0.67	0.79	0.68	2.43	2.82

Figure 4(A) illustrates that, even in undersampled regimes, the BMTI estimates are not only accurate but also smooth, two characteristics which for instance make it particularly well-suited for free energy reconstructions in physical chemistry applications. In this context, it is necessary to estimate the derivatives of thermodynamic potentials such as the free energy, so a spiky estimated free energy profile would yield unphysically high forces. In fact, it is among the most accurate in capturing the NLD differences between the main minimum, where the path starts, and the highest saddle point or the second-lowest minimum. While comparably accurate, k NN-based methods are visibly rougher and noisier. On the other hand, GKDE is smooth but again, it underestimates NLD differences. While aKDE looks preferable to GKDE in this case, by looking at table 1 and figure 3 we observe that standard GKDE is more robust than aKDE when D increases.

The data-efficiency of the various methods is compared in figure 4(B) by tracking the MAE as the sample size increases on the 6 dimensional dataset. As a reference to another state-of-the-art method, despite not being a direct competitor, we also include a modern real NVP NFs model [58] as implemented in [105]. We train the NF for 20 epochs, so that it approaches convergence; we did not push the number of epochs further because, already with this choice, the computation times for NFs are one-to-two orders of magnitude larger than for BMTI. BMTI is the best performing of all the competing methods, maintaining a low MAE for all sample sizes. While BMTI shows a better performance in an undersampling regime, the big NF model outperforms BMTI for very large samples.

By looking at all the benchmarks, we observe that the methods restricting to the intrinsic manifold (BMTI and PAK) are the only ones consistently among top performers. Standard k NN is competitive only in cases in which the ID is close or equal to D , i.e. precisely when it computes volumes on the intrinsic manifold. All other methods break down as D increases. The only exception is GKDE in the 20d-A potential, where it curiously seems to implicitly restrict to the intrinsic manifold despite not having any notion of it.

Finally, figure 5(A) compares the time scaling as a function of the number of points N for all the various algorithms. BMTI uses the conjugate gradient sparse iterative solver, which shows a slight deviation from linear time scaling (as discussed in section C.5 of the SM). The GKDEs scale quadratically, while all the other kernel methods scale linearly—although PAK with a prefactor at least one order of magnitude larger than standard k NN. Finally, training NFs has a linear cost but is computationally one-to-two orders of magnitude slower than BMTI.

By analysing this in conjunction with the other performance metrics, we can conclude that BMTI stands out as a method of choice:

- (i) for datasets up to moderate sizes of $\mathcal{O}(10^5)$ points, but can be preferred to deep neural network estimators even for larger sizes, when time efficiency is important;
- (ii) for multimodal densities (since in trivial cases simple parametric models might be a better choice);
- (iii) when a quantitative control over the accuracy and smoothness on the log-density is deemed important (this is the case, for example, in many physical chemistry or ML applications).

The performance of a regularised BMTI estimator is benchmarked in figure 5(B) for the case in which the regularising log-likelihood in equation (27) is that of the k NN-based PAK estimator from [35]. The benchmark dataset is obtained from the Mueller–Brown potential with an inverse temperature $\beta = 0.07$, which is double w.r.t. the one used in figure 4(A), $\beta = 0.035$. In this case, the three main wells of the potential are sampled separately, since the saddle points on the transition paths between the basins are not populated due to the small size of the sample and the low temperature. A strictly-local method, such as the PAK estimator, is not affected by the disconnectedness of the NG, since for each point the NLD estimate $\hat{F}_i$ is independent; therefore the parity plot results in a single connected cloud. The parity plot of the BMTI estimator, which is global, presents instead three disconnected clouds (figure 5(B2)). By mixing the likelihoods for the two estimators as in equation (27), the parity plot becomes unimodal (figure 5(B3)), and it is actually even slimmer than the case of PAK alone, especially for high density values. This improved performance is even clearer by looking at panel B4 of the same figure, in which the MAE is plotted as a

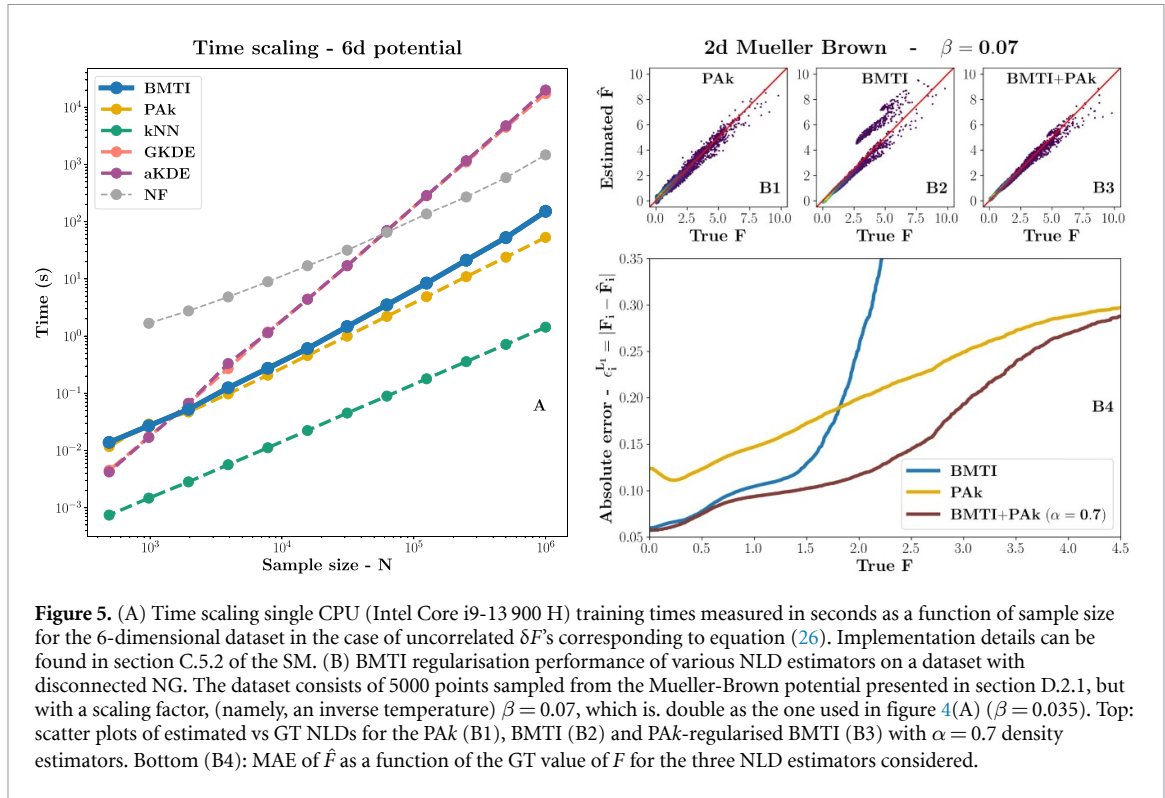


Figure 5. (A) Time scaling single CPU (Intel Core i9-13 900 H) training times measured in seconds as a function of sample size for the 6-dimensional dataset in the case of uncorrelated δF 's corresponding to equation (26). Implementation details can be found in section C.5.2 of the SM. (B) BMTI regularisation performance of various NLD estimators on a dataset with disconnected NG. The dataset consists of 5000 points sampled from the Mueller-Brown potential presented in section D.2.1, but with a scaling factor, (namely, an inverse temperature) $\beta = 0.07$, which is, double as the one used in figure 4(A) ($\beta = 0.035$). Top: scatter plots of estimated vs GT NLDs for the PAK (B1), BMTI (B2) and PAK-regularised BMTI (B3) with $\alpha = 0.7$ density estimators. Bottom (B4): MAE of $\hat{F}$ as a function of the GT value of F for the three NLD estimators considered.

function of the GT NLD: the regularised BMTI estimator outperforms PAK basically across the whole range. Note that even very small values of the mixing scalar α in equation (27), such as 10^{-2} , or even 10^{-3} , are enough to regularise the BMTI estimator and heal it from the disconnectedness. Moreover, this regularisation corrects the singularity of the matrix $\mathbf{A}$ in equation (6), as discussed in section C.5.2 of the SM.

5. Conclusions

We have presented BMTI, a nonparametric data-efficient method to estimate smooth log-density landscapes even in high dimensional spaces. BMTI is based on a first estimation of log-density differences between neighbouring points, and a subsequent (implicit) integration of such differences. Its key ingredients are the possibility of restricting its operation to the intrinsic data manifold, its point-adaptive nature and a rigorous error control. Such features make BMTI accurate and efficient, as we demonstrate through numerical experiments. These characteristics make it suitable candidates for physical or ML applications [106, 107]. Moreover, we stress that this work also introduces an adaptive nonparametric log-density gradient estimator and a graph-based integration procedure, which can be seen as interesting stand-alone methods.

Importantly, BMTI can be made robust against cases in which the NG is disconnected, a situation in which thermodynamic integration schemes would be doomed to fail. To avoid this pitfall, the BMTI log-likelihood can be mixed additively with the log-likelihood of a strictly-local and normalised NLD estimator.

Finally, the BMTI framework also naturally incorporates a way to estimate the log-density uncertainties. The theoretical derivation is rigorous under the assumption of Gaussian noise (equation (2)), but the numerical implementation is approximate (we estimate the inverse of the covariance matrix assuming the inverse is diagonal-dominated). Much of our current effort is in the direction of finding better approximations and numerical solutions to estimate this inverse.

Acknowledgments

This work is supported in part by funds from the European Union's Horizon 2020 research and innovation program (Grant Number 824143, MaX 'Materials design at the eXascale' Centre of Excellence) and by NextGenerationEU through the Italian National Centre for HPC, Big Data, and Quantum Computing (Grant Number CN00000013). The authors would like to thank S De Gironcoli (SISSA), J. Henin (CNRS), F. Marinelli (MCW), T D Swinburne (CNRS), M C Marinica (CEA), F Pellegrini (SISSA), Claudia Biancotti (BdI), Lorenzo Fant (OGS), Laura Zichi (Harvard), Iuri Macocco (UPF) and Claudio Leone (ICTP) for

helpful discussions and feedback. The authors are grateful to one of the anonymous reviewers for their important suggestion to explore the use of efficient solvers for the BMTI linear system.

The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of Banca d'Italia.

Code and data availability statement

Our code for the BMTI, kNN and PAk density estimators is openly available in DADapy [20] at <https://github.com/sissa-data-science/DADapy>.

All datasets that support the findings of this study, described in section D of the SM, are openly available on Zenodo at <https://doi.org/10.5281/zenodo.15002318> [108].

ORCID iDs

Matteo Carli  <https://orcid.org/0000-0002-8860-3062>

Alex Rodriguez  <https://orcid.org/0000-0002-0213-6695>

Alessandro Laio  <https://orcid.org/0000-0001-9164-7907>

Aldo Glielmo  <https://orcid.org/0000-0002-4737-2878>

References

- [1] Silverman B W 1986 *Density Estimation for Statistics and Data Analysis* vol 26 (CRC Press)
- [2] David W S 2015 *Multivariate Density Estimation: Theory, Practice and Visualization* 2nd edn (Wiley) (<https://doi.org/10.1002/9780470316849>)
- [3] Glielmo A, Husic B E, Rodriguez A, Clementi C, Noé F and Laio A 2021 Unsupervised learning methods for molecular simulation data *Chem. Rev.* **121** 9722–58
- [4] Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning Data Mining, Inference and Prediction* 2nd edn (Springer) (<https://doi.org/10.1007/978-0-387-84858-7>)
- [5] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the em algorithm *J. R. Stat. Soc. B* **39** 1–22
- [6] Bengio Y, Courville A and Vincent P 2013 Representation learning: a review and new perspectives *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1798–828
- [7] Lecun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44
- [8] Schmidhuber J 2015 Deep learning in neural networks: An overview *Neural Netw.* **61** 85–117
- [9] Christopher M B 2006 *Pattern Recognition and Machine Learning* 1st edn (Springer) (<https://doi.org/10.5555/1162264>)
- [10] Julian Izenman A 1991 Recent Developments in Nonparametric Density Estimation *J. Am. Stat. Assoc.* **86** 205
- [11] Izenman A J 2008 *Modern Multivariate Statistical Techniques* vol 1 (Springer)
- [12] Parzen E 1962 On the estimation of probability density functions and mode *Ann. Math. Stat.* **33** 1065–76
- [13] Fix E and Hodges J L 1951 Discriminatory analysis. nonparametric discrimination: consistency properties *Report 4* (Project Number 21-49-004) (USAF School of Aviation Medicine, Randolph Field, Texas) (<https://doi.org/10.2307/1403797>)
- [14] Bellman R 1961 *Adaptive Control Processes: A Guided Tour* (Princeton University Press) (<https://doi.org/10.1515/9781400874668>)
- [15] Jerome H F 1997 On bias, variance, 0/1-loss and the curse-of-dimensionality *Data Min. Knowl. Discov.* **1** 55–77
- [16] Charles J S 1984 An asymptotically optimal window selection rule for kernel density estimates *Ann. Stat.* **12** 1285–97
- [17] Turlach B A 1993 Bandwidth selection in kernel density estimation: A review *CORE and Institut de Statistique* (available at: <https://ideas.repec.org/p/wop/humbse/9307.html>)
- [18] Heidenreich N-B, Schindler A and Sperlich S 2013 Bandwidth selection for kernel density estimation: a review of fully automatic selectors *AStA Adv. Stat. Anal.* **97** 403–33
- [19] Fukunaga K and Larry D H 1975 The estimation of the gradient of a density function, with applications in pattern recognition *IEEE Trans. Inf. Theory* **21** 32–40
- [20] Glielmo A et al 2022 Dadapy: Distance-based analysis of data-manifolds in python *Patterns* **3** 100589
- [21] Pope P E, Zhu C, Abdelkader A, Goldblum M and Goldstein T 2021 The intrinsic dimension of images and its impact on learning (arXiv:2104.08894)
- [22] Korn F, Pagel B-U and Faloutsos C 2001 On the “dimensionality curse” and the “self-similarity blessing” *IEEE Trans. Knowl. Data Eng.* **13** 96–111
- [23] Beygelzimer A, Kakade S and Langford J 2006 Cover trees for nearest neighbor *Proc. 23rd Int. Conf. on Machine Learning (ICML'06)* (Association for Computing Machinery) pp 97–104
- [24] Maccoco I, Glielmo A, Grilli J and Laio A 2023 Intrinsic dimension estimation for discrete metrics *Phys. Rev. Lett.* **130** 067401
- [25] Facco E, Pagnani A, Tea Russo E and Laio A 2019 The intrinsic dimension of protein sequence evolution *PLoS Comput. Biol.* **15** e1006767
- [26] Konz N, Hanxue G, Dong H and Mazurowski M A 2022 The intrinsic manifolds of radiological images and their role in deep learning *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* (Springer) pp 684–94
- [27] Varghese A, Santos-Fernandez E, Denti F, Mira A, and Mengersen K 2022 On the intrinsic dimensionality of covid-19 data: a global perspective (arXiv:2203.04165)
- [28] Ansuini A, Laio A, Macke J H and Zoccolan D 2019 Intrinsic dimension of data representations in deep neural networks *Advances in Neural Information Processing Systems* p 32
- [29] Camastra F and Staiano A 2016 Intrinsic dimension estimation: advances and open problems *Inf. Sci.* **328** 26–41
- [30] Facco E, D'Errico M, Rodriguez A and Laio A 2017 Estimating the intrinsic dimension of datasets by a minimal neighborhood information *Sci. Rep.* **7** 1–11
- [31] Erba V, Gherardi M and Rotondo P 2019 Intrinsic dimension estimation for locally undersampled data *Sci. Rep.* **9** 1–9

- [32] Denti F, Doimo D, Laio A and Mira A 2022 The generalized ratios intrinsic dimension estimator *Sci. Rep.* **12** 20005
- [33] Bac J, Mirkes E M, Gorban A N, Tyukin I and Zinovyev A 2021 Scikit-dimension: a python package for intrinsic dimension estimation *Entropy* **23** 1368
- [34] Ozakin A and Gray A 2009 Submanifold density estimation *Advances in Neural Information Processing Systems* vol 22, ed Y Bengio, D Schuurmans, J Lafferty, C Williams and A Culotta (Curran Associates, Inc.) (available at: <https://dl.acm.org/doi/abs/10.5555/2984093.2984248>)
- [35] Rodriguez A, D'Errico M, Facco E and Laio A 2018 Computing the free energy without collective variables *J. Chem. Theory Comput.* **14** 1206–15
- [36] Liu Q, Jiaze X, Jiang R and Hung Wong W 2021 Density estimation using deep generative neural networks *Proc. Natl Acad. Sci.* **118** e2101344118
- [37] Horvat C and Pfister J-P 2023 Density estimation on low-dimensional manifolds: an inflation-deflation approach *J. Mach. Learn. Res.* **24** 61–61
- [38] Kobayev I, Prince S J D and Brubaker M A 2020 Normalizing flows: an introduction and review of current methods *IEEE Trans. Pattern Anal. Mach. Intell.* **43** 3964–79
- [39] Fukunaga K 1990 *Introduction to Statistical Pattern Recognition* (Academic) (<https://doi.org/10.5555/92131>)
- [40] Park B U and Marron J S 1990 Comparison of data-driven bandwidth selectors *J. Am. Stat. Assoc.* **85** 66–72
- [41] Pearson K 1894 Contributions to the mathematical theory of evolution *Phil. Trans. R. Soc. A* **185** 71–110
- [42] Silverman B W and Jones M C 1989 E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951) *Int. Stat. Rev.* **57** 233
- [43] Rosenblatt M 1956 Remarks on some nonparametric estimates of a density function *Ann. Math. Stat.* **27** 832–7
- [44] Cacoullos T 1964 Estimation of a Multivariate Density. *Tech. Report; No. 40.* (University of Minnesota, Department of Statistics) (available at: <https://hdl.handle.net/11299/199362>)
- [45] Breiman L, Meisel W and Purcell E 1977 Variable kernel estimates of multivariate densities *Technometrics* **19** 135–44
- [46] Chris Jones M, Marron J S and Sheather S J 1996 A brief survey of bandwidth selection for density estimation *J. Am. Stat. Assoc.* **91** 401–7
- [47] Zhao P and Lai L 2022 Analysis of knn density estimation *IEEE Trans. Inf. Theory* **68** 7971–95
- [48] Abramson I S 1982 On bandwidth variation in kernel estimates-a square root law *Ann. Stat.* **10** 1217–23
- [49] Abramson I S 1982 Arbitrariness of the pilot estimator in adaptive kernel methods *J. Multivar. Anal.* **12** 562–7
- [50] Fukunaga K and Flick T E 1982 A parametrically-defined nearest neighbor distance measure *Pattern Recognit. Lett.* **1** 3–5
- [51] Peter H and Marron. J S 1988 Choice of kernel order in density estimation *Ann. Stat.* **16** 161–73
- [52] Myles J P and Hand D J 1990 The multi-class metric problem in nearest neighbour discrimination rules *Pattern Recognit.* **23** 1291–7
- [53] Ord J K 1978 How many trees in a forest *Math. Sci.* **3** 23–33
- [54] Polianskii V, Luca Marchetti G, Kravberg A, Varava A, Pokorniy F T and Kragic D 2022 Voronoi density estimator for high-dimensional data: Computation, compactification and convergence *Proc. 38th Conf. on Uncertainty in Artificial Intelligence (Proc. of Machine Learning Research* vol 180), ed J Cussens and K Zhang (PMLR) pp 1644–53 (available at: <https://proceedings.mlr.press/v180/polianskii22a.html>)
- [55] Luca Marchetti G, Polianskii V, Varava A, Pokorniy F T and Kragic D 2023 An efficient and continuous voronoi density estimator *Int. Conf. on Artificial Intelligence and Statistics* (PMLR) pp 4732–44 (available at: <https://proceedings.mlr.press/v206/marchetti23a.html>)
- [56] Kingma D P and Welling M 2013 Auto-encoding variational Bayes (arXiv:1312.6114)
- [57] Goodfellow I, Pouget-Abadie J, Mirza M, Bing X, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Information Processing Systems* p 27
- [58] Dinh L, Sohl-Dickstein J and Bengio S 2016 Density estimation using real nvp (arXiv:1605.08803)
- [59] Papamakarios G, Pavlakou T and Murray I 2017 Masked autoregressive flow for density estimation *Advances in Neural Information Processing Systems* p 30
- [60] Liu Q, Jiaze X, Jiang R and Hong Wong W 2021 Density estimation using deep generative neural networks *Proc. Natl Acad. Sci. USA* **118** e2101344118
- [61] Hyvärinen A 2005 Estimation of non-normalized statistical models by score matching *J. Mach. Learn. Res.* **6** 695–709 (available at: <https://jmlr.org/papers/v6/hyvarinen05a.html>)
- [62] Marc'Aurelio Ranzato Y-L B, Chopra S and LeCun Y 2007 A unified energy-based framework for unsupervised learning *Proc. 11th Int. Conf. on Artificial Intelligence and Statistics (Proc. of Machine Learning Research* vol 2) (San Juan, Puerto Rico) ed M Meila and X Shen (PMLR) pp 371–9 (available at: <https://proceedings.mlr.press/v2/ranzato07a.html>)
- [63] Pietrucci F 2017 Strategies for the exploration of free energy landscapes: unity in diversity and challenges ahead *Rev. Phys.* **2** 32–45
- [64] Ester M, Kriegel H-P, Sander J, Xiaowei X et al 1996 A density-based algorithm for discovering clusters in large spatial databases with noise *kdd* vol 96 pp 226–31
- [65] Comaniciu D and Meer P 1999 Mean shift analysis and applications *Proc. 7th IEEE Int. Conf. Computer Vision* vol 2 (IEEE) pp 1197–1203
- [66] Yang C, Duraiswami R, DeMenthon D and Davis L 2003 Mean-shift analysis using quasineutron methods *Proc. 2003 Int. Conf. on Image Processing (Cat. No. 03CH37429)* vol 2 (IEEE) p II–447
- [67] LeCun Y, Chopra S, Hadsell R, Ranzato A and Huang F J 2006 A tutorial on energy-based learning
- [68] Kulczycki P and Charytanowicz M 2010 A complete gradient clustering algorithm formed with kernel estimators *Int. J. Appl. Math. Comput. Sci.* **20** 123–34
- [69] Rodriguez A and Laio A 2014 Clustering by fast search and find of density peaks *Science* **344** 1492–6
- [70] d'Errico M, Facco E, Laio A and Rodriguez A 2021 Automatic topography of high-dimensional data sets by non-parametric density peak clustering *Inf. Sci.* **560** 476–92
- [71] Tuckerman M 2010 *Statistical Mechanics: Theory and Molecular Simulation* 2nd edn (Oxford University Press) (<https://doi.org/10.1093/oso/9780198825562.002.0003>)
- [72] Laio A and Parrinello M 2002 Escaping free-energy minima *Proc. Natl Acad. Sci.* **99** 12562–6
- [73] Weinan E and Vanden-Eijnden E 2004 Metastability, conformation dynamics and transition pathways in complex systems *Multiscale Modelling and Simulation* (Springer) pp 35–68
- [74] Chipot C and Pohorille A 2007 *Free Energy Calculations Theory and Applications in Chemistry and Biology* 1 edn (Springer) (<https://doi.org/10.1007/978-3-540-38448-9>)

- [75] Saremi S, Mehrjou A, Scholkopf B and Hyvärinen A 2018 Deep energy estimator networks (arXiv:1805.08306)
- [76] Whye Teh Y, Welling M, Osindero S and Hinton G E 2003 Energy-based models for sparse overcomplete representations *J. Mach. Learn. Res.* **4** 1235–60 (available at: <https://www.jmlr.org/papers/v4/teh03a.html>)
- [77] LeCun Y, Huang F J and Bottou L 2004 Learning methods for generic object recognition with invariance to pose and lighting *Proc. 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2004 CVPR 2004* vol 2 (IEEE) p II–104
- [78] Kirkwood J G 1935 Statistical mechanics of fluid mixtures *J. Chem. Phys.* **3** 300–13
- [79] Dehez F, Tarek M and Chipot C 2007 Energetics of ion transport in a peptide nanotube *J. Phys. Chem. B* **111** 10633–5
- [80] Ciccotti G, Kapral R and Vanden-Eijnden E 2005 Blue Moon sampling, vectorial reaction coordinates and unbiased constrained dynamics *ChemPhysChem* **6** 1809–14
- [81] Hénin J, Fiorin G, Chipot C and Klein M L 2010 Exploring multidimensional free energy landscapes using time-dependent biases on collective variables *J. Chem. Theory Comput.* **6** 35–47
- [82] Comer J, Gumbart J C, Hénin J, Lelièvre T, Pohorille A and Chipot C 2015 The adaptive biasing force method: everything you always wanted to know but were afraid to ask *J. Phys. Chem. B* **119** 1129–51
- [83] Marinova V and Salvalaglio M 2019 Time-independent free energies from metadynamics via mean force integration *J. Chem. Phys.* **151** 164115
- [84] Lelièvre T, Rousset M and Stoltz G 2010 *Free Energy Computations A Mathematical Perspective* (Imperial College Press) (<https://doi.org/10.1142/p579>)
- [85] Alrachid H and Lelièvre T 2015 Long-time convergence of an adaptive biasing force method: variance reduction by Helmholtz projection *SMAI J. Comput. Math.* **1** 55–82
- [86] Hénin J 2021 Fast and accurate multidimensional free energy integration *J. Chem. Theory Comput.* **17** 6789–98
- [87] Abrams C F and Vanden-Eijnden E 2010 Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics *Proc. Natl Acad. Sci.* **107** 4961–6
- [88] Chen M, Cuendet M A and Tuckerman M E 2012 Heating and flooding: a unified approach for rapid generation of free energy surfaces *J. Chem. Phys.* **137** 024102
- [89] Maragliano L and Vanden-Eijnden E 2008 Single-sweep methods for free energy calculations *J. Chem. Phys.* **128** 1–10
- [90] Marinelli F and Faraldo-Gómez J D 2021 Force-correction analysis method for derivation of multidimensional free-energy landscapes from adaptively biased replica simulations *J. Chem. Theory Comput.* **17** 6775–88
- [91] Hestenes M R and Stiefel E 1952 Methods of conjugate gradients for solving linear systems *J. Res. Natl Bur. Stand.* **49** 409–36
- [92] Penrose R 1956 On best approximate solutions of linear matrix equations *Math. Proc. Camb. Phil. Soc.* **52** 17–19
- [93] Carli M and Laio A 2021 Statistically unbiased free energy estimates from biased simulations *Mol. Phys.* **119** e1899323
- [94] Comaniciu D and Meer P 2002 Mean shift: A robust approach toward feature space analysis *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 603–19
- [95] Jaccard P 1912 The distribution of the flora in the alpine zone. 1 *New Phytol.* **11** 37–50
- [96] Virtanen P et al (SciPy 1.0 Contributors) 2020 SciPy 1.0: fundamental algorithms for scientific computing in python *Nat. Methods* **17** 261–72
- [97] Carli M 2022 Nonparametric density estimation methods and applications to molecular simulations (available at: <https://hdl.handle.net/20.500.11767/127569>)
- [98] Müller K and Brown L D 1979 Location of saddle points and minimum energy paths by a constrained simplex optimization procedure *Theor. Chim. Acta* **53** 75–93
- [99] Granata D, Camilloni C, Vendruscolo M and Laio A 2013 Characterization of the free-energy landscapes of proteins by nmr-guided metadynamics *Proc. Natl Acad. Sci.* **110** 6817–22
- [100] Baftizadeh F, Pietrucci F, Biarnés X and Laio A 2013 Nucleation process of a fibril precursor in the c-terminal segment of amyloid- β *Phys. Rev. Lett.* **110** 168103
- [101] Zerze G H, Miller C M, Granata D and Mittal J 2015 Free energy surface of an intrinsically disordered protein: comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics *J. Chem. Theory Comput.* **11** 2776–82
- [102] Menne T and Walmsley M 2022 Adaptive width kde with Gaussian kernels (available at: <https://github.com/mennthor/awkde>)
- [103] Ian S A 1984 Adaptive density flattening—a metric distortion principle for combating bias in nearest neighbor methods *Ann. Stat.* **12** 880–6
- [104] Demortier L and Lyons L 2002 Everything you always wanted to know about pulls *CDF Note* (available at: https://lucdemortier.github.io/assets/papers/cdf5776_pulls.pdf)
- [105] Stimper V, Liu D, Campbell A, Berenz V, Ryll L, Schölkopf B and Miguel Hernández-Lobato J 2023 normflows: A PyTorch package for normalizing flows *J. Open Source Softw.* **8** 5361
- [106] Yilun D and Mordatch I 2019 Implicit generation and modeling with energy based models *Advances in Neural Information Processing Systems* p 32
- [107] Hénin J, Lelièvre T, Shirts M R, Valsson O and Delemotte L 2022 Enhanced sampling methods for molecular dynamics simulations [article v1.0] *Living J. Comput. Mol. Sci.* **4** 1583
- [108] Carli M, Rodríguez A, Laio A and Glielmo A 2025 Datasets to reproduce the results in the paper “density estimation via binless multidimensional integration” by Carli M and Glielmo A and Rodríguez A and Laio A *Zenodo* (<https://doi.org/10.5281/zenodo.15002318>)