Unione europea
Fondo sociale europeo

REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA

FSE 2014-2020
POR SARDEGNA

University of Cagliari   University of Sassari

PhD in Chemical Sciences and Technologies

Cycle XXXII

Discrete coarse-grained modelling of adsorption and diffusion in host-guest systems

Scientific disciplinary sector

CHIM/02

Presented by:	Giovanni Pireddu
PhD Coodinator:	Prof. Stefano Enzo
Tutors:	Prof. Pierfranco Demontis
	Dr. Federico G. Pazzona

Final exam academic year 2018 – 2019

Thesis defended during the session of February 2020

To my family,

my rock-solid and heartwarming sanctuary.

Abstract

Representing molecular systems above the microscale is a challenging task. The widely-used atomistic methods are very accurate, but at the same time, very limited in terms of efficiency. In this thesis, I report different methodologies to represent adsorption and diffusion occurring in host-guest systems on larger scales, through discrete models. First, I report a data-driven approach for the definition of molecular states based on local atomistic patterns. Second, I propose another method that makes use of the occupancies i.e. local amounts of guest species. Molecular systems are mapped into lattice models equipped with coarse-grained thermodynamics and a local operator, which represents the dynamics. These methods are validated in different ways on several molecular systems, and provide an accurate reproduction of the reference atomistic properties. Moreover, they unveiled interesting physicochemical insights while being strikingly more efficient than their atomistic counterpart.

Acknowledgements

Financial support

I gratefully acknowledge the Sardinian Regional Government for the financial support of my PhD scholarship (P.O.R. Sardegna F.S.E. - Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014-2020, Axis III Education and training, Thematic goal 10, Investment priority 10ii), Specific goal 10.5.

Academic staff

I would like to thank my supervisor Prof. Pierfranco Demontis for advising me through this path in different ways: scientifically, professionally and personally. I also thank him for the support in pursuing my scientific curiosities that arose during this period. I am profoundly grateful to my co-supervisor Dott. Federico G. Pazzona for mentoring me constantly. I have learnt a great deal from our scientific and general conversations. I am thankful for his availability and kind help during both great and terrible times. I owe to Federico what I learned about approaching scientific problems, because of his analytical and creative mentality. I thank my other office colleagues, Andrea, Alberto and Marcello, for all the conversations we had together, and for being very supportive during the last years.

I would like to thank Prof. Michele Ceriotti for welcoming me in the COSMO group in Lausanne, where I learned a lot of new techniques and broadened my horizons as a scientist. I am very grateful to Prof. Rocio Semino for mentoring me during my Swiss visit and for being very supportive from several points of view.

I am also thankful to Prof. dr hab. Magdalena Załuska-Kotur for welcoming me in her research group, for her help, support and for the marvellous collaboration we built together. My stay at IF-PAN in Warsaw was also amazing because of the other office friends, with whom I spent great coffee breaks and lunches together.

Contents

Preface **9**

1 Introduction **10**

1.1 The representation problem: keeping what matters 10

1.2 Coarse-graining strategies 16

1.3 Outline of the thesis 19

2 Theoretical Background **20**

2.1 Molecular representations 20

2.1.1 Classical atomistic representation 20

2.1.2 Local occupancy-based representations 22

2.1.3 Representations for machine learning 23

2.2 Simulation methods 26

2.2.1 Statistical mechanics background 26

2.2.2 Monte Carlo 30

2.2.3 Molecular dynamics 33

2.3 Materials studied in this thesis 37

2.3.1 ITQ-29 37

2.3.2 LTA-ZTC 38

2.3.3 Graphene and related materials 40

3 Machine learning molecular states: adsorption patterns of CO_2 in zeolites 42

3.1 Introduction . . . 42
3.2 Smooth overlap of atomic positions . . . 43
3.3 Dimensionality reduction . . . 45
3.4 Probabilistic analysis of molecular motifs . . . 46
3.5 Results . . . 48
3.5.1 Molecular patterns . . . 48
3.5.2 Static properties . . . 50
3.5.3 Backmapping and network model construction . . . 52
3.6 Conclusions and perspectives . . . 54

4 Local free energy approximations for the coarse-graining of adsorption phenomena: the interacting pair approximation 57

4.1 Introduction . . . 57
4.2 Local, Coarse-grained interactions . . . 62
4.3 Coarse-graining under the Interacting Pair Approximation . . . 66
4.4 Comparison with previous models . . . 74
4.5 Simulations and discussion . . . 76
4.5.1 Lattice-gas with repulsive interactions . . . 76
4.5.2 Lennard-Jones particles under the influence of an external field . . . 82
4.6 Conclusions . . . 87
4.7 Connection between fine-grained and coarse-grained global occupancy distribution . . . 89
4.8 Estimation of the interaction terms . . . 91
4.8.1 'One-chemical-potential-at-a-time' (OCT) . . . 91
4.8.2 'Choose-the-best-ratio' (CBR) . . . 92

4.8.3 Probability threshold 92

5 Spatial Coarse-graining of Methane Adsorption in Graphene Materials 94

5.1 Introduction 94

5.2 Coarse-grained model 96

5.2.1 Data pre-processing at low T 102

5.3 Results and discussion 104

5.3.1 Methane on single layer graphene. 107

5.3.2 Methane between two graphene layers. 109

5.4 Conclusions 114

6 Scaling-up Simulations of Diffusion in Microporous Materials 116

6.1 Introduction 116

6.2 Methods 119

6.2.1 Coarse-grained structure 120

6.2.2 Thermodynamics 122

6.2.3 Elementary events for diffusion 123

6.2.4 Dynamical correlations 127

6.2.5 Numerical simulations 129

6.3 Results and discussion 133

6.3.1 Jump rates modelling 133

6.3.2 Dynamical correlations 136

6.3.3 Diffusivity 138

6.3.4 Decay of occupancy correlations 142

6.3.5 Computational speedup 144

6.4 Conclusions 146

7 Conclusions and future perspectives 149

7.1 An *a posteriori* overview 149

7.2 Future perspectives 151

A Supplementary material for Chapter 4 154

A.1 Occupancy distributions for the lattice-gas system 154

A.2 Lennard-Jones system 158

A.2.1 Coarse-graining of a lone pair of pores under the NIPA . . . 158

A.2.2 Coarse-graining of the full LJ systems 159

A.3 Estimation of the interaction terms — Missing entries in the CG interaction matrix 159

B Supplementary material for Chapter 6 164

B.1 Sensitivity tests 164

B.2 Fitted models for the coarse-grained simulations 167

B.2.1 ITQ-29 168

B.2.2 ZTC 168

B.3 Static properties 169

B.4 Boltzmann-Matano simulations of the ITQ-29 system 173

Preface

This thesis summarizes the developments and the results obtained by me and my collaborators in the context of the main project of my PhD program. During the last three years, I tried to contribute to the challenge of representing real-world molecular systems via theoretical and computational tools. The path that conducted to the results and the progress reported in this work was never straight. Many of the approaches I explored are not present in this thesis, since they didn't lead to consistent or generally relevant results. The remaining ones were (from slightly to radically) modified during the research process and eventually led to something scientifically relevant.

The project on which I report in this thesis is far from being complete, and it led to more questions rather than answers. However, I think that the works presented here have a scientific value up to some extent. In some cases, such value was also recognized by the community and it resulted in the publication of peer-reviewed articles.

Apart from its scientific value, this path was particularly enriching for me both from a professional and a personal point of view. At the time of this writing, I find myself very different from three years ago. The transformation that this experience induced in me was not painless, but eventually led to satisfaction and a sense of personal growth.

Chapter 1

Introduction

1.1 The representation problem: keeping what matters

"Everything should be made as simple as possible, but not simpler."

– A. Einstein, 1950

Understanding and rationalizing the fundamental properties of a chosen system is probably the main focus of any researcher in natural sciences. Specifically, theoretical studies attempt to describe nature by using approaches from the formal sciences such as mathematics and logic [1]. Formalizing real-world observations through a set of abstract *representations* and *laws* is the main way to rationalize the phenomena and properties of interest.

The discovery and validation of theoretical models is also accelerated by the extensive use of computer simulations of the phenomena of interest. Computational results along with experimental observations provide a further reference that can be used to test and improve theory and models. The reliability of computer simulations strictly depends on which theoretical representation of the physical systems under study is chosen to be used.

However, finding the appropriate way to formalize real-world phenomena is not straightforward and it is the subject of a considerable number of studies. In principle, one could choose the most detailed and thus accurate formalization possible as the most complete description available. Unfortunately, in many scenarios, this may not be the optimal choice for different reasons: because such formalization is simply not available in that specific context, or because it is practically impossible to use it. The first case usually originates from the presence of a theoretical gap, which hinders the modeler from applying the chosen theory to a specific problem. The second case — the one that this thesis tries to cope with — is generally caused by the overwhelming amount of information that the most complete description would require. So, despite being available, the most complete representation may carry too many details which might be unnecessary or redundant for the specific problem we are facing. This is, for example, the case of large-scale phenomena, which usually cannot be analysed starting from first principles, since that would simply be an unfeasible task.

At this point, the need for a more effective and efficient representation becomes evident. The shift from a high-detailed to a less-detailed description can be imagined as an *abstraction* process. One could imagine reiterating such a representation shift an arbitrarily large number of times, thus obtaining a set of different possible descriptions of the same system.

We can metaphorically think about it as a set of different lenses, each one with a specific resolution or magnification power. Imagine having to analyse a fruit, say an apple. If we observe it through the highest resolution lens we have, we will be able to see a large number of details and to fully characterize any portion of it. This kind of representation turns out to be very useful if we are interested in phenomena occurring in the *microscale*, such as biochemical reactions. But as humans are limited, we can carry a limited amount of information in our memory and our hardware, and we cannot even think of describing an entire apple by tracking every single atom of it.

At the opposite side, we can use the lowest-resolution lens and see the apple as a point. This kind of description may be useful to a process engineer who is interested in predicting the apple flow in a packaging factory. This is a very lightweight description which may seem to be the optimal choice in terms of memory, but might be too abstract to be useful. At this resolution, an apple would be indistinguishable from an orange and such classification error could be unacceptable for the kind of analysis we want to perform.

This metaphor serves to make the reader aware of the representation problem and of the many questions that arise from it. In my personal opinion, the most important ones can be resumed as the following:

- which criteria should we use to define the derived representations? — How do we construct different lenses?
- which representation should we use to cope with a specific problem? — How do we choose the appropriate pair of lenses in a specific context?

Both of the questions are still open, in the sense that there is no general answer. However, when defining the mapping from a given representation into another we could follow some general indications, such as the fact that a derived representation should be as consistent as possible with the original one. It is also desirable that such mappings are as general as possible, i.e. we should be able to reduce to a minimum the case-specific assumptions and decisions that could bias the definition of the low-resolution representation.

Up to now, I implicitly suggested that the shift should be performed from a high-detailed representation to a lower-detailed representation and not the opposite. This is denoted as a *bottom-up* approach, while the opposite shift is called *top-down* approach. The choice of following a bottom-up approach rather than a top-down modelling is motivated by very simple and general arguments. In the bottom-up approach, we use a set of pre-existing information which we assume to be true, from which we carry out a selection of relevant features in order to obtain a derived set of information. Practically, this can be pictured as an interpolation problem, as we define the new set of information within the preexisting set that we already possess. In the reverse approach, we would have to generate from scratch a set of details we do not have. This is a way more difficult problem and it can be imagined as an extrapolation process.

Furthermore, when defining the derived representations of the chosen system, we expect to obtain a match between the physical properties at both levels of description. In other words, we can measure the *accuracy* by comparing the properties of the derived representations with the *reference* ones. Choosing a top-down approach means that we have to start from an abstract, low-detailed picture of the system and use its properties as our reference. Conversely, when pursuing a bottom-up approach we should start from a high-detailed reference representation which is usually related with accurate and realistic properties. Thus, bottom-up approaches should be preferred since they are based on more reliable reference data as compared with the ones we would use for the top-down methods.

In the literature, bottom-up representation shifts are usually called *coarse-graining* methods. From now on, I will use the abbreviation CG to indicate both "coarse-graining" and "coarse-grained".

In this thesis, I want to provide a set of methods to perform the CG modelling of two phenomena typical of host-guest systems: adsorption and diffusion of the guest species. The main idea is to ideally fill the gap between microscopic and macroscopic descriptions of such phenomena, with a set of possible mesoscopic representations. This is because the methodologies to represent such phenomena in the *microscale* (lengths $\lesssim 1$ nm) and *macroscale* (lengths $\gtrsim 1$ μm) are already well-established and widely used. The most accurate methods to characterize the dynamics of host-guest systems in the microscale include ab-initio molecular dynamics (AIMD) and path integral molecular dynamics (PIMD), which allow introducing quantum effects in atomistic simulations [2, 3]. If quantum effects can be ignored, we can usually employ classical molecular dynamics (MD) methods [4]. The adsorption static properties are usually studied using grand-canonical Monte Carlo methods (GCMC) [5]. Both MD and MC methods will be briefly introduced in the following chapter. Despite the great accuracy related to such methods, they are usually limited to the microscale, mainly because of the computational effort required or because of numerical stability.

On the macroscale, the methods used to investigate such phenomena are usually variants of computational fluid dynamics (CFD) approaches [6]. CFD methods attempt to numerically and iteratively solve the Navier-Stokes equations, which describe fluid flows under the assumption that the flowing substance can be described as a continuum. Unfortunately, there is no general procedure to parameterize CFD methods to represent host-guest systems. Usually, the parameters involved are obtained by numerically fitting a certain number of macroscopic properties of the systems under study, i.e. through a top-down approach. Such approaches might be useful for practical uses, but they require to be applied very carefully as they are not based on the underlying physics of the system but rather on some property matching criteria. For this reason, macroscopic methods may reproduce macroscopic properties correctly, but there is no guarantee that the real underlying mechanisms will be physically meaningful. Recently, efforts have been made to introduce bottom-up approaches or hybrid simulations where the CFD methods are made consistent with atomistic MD simulations, but such strategies are still in the early stages [7].

This motivates the idea of researching new representations that could ideally lie between the microscopic and macroscopic methods, but with the condition that they should be parameterized upon the microscopic results. In particular, this thesis wants to demonstrate the possibility of defining *discrete* CG models of the reference systems. This means that the set of possible CG states is *countable.* As a consequence, if the CG dynamics are defined, they should be based on abrupt changes from one state to the other.

We chose to use a discrete modelling approach for different reasons. In general, discrete models require less computational effort if compared with the continuous counterparts. Also, due to their simplicity, discrete models have known analytic solutions more often. Conversely, continuous models usually require the use of differential equations, which are analytically solvable in a narrower range of cases. For this reason, many continuous models have to undergo discretization in order to be simulated.

Despite these pros, discrete models often suffer from their limited resolution, and may produce unrealistic results. For example, one of the first attempts to reproduce a fluid through a *cellular automaton* (CA), was the HPP model [8]. Unfortunately, the model was unrealistically anisotropic in its behaviour due to the insufficient degree of *rotational symmetry.* For this reason, the HPP model was not able to reproduce the Navier-Stokes equations in the macroscopic limit [9]. Despite such failures, there are many notable examples of successful discrete representations of physical systems, such as the famous Ising model, originally used to represent magnetic properties of matter [10]. Another important mention has to be made to the Markov state modelling of molecular systems approaches developed by Noe et al [11, 12]. Their methods are particularly suitable for the representation of systems of biochemical importance, such as the study of protein folding and the analysis of metastable states. Such methods allow defining discrete models of the reference systems by partitioning the configuration space in disjoint states using general statistical assumptions, and the transitions from one state to the other can be used to study or represent the molecular kinetics.

Strictly speaking, no discrete model of nature is completely realistic, since — at least according to classical theories such as general relativity— space-time is continuous. However, there are scenarios in which the discrete approach yields a realistic approximation of the chosen problem. It is the case of systems characterized by the fact that, under the time resolution we adopted to observe the dynamical evolution, the transformation from one state to the other is practically abrupt.

1.2 Coarse-graining strategies

This section aims to provide the reader with an overview of the main CG methods that are currently employed in the scientific community. Usually, and within this thesis as well, the CG of molecular systems involves the definition of a *mapping function* which transforms the original atomic coordinates into states expressed in terms of CG variables. In a certain sense, CG strategies share some common features and can be ideally resumed in the scheme depicted in Fig1.1.

However, they might present some fundamental differences when applied in practice. For this reason, CG methods can be grouped in several ways according to the respective similarities and differences. For our purposes, it is useful to introduce the distinction between *topological* and *spatial* approaches.

In the first case, atomic positions are grouped into beads and equipped with a structure of connections that ends up with CG objects representing fine-grained molecules at a coarser resolution — such objects are still "molecules", in the sense

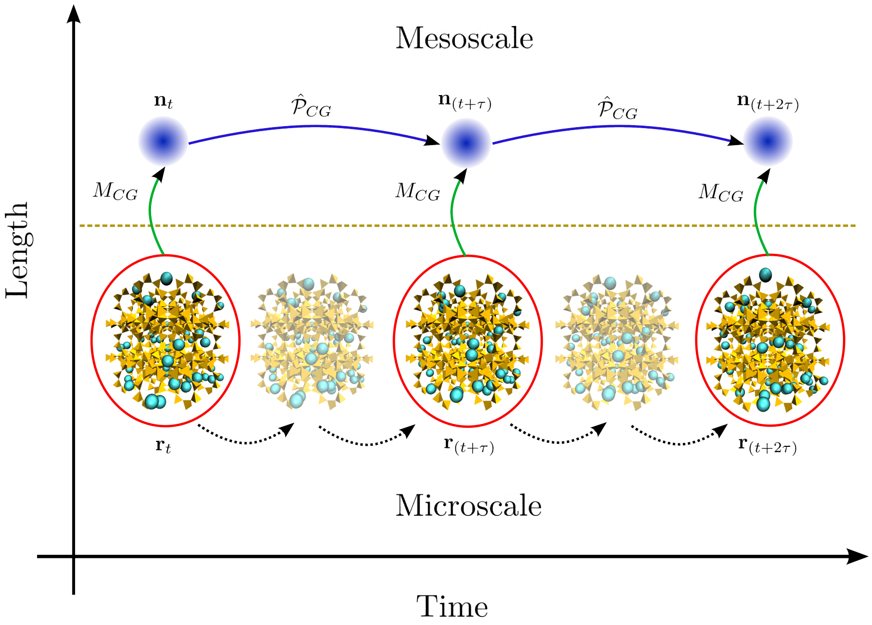

Figure 1.1: A general scheme representing the coarse-graining of a molecular system. The x-axis represents the simulated time, while the y-axis represents the length scale. In the lower part, the system is represented at atomistic resolution in the microscale, where each configuration is represented by the atomic positions $\mathbf{r}$, and the evolution is governed by the atomistic laws of motion (dashed black arrows). A subset of configurations are mapped (green arrows) to a mesoscopic coarse-grained representation $\mathbf{n}$ through the operator M_{CG}. In the mesoscale, the CG dynamics are represented by the application of a CG operator $\hat{\mathcal{P}}_{CG}$.

that, at CG level, beads' positions, orientations, and momenta are still defined, and still represent the observables that undergo dynamical evolution [13–16]. CG objects should be intended molecules in a broad sense, as they might also include artificial features such as virtual sites [17, 18]. Then, the fine-grained (FG) force-field is mapped into an effective force-field; this can be performed according to different methods such as force-matching and Boltzmann inversion techniques [13, 14, 19–23]. At CG resolution, molecular objects will evolve through the same evolution algorithm as the one used to simulate the system in the fine-grained scale, usually a molecular dynamics algorithm which, in case, might include modifications, like handling friction and random forces as in the case of dissipative-particle dynamics [24]. However, CG and FG evolution algorithms share the same nature.

In the case of spatial coarse-graining, the system is partitioned into a grid of cells, which tessellate the entire simulation space; then the FG observables of molecules (such as position and momenta), as they are extracted from some fine-grained source of data (GCMC or MD simulation trajectories), are mapped to a CG state that can represent different features associated to each cell [25–32]. A very common kind of CG state for spatial approaches is the local amount of species i.e. the *occupancy*, but also other information can be considered such as local charge, polarization, local mass distribution etc. Therefore, this strategy leads to the definition of objects that are no longer "molecules", but they rather are "cells", geometrical entities that are supposed to neither change their shape nor their position in space during the time-evolution of the coarse-grained system; in this case, at a coarse-grained level, it is the CG states that undergo dynamical evolution. Finally, the CG representation is equipped with an evolution algorithm that updates the CG states, thus mimicking as closely as possible the dynamical evolution of the fine-grained counterpart. Usually, such algorithm is stochastic and is based upon schemes that might not have anything in common with molecular dynamics-based algorithms, but might rather resort to kinetic Monte Carlo or cellular automata rules [33, 34]. Such rules can be conveniently parameterized according to the state-change rates observed from the reference fine-grained simulations.

The subdivision between spatial and topological approaches proposed in this thesis is in analogy with the definitions of field-based and particle-based approaches to mesoscopic representations, usually present in the soft matter literature.

The two approaches present profound differences, but they are conciliable. In fact, examples of hybrid simulations, where molecular-based and spatially coarse-grained representations coexist in the same simulation environment, are present in the recent literature[35–37]. For example, the definition of coarse-grained potential may involve both molecular-like contributions (such as particle positions, bonds between beads etc.), and field contributions, which depend on other coarse-grained variables such as the local density of particles.

1.3 Outline of the thesis

The remaining part of this thesis is organized as follows. In the next chapter, I will introduce the main background concepts that should provide the reader with the basic tools to understand the core parts of this thesis. The third chapter is dedicated to a preliminary work devoted to the automatic definition of CG states through a data-based method. The fourth chapter is a re-adaptation of a published work concerning the coarse-graining of host-guest systems where the interacting pair approximation (IPA) is introduced [38]. The fifth chapter reports another published article dedicated to the structural generalization of the IPA approach, focused on the modelling of more complex host environments [39]. The sixth chapter is an adaptation of another article, which was just accepted by the editor at the time of this writing [40]. Such work completes the coarse-graining of adsorption and diffusion in host-guest systems by providing a dynamical evolution algorithm which can be used to represent mass-transfer mechanisms. Finally, in the last chapter I will draw the conclusions concerning the work here presented, and propose future perspectives.

Chapter 2

Theoretical Background

2.1 Molecular representations

2.1.1 Classical atomistic representation

Molecular systems can be represented under different levels of detail. The choice of a specific representation is usually arbitrary and depends on the type of systems and phenomena investigated. We can introduce an example molecular system $\mathcal{S}$, which can be a single molecule, a full crystal or, more generally speaking, just a bunch of atoms.

The most detailed and accurate way to represent and characterize $\mathcal{S}$ is by introducing its wave function $\Psi_{\mathcal{S}}$, which ideally contains all the information related to its static and dynamical properties. Unfortunately, the closed form of Ψ is known only for very simple systems under specific boundary conditions. Several approximate methods for the estimation of $\Psi_{\mathcal{S}}$ exist, each one introducing specific assumptions and with a limited range of applications. Furthermore, the use of such methods is usually limited to small systems because of the large computational effort required.

However, this thesis focuses on the representation of phenomena occurring in spatial and time scales which are usually larger and sufficiently separated from the ones where quantum effects need to be simulated explicitly.

If such assumption holds, we can treat the systems as fully classical ones. In this way, all quantum effects, such as the consequences of Heisenberg's uncertainty principle, can be ignored thus allowing us to uniquely define the positions $\mathbf{r}^N$ and the momenta $\mathbf{p}^N$ of the N particles constituting the system. The use of the superscript in $\mathbf{r}^N$ and $\mathbf{p}^N$ is a shortcut to indicate the position and momenta vectors of all the N particles present in $\mathcal{S}$.

Furthermore, we can restrict our representation to the atomic nuclei, ideally intended as single particles.

Thus, treating our system as a mechanical one, the state of $\mathcal{S}$ can be completely represented by the pair $\left(\mathbf{r}^N, \mathbf{p}^N\right)$, that indicates a specific point $\mathbf{\Gamma}_{\mathcal{S}} = \left(\mathbf{r}^N, \mathbf{p}^N\right)$ in the system's phase space $\{\mathbf{\Gamma}_{\mathcal{S}}\}$, which is the space of all possible positions $\mathbf{r}^N$ and momenta $\mathbf{p}^N$ of the system.

This kind of representation defines the microscopic state or *microstate* of the system within the classical mechanics' approximation. It is particularly important since it constitutes the standard input for classical force-fields which are widely used in computational simulations like classical molecular dynamics (MD) and Monte Carlo (MC) methods, which will be explained more in detail in the following section.

This atomistic, classical representation of $\mathcal{S}$ is also the starting point for the development of all the derived representation that will be introduced in this thesis. A sketch representing an example molecular system, under the lens of the different representations used in this thesis, is shown in Fig. 2.1.

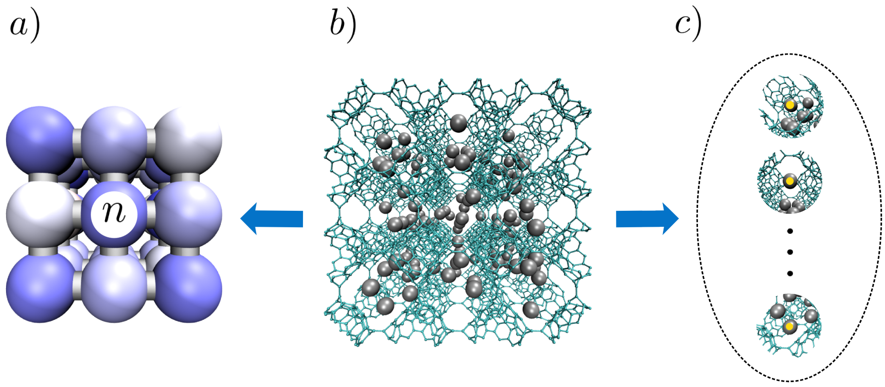

Figure 2.1: Sketch of the same molecular system depicted according to the different kinds of representations used in this thesis. Subfigure *a* represents the occupancy-based network of cells; subfigure *b* shows the atomistic representation; subfigure *c* shows a decomposition in atom-centred local environments, a typical input of machine-learning models.

2.1.2 Local occupancy-based representations

When considering host-guest systems, a possible approach for representing each configuration is to consider the local amount of guest species, which we call *occupancy*[38–41]. In order to do this, the reference system needs to be divided into a set of non-overlapping subvolumes, which we will refer to as *cells*.

The division in cells can be performed in several ways and the choice is usually made on the basis of the host's structure. For example, an ordered porous material can be conveniently divided partitioned according to a regular tiling in such a way that every cell embeds a single pore or a group of pores ($2 \times 2 \times 2$, $3 \times 3 \times 3$, etc.).

Disordered hosts generally require more complicated partitioning schemes. A possible strategy could be to divide the space with clustering techniques applied to the distribution of guest molecules in the space covered by the host. In this way, it is possible to find potential centres for the cells, located where the guest molecules tend to accumulate. Then, on the basis of the clusters' centres, one can construct a Voronoi tessellation to fill the entire simulation space with cells.

In both cases, the reference molecular configuration is mapped to a set of N_c distinct cells, which can be connected on the basis of the interaction network and may represent the directions of particle motion.

After partitioning, the state of the system can be represented by a configuration of occupancies which can be embedded in a vector $\mathbf{n} = (n_1, n_2, ..., n_{N_c})$. The components of such a vector represent the occupancies, which are equal to the total number of guest species located inside of each cell. The occupancy value for the i-th cell can be obtained using the following formula

$$n_i = \sum_{j=1}^{N_p} \theta_i(\mathbf{r}_j), \tag{2.1}$$

where $\mathbf{r}_j$ is the position vector of the j-th particle and $\theta_i(\mathbf{r}_j)$ is a membership function, which is equal to 1 if $\mathbf{r}_j$ is included in the volume covered by the cell i, and 0 otherwise.

This level of description is particularly convenient for the coarse-grained representation of phenomena which depend on the distribution of guest molecules, within the host environment. In fact, this is the starting point for the representation of adsorption and diffusion phenomena through lattice models.

2.1.3 Representations for machine learning

Mapping from atomistic representations to derived descriptions is generally a difficult task that requires some knowledge of the reference system and the phenomena which are the object of study. Humans can be very accurate in understanding and rationalising molecular structures but, even in the best scenarios, their capabilities are limited and can be subject to biases and logical fallacies. To cope with this kind of issues, it is necessary to develop strategies based on an agnostic perspective, which would be unbiased by definition; but with the capability of reproducing the chosen systems and properties with satisfactory accuracy.

In the last years, machine learning (ML) methods have explored several ways of describing molecular structures for different purposes, such as classification, potential models development, rationalization of structure-property relationships etc [42]. Before being employed for solving specific tasks, ML methods require a *training set* of molecular configurations $\{\mathcal{S}\}$. This set is then used to tune the parameters embedded in the chosen ML model.

To be used for the training, $\{\mathcal{S}\}$ has to be converted in a set of appropriate inputs $\{\mathbf{x}\}$. Such inputs can also contain the properties $\mathbf{y}$ related to the reference structures, which have to be fed to the ML machinery if the goal is to learn the relation between an input structure and the relative properties.

The choice of the kind of molecular inputs, usually called *descriptors*, that are used to feed the ML model is particularly crucial and may dramatically affect the performance. For certain tasks, local descriptors can be used instead of providing the full reference structures. This is, for example, the case of learning a set of properties that can be safely divided into local contributions or if the goal is to define and compare local structural patterns. The decomposition of a reference molecular system in local environments is depicted in Fig. 2.1.

In principle, atomic cartesian coordinates could be used as descriptors for the ML model, but the downside is that physically indistinguishable configurations could be misinterpreted as different ones. For example, if a molecular structure is rigidly translated from the starting configuration, the atomistic coordinates would change and the final structure would be considered different from the starting one. There are three kinds of transformations with respect to which a good local descriptor should be invariant to: translation, rotation and permutation of the same-elements atoms. These transformations are depicted in Fig. 2.2 with an example molecular system composed by two CO_2 molecules.

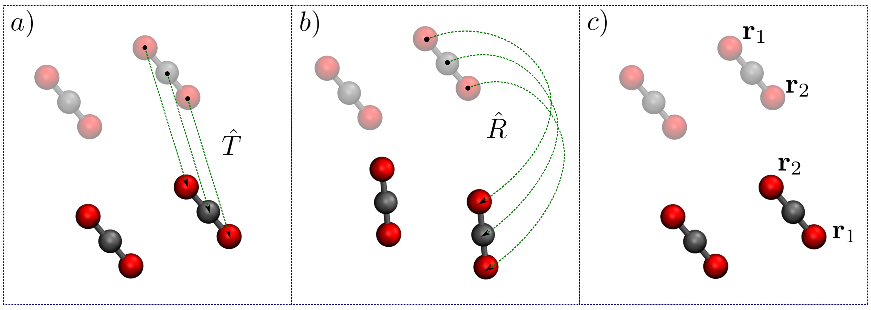

Figure 2.2: Transformations which should be encoded in an appropriate descriptor of a molecular system: translation $\hat{T}$ (subfigure *a*), rotation $\hat{R}$ (subfigure *b*), permutation of the same element-atoms (subfigure *c*).

Popular examples of local descriptors for machine learning are internal coordinates such as distances and angles between atoms, which are naturally invariant to translation and rotation. However, the choice of which internal coordinates should be considered relevant depends on some prior understanding of the molecular systems and phenomena which have to be investigated.

Further general-purpose local descriptors have been introduced with the scope to provide a more general and unbiased encoding of local atomic configurations. It is the case of Behler-Parrinello's symmetry functions and SOAP (smooth overlap of atomic positions) descriptors [43, 44]. The latter kind of descriptor is the one that was used in this thesis for the analysis of local CO_2 adsorption patterns in the ITQ-29 zeolite. Such descriptors do not require any detailed prior physico-chemical knowledge of the molecular study, and require the tuning of only a few general parameters (which elements should be analysed, the cutoff radius of each descriptor, and the set of basis functions that should be used). SOAP and Symmetry functions are successfully employed for the statistical analysis of molecular configurations through ML algorithms, but also as input set for ML potentials that can be used for molecular simulations [45–47].

2.2 Simulation methods

2.2.1 Statistical mechanics background

In this paragraph, I would like to provide the reader with a very brief and intuitive picture of a few, but very relevant concepts from statistical mechanics, which are crucial to understand the rest of this thesis. A complete and rigorous derivation of the formalism introduced in this paragraph is beyond the scope of this thesis.

As shown in the previous section, a molecular system can be characterized by different levels of description. From a macroscopic point of view, the state of a system can be defined on the basis of some properties such as the volume V, the temperature T, the amount N of substance, or of substances $\{N_i\}$, constituting the system etc. For example, if a system is isolated with respect to its environment, the most concise way to characterize its macroscopic state or *macrostate* is to define the triplet (N, V, E), where E is the total energy.

In general, there can be many microstates associated to the same macrostate. For non-trivial systems, enumeration and systematic study of all accessible microstates within the classical approximation is impossible, since the phase space of a mechanical system is constituted by an infinite number of points Γ, each one related to a specific microstate.

However, a detailed analysis of molecular systems can still be conducted by introducing the concept of *ensemble* [48]. We first postulate that the average properties of a system are equal to the properties in the thermodynamic limit — i.e. in the limit for $N \to \infty$. We also introduce the idea of *ensemble* $\mathscr{S} = \{\mathcal{S}_{(N,V,E)}\}$, which is a set of molecular systems $\mathcal{S}$ defined by the same macrostate (N, V, E). This particular kind of ensemble is called *microcanonical* ensemble.

Practically, one could ideally imagine producing several replicas of $\mathcal{S}$ with the same macrostate, but with different starting microstates. Allowing the different systems to evolve following the same classical mechanics' laws, it is possible to obtain a set of uncorrelated systems that can be imagined as independent walkers in the phase space $\{\mathbf{\Gamma}_{\mathcal{S}}\}$, each one drawing its own trajectory.

In the limit of infinite replicas, the phase space distribution of the walkers will be equal to the probability distribution of the states in the phase space. In the (N, V, E) case, the probability of finding the system in a specific state $(\mathbf{r}^N, \mathbf{p}^N)$ is uniform over all the configurations *which total energy belongs to an infinitesimally narrow energy interval centered in E*. The microcanonical probability density function is defined as follows:

$$\rho(\mathbf{r}^N, \mathbf{p}^N) = \frac{1}{N! h^{3N} \Omega_{(N,V,E)}}, \tag{2.2}$$

where $\Omega_{(N,V,E)}$ is the microcanonical partition function, also called *degeneracy*, which is equal to the number of states compatible with the triplet (N, V, E); h is Planck's constant, which represents the minimum amount of action possible; $1/N!$ corrects over-counting configurations with indistinguishable particles. The partition function $\Omega_{(N,V,E)}$ is related to the entropy S through the following equation:

$$S = k_B \ln \Omega_{(N,V,E)}, \tag{2.3}$$

where k_b is the Boltzmann's constant. Such equation is particularly important, since it is related to the *equilibrium state* of the system. The entropy S is said to be the *characteristic state function* of the microcanonical ensemble. Every ensemble is related to its characteristic state function through equations similar to the one shown in Eq. 2.3. Characteristic state functions are among the most difficult quantities that one would want to estimate through molecular simulations. This is because they are strictly connected to the knowledge of the partition functions, which are related to volumes in phase space.

Since we assume the distribution in Eq. 2.2 to be uniform, we expect that, after a sufficiently long time, the systems will visit every accessible microstate equally.

This is called the *ergodic hypothesis* and it is crucial for justifying the consistency of molecular simulations.

The main implication of ergodicity is the equivalence between ensemble (or phase space) averages and time averages. For example, considering an observable O, its ensemble average can be written as:

$$\langle O \rangle = \int O(\mathbf{r}^N, \mathbf{p}^N)\rho(\mathbf{r}^N, \mathbf{p}^N)d\mathbf{r}^N d\mathbf{p}^N, \tag{2.4}$$

which, considering a sufficiently long time window, is equal to the time average of O

$$\langle O \rangle = \lim_{n_t \to \infty} \frac{1}{n_t} \sum_{t=1}^{n_t} O(\mathbf{r}_t^N, \mathbf{p}_t^N), \tag{2.5}$$

where we intend to average over a time series composed of n_t microstates of the system. For our purposes, we will *assume* the ergodic hypothesis to be true for all the molecular systems we consider.

Ergodicity implicitly suggests two possible strategies for calculating the average properties through molecular simulations:

- to generate a sufficient number of configurations directly from $\rho(\mathbf{r}^N, \mathbf{p}^N)$, to compute the properties for each configuration and then, divide by the total number of generated microstates;

- to let the system evolve iteratively from a starting configuration according to its laws of motion, to compute the properties for each microstate visited and then, divide by the total number of iterations.

In the microcanonical ensemble, the constraint of constant total energy E defines a specific hypersurface in phase space. This means that the system can only visit the microstates which are comprised within such surface.

If the system is coupled with a thermostat at a constant temperature T, instead of having its total energy fixed, all the states of the phase space become accessible. This is due to the energy fluctuations of the system induced by the coupling with the thermostat. In this case, the system's macrostate will be defined by the triplet (N, V, T). The ensemble of systems having the same (N, V, T) triplet is called *canonical ensemble*. In this case, the probability density related to a particular microstate reads

$$\rho(\mathbf{r}^N, \mathbf{p}^N) = \frac{e^{-\beta H(\mathbf{r}^N, \mathbf{p}^N)}}{N! h^{3N} Z_{(N,V,T)}}, \tag{2.6}$$

where $H(\mathbf{r}^N, \mathbf{p}^N)$ is the energy of the system, β is $(k_B T)^{-1}$. $Z_{(N,V,T)}$ is the canonical partition function, defined as

$$Z_{(N,V,T)} = \frac{1}{N! h^{3N}} \int e^{-\beta H(\mathbf{r}^N, \mathbf{p}^N)} d\mathbf{r}^N d\mathbf{p}^N. \tag{2.7}$$

The characteristic state function of the canonical ensemble is the Helmholtz free energy A, which is related to the partition function through

$$A = -\beta^{-1} \ln Z_{(N,V,T)}. \tag{2.8}$$

The canonical ensemble is particularly convenient for the simulation of thermally-activated processes, such as the diffusion of guest species on solid surfaces. For this reason, the canonical ensemble will be employed for the simulation and analysis of pore-to-pore molecular jumps in microporous materials.

If the systems we consider are also allowed to exchange particles with a reservoir, the ensemble is called *grand-canonical*. In this case, the ensemble is defined by the triplet (μ, V, T), where μ is the chemical potential associated with the species which compose the system. The probability distribution associated with this ensemble reads

$$\rho(\mathbf{r}^N, \mathbf{p}^N) = \frac{e^{-\beta\left[H(\mathbf{r}^N, \mathbf{p}^N) - \mu N\right]}}{N! h^{3N} \Xi_{(\mu,V,T)}}, \tag{2.9}$$

where $\Xi_{(\mu,V,T)}$ is the grand-canonical partition function which can be expressed as a function of the canonical partition function $Z_{(N,V,T)}$:

$$\Xi_{(\mu,V,T)} = \sum_{N=0}^{\infty} e^{\beta\mu N} Z_{(N,V,T)}. \tag{2.10}$$

The grand-canonical partition function is related to the grand potential Φ through

$$\Phi = -\beta^{-1} \ln \Xi_{(\mu,V,T)}. \tag{2.11}$$

In this thesis, the grand-canonical ensemble will be employed for the study of static properties of host-guest systems, such as the distribution of guest molecules in the host environment.

2.2.2 Monte Carlo

The term Monte Carlo (MC) is used to denote a family of methods that makes extensive use of random numbers. The main idea behind MC methods is that it is possible to randomly sample a volume in a given d-dimensional space in order to obtain an estimate of an integral defined in such space.

MC techniques are particularly useful for the estimation of integral quantities in spaces with high dimensionality. An example molecular system composed by $N = 1000$ particles is related to a 6000-dimensional phase space — this is because we have to consider $3N$ position components + $3N$ velocity components — and it would be impossible to evaluate any integral quantity with deterministic quadrature- or grid-based techniques. Even if we just consider the *configuration space*, i.e. the set of all possible particle positions, the dimensionality would still be $3N$.

To introduce the MC techniques, let us consider for the moment a 2-dimensional problem. Fig. 2.3 shows an area V enclosed in a bigger orthogonal parallelogram

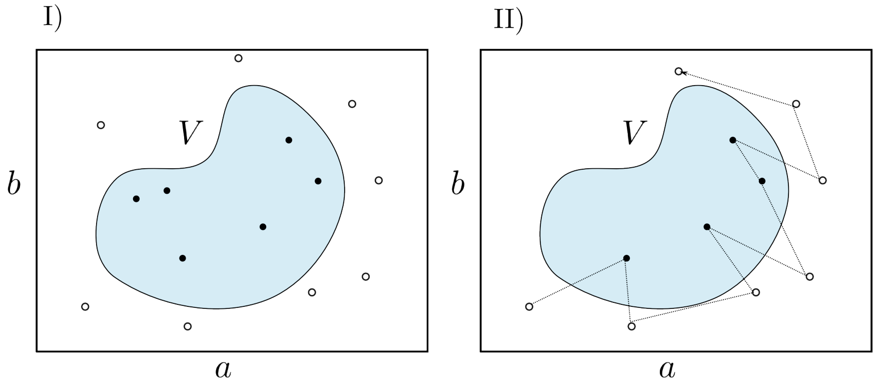

Figure 2.3: Two different Monte Carlo strategies for the estimation of the area V enclosed in an orthogonal parallelogram of area $A = ab$. Subfigure (I) represents the random uncorrelated generation of points, while subfigure (II) represents a random walk method.

of area $A = ab$. The estimation of V through MC can be performed by generating a set of random positions and checking if the point falls within V (subfigure I).

After a sufficient number of point generations we can estimate V by using the ratio between the n_V points falling in the chosen volume and the total number of points N_{trials}

$$V = \lim_{N_{\text{trials}} \to \infty} \frac{n_V ab}{N_{\text{trials}}}. \tag{2.12}$$

Considering molecular systems, this is equivalent to generating random molecular positions from scratch and then evaluating a certain set of properties for each configuration.

Despite its simplicity, this method is not normally used for molecular systems because it would easily lead to particle overlaps and it would be not very handy if we required some spatial constraints to be fulfilled. An easy solution to these practical problems is to start from a first configuration and then proceed iteratively with the exploration of available space by applying random perturbations. This concept is depicted in the subfigure II of Fig. 2.3 as a random trajectory. This procedure generates a correlated motion since every new point depends on the previous one. For this reason, this kind of method requires a sufficient number of points such that correlations are amply lost during the random walk. Finally, after a sufficient number of generations, we can employ Eq. 2.12 for the estimation of V.

MC simulations of molecular systems involve the creation of a set of molecular configurations $\{\chi\}$. Assuming that such configurations are drawn from the correct probability distribution of the chosen ensemble, the estimation of the average value of a certain property O — already rigorously defined in Eq. 2.4 — through MC simulations can be carried out through the following formula

$$\langle O \rangle = \lim_{N_{\text{trials}} \to \infty} \frac{1}{N_{\text{trials}}} \sum_{\{\chi\}} O_{\chi}, \tag{2.13}$$

which is conceptually analogous to Eq. 2.12.

In general, not all the configurations are equally important for the estimation of the selected properties. The most relevant regions are the ones for which the selected properties are close to the relative average values and for which the probability density values are relatively high [5]. This fact is naturally evident in Eq. 2.4. A smart algorithm would mainly focus on such regions instead of exploring uniformly the whole configuration space. This is the core idea behind the so-called *importance sampling.*

The most popular MC method for simulating molecular systems is the Metropolis-Hastings algorithm. The idea is to perform a Markov chain walk in configuration space through iterative stochastic perturbations applied to the starting configuration. Possible perturbations are be atomic displacements, rigid rotations, and — for grand-canonical ensemble simulations — insertion/deletion of molecules.

The generation of the configuration χ' depends only on the previous configuration χ, and is drawn from the conditional distribution that we indicate as $g(\chi' \mid \chi)$, which represents the probability of *proposing* χ' as the next state. The molecular configuration we generated from $g(\chi' \mid \chi)$ will then be accepted (meaning that χ' becomes the new current configuration) or rejected (meaning that χ remains the current configuration, and χ' is discarded) on the basis of the *acceptance* probability, which depends on the difference in potential energy between χ' and χ, if $g(\chi' \mid \chi)$ is uniform.

According to the Metropolis-Hastings algorithm, χ' will be accepted with the following probability:

$$\alpha(\chi' \mid \chi) = \begin{cases} 1, & \text{if } \Delta V_{\chi',\chi} \leq 0 \\ e^{-\beta \Delta V_{\chi',\chi}}, & \text{otherwise. ,} \end{cases} \tag{2.14}$$

where $\Delta V_{\chi',\chi}$ is the difference in potential energy between the state χ' and χ. Resuming, the probability that this transition will effectively occur is $P(\chi' \mid \chi) = g(\chi' \mid \chi) \cdot \alpha(\chi' \mid \chi)$. This evolution scheme leads the generated stochastic trajectory to produce, after a sufficient number of iterations, the desired equilibrium distribution since it satisfies the *detailed balance* condition. Such condition ensures that the the Markov chain is *stationary*, with the correct equilibrium distribution as its marginal probability distribution. Considering the example states introduced before, such condition reads

$$\rho_\chi P(\chi' \mid \chi) = \rho_{\chi'} P(\chi \mid \chi'), \tag{2.15}$$

with ρ_χ and $\rho_{\chi'}$ being the equilibrium probabilities related to the two molecular states. Using Eq. 2.15 and imposing the simmetry $g(\chi' \mid \chi) = g(\chi \mid \chi')$, it is easy to show that the Metropolis-Hastings rule satisfies the detailed balance condition.

2.2.3 Molecular dynamics

Unlike MC methods, molecular dynamics (MD) methods focus on the realistic simulation of molecular systems' time evolution. For this reason, MD methods are widely used to study dynamical properties such as diffusion, time correlations etc. while sampling the correct static properties if the ergodic hypothesis holds. MD simulations are performed by numerically integrating the laws of motion of the chosen system.

In principle, the laws of motion of a mechanical system are described by Newton's second law, which for a generic i-th particle reads

$$m_i \ddot{\mathbf{r}}_i = \mathbf{f}_i, \tag{2.16}$$

where m_i is the mass, the vector $\ddot{\mathbf{r}}_i$ is the second-order time derivative of the position i.e. the acceleration, $\mathbf{f}_i = -\nabla_i V$ is the force, and V is the potential energy.

A more abstract and general formalism for the laws of motion is provided by Hamilton's equations. We introduce the *Hamiltonian* $\mathcal{H}$, which corresponds to the total energy of the system. For a closed system, $\mathcal{H} = \mathcal{K} + \mathcal{V}$, with $\mathcal{K}$ being the total kinetic energy and $\mathcal{V}$ the total potential energy. Within the classical mechanics formalism, the kinetic energy of each particle is $K_i = \mathbf{p}_i^2/2m_i$ and the potential energy $V(\mathbf{r}^N)$ depends on the interparticle interactions. The relation between the Hamiltonian and the position and momenta vectors is the following

$$\begin{cases} \frac{d\mathbf{p}_i}{dt} = -\frac{\partial \mathcal{H}}{\partial \mathbf{r}_i} \\ \frac{d\mathbf{r}_i}{dt} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}_i}. \end{cases} \tag{2.17}$$

Unfortunately, there is no general analytic solution to the laws of motion of a many-body system. For this reason, Hamilton's or Newton's equations should be solved numerically by employing a finite time interval of Δt. However, the numerical integrator should fulfil some conditions in order to provide realistic molecular trajectories:

- the integrator must be convergent, which means that in the limit of $\Delta t \to 0$, the algorithm should reproduce perfectly the Hamiltonian dynamics;

- time reversibility or time-symmetry should be satisfied, which means that by reversing all the velocity vectors we expect to obtain an identical but reversed trajectory;

- the algorithm should be symplectic, which implies that phase space volume is preserved — in agreement with Liouville's theorem for classical mechanics [48].

Verlet's algorithm satisfies all such conditions and is widely used for its simplicity and stability [49]. Basically, it makes use of the positions at time $t - \Delta t$ and t, and of the acceleration at time t. The positions are updated through the following equation

$$\mathbf{r}_i(t + \Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t - \Delta t) + \Delta t^2 \mathbf{a}_i(t). \tag{2.18}$$

Velocities are not explicitly used during the integration, but can be evaluated with the finite different method, in its *central* version

$$\mathbf{v}_i(t) = \frac{\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t - \Delta t)}{2\Delta t}. \tag{2.19}$$

A more commonly used integrator, that can be thought as a variant of Verlet's algorithm is the so-called velocity Verlet. The main difference is that this algorithm does not use any information coming from the $t - \Delta t$ step, but it rather involves the evaluation of the velocity at the half step $t + \frac{1}{2}\Delta t$. In this case, each integration step prescribes the following scheme:

1. half-step velocities are estimated using $\mathbf{v}_i(t + \frac{1}{2}\Delta t) = \mathbf{v}_i(t) + \frac{1}{2}\mathbf{a}_i(t)\Delta t$;

2. atomic positions are updated with $\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \frac{1}{2}\Delta t)\Delta t$;

3. new forces and acceleration vectors are evaluated considering the new atomic positions;

4. new velocities are computed as $\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i(t) + \frac{\Delta t}{2}(\mathbf{a}_i(t) + \mathbf{a}_i(t+\Delta t))$.

The aforementioned numerical integrators can be used for the simulation of molecular systems within the microcanonical ensemble. Modifications of such schemes can be performed to embed a *thermostat* in order to simulate the chosen systems in the canonical ensemble. A variety of methods exist to control the temperature, such as velocity-rescaling, Andersen thermostat, Nosé-Hoover thermostat etc. Each method uses different approaches to control the temperature by tuning the velocities of the particles, because the two quantities are related to each other by the following equation

$$T(t) = \sum_{i=1}^{N_p} \frac{m_i v_i^2(t)}{3k_B N_p}, \tag{2.20}$$

where $T(t)$ is the instantaneous temperature.

In this thesis, the Nosé-Hoover thermostat was chosen to control the temperature in the canonical MD simulations [50, 51]. In particular, this method is completely deterministic and ensures that we sample the correct canonical fluctuations once the system has reached the desired temperature. Basically, this approach employs a modified version of Newton's equation of motion (shown in Eq. (2.16)) by subtracting a friction force which parameterized through the term ζ. The modified Newton's equation reads

$$m_i \ddot{\mathbf{r}}_i = \mathbf{f}_i - \zeta m_i \mathbf{v}_i. \tag{2.21}$$

The dynamics of the friction term are defined by

$$\dot{\zeta} = \frac{1}{Q}\left(\sum_{i=1}^{N_p} \frac{\mathbf{p}_i^2}{m_i} - 3N_p k_B T\right), \tag{2.22}$$

where T is the desired temperature and Q is a factor introduced to speed-up or slow-down the dynamics of the friction term. In other terms, by tuning Q one can control how fast the system should converge, or equilibrate, to the desired temperature.

2.3 Materials studied in this thesis

This section aims to provide the reader with general information concerning the host materials investigated in this thesis.

2.3.1 ITQ-29

The ITQ-29 framework is a famous, relatively simple zeolite [52]. Zeolites are porous, silica-based crystalline materials. Generally speaking, they're constituted by corner-sharing silica tetrahedra, which can be isomorphically substituted by heteroatoms such as Ge, P and Al [53]. Zeolites can also accommodate ions, such as Ca and Na, which tend to localize in particular crystallographic positions of the host framework. This class of materials includes a wide variety of frameworks, which can be used for different scopes, such as gas separation, heterogeneous catalysis, ion exchange etc [53]. Specifically, the ITQ-29 zeolite is a pure-silica zeolite belonging to the Linde-type A (LTA) topology class. The ITQ-29 is constituted by a lattice of cubic unit cells with lattice parameter $a = 11.91$ Å . The unit cell and a $3 \times 3 \times 3$ supercell arrangement of this zeolite are represented in Fig.2.4.

Each unit cell includes a cavity (also called pore, or cage), centered in the middle of the cell, that can host certain guest chemical species, such as small gas molecules. Each pore is connected to 6 neighbouring pores through narrow openings, or windows, sitting on the faces of the unit cell cube.

Due to its relative simplicity, this material facilitates the modelling and the computer simulations. At the same time, it represents an important example of typical materials, which are used nowadays for gas-separation membranes. For those reasons, this material was chosen as the hosting framework for the validation of various coarse-graining techniques introduced in this thesis.

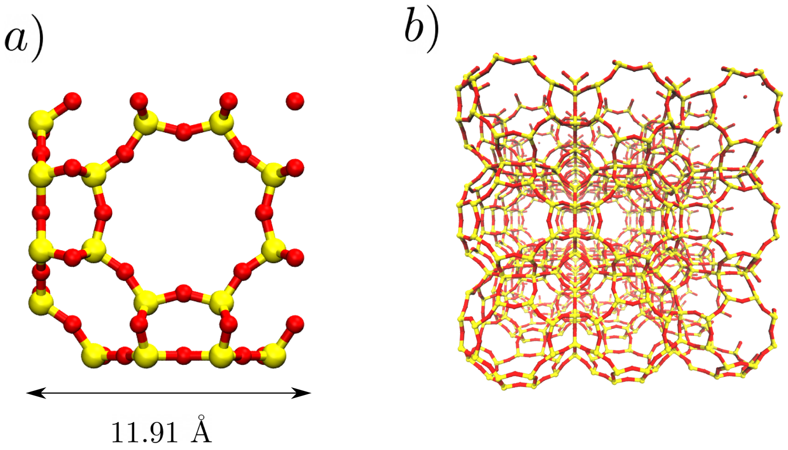

Figure 2.4: Atomistic representations of the ITQ-29 zeolite. Subfigure *a* illustrates the unit cell in a ortographic projection, while the subfigure *b* depicts a $3 \times 3 \times 3$ supercell in a perspective projection. In both representations the Si atoms are represented by yellow spheres, while O atoms are depicted as red spheres.

2.3.2 LTA-ZTC

Zeolite-templated-carbons (ZTCs) are a relatively new class of microporous materials, entirely made of carbon, synthesized using zeolites as sacrificial templates [54]. The synthesis of such materials can be summarized in two steps: i) the inclusion and carbonization of a carbon precursor in the zeolitic framework; ii) the removal of the zeolitic framework with a specific chemical treatment. Due to their peculiar properties, ZTCs have been used in many applications such as hydrogen storage, methane storage, CO_2 capture, liquid-phase adsorption, catalysis and fuel cells [54–56].

In the works presented in this thesis, we considered a carbon material that is derived from the templating of the ITQ-29 (LTA) zeolite, which was introduced in the previous paragraph. To our best knowledge, such a material has never been synthesized. However, its structure was theorized by Braun et al. in a recent work [57] and is currently available on `materialscloud.org`[58]. Atomistic

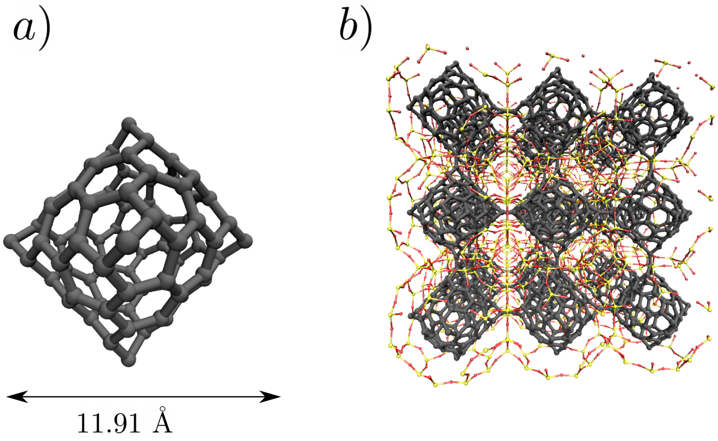

Figure 2.5: Atomistic representations of the LTA-ZTC. Subfigure *a* illustrates the unit cell in a ortographic projection, while the subfigure *b* depicts an overlap between $3 \times 3 \times 3$ supercells of LTA-ZTC and the precursor zeolite, in a perspective projection. In both representations, C atoms are represented as grey spheres, Si atoms are represented by yellow spheres, and O atoms are depicted as red spheres.

representations of the ZTC unit cell and a supercell arrangement (in comparison with the zeolitic precursor), are illustrated in Fig.2.5.

This material presents an ordered porous structure, with the same connectivity as the LTA zeolite, and with pore centres shifted by 5.95 Å along the lattice vectors, with respect to the precursor centres. Despite the similarities in terms of pores arrangement and connectivity, this material presents profound differences compared to the LTA. Except for the trivial differences in terms of chemical composition, this material presents larger pore volumes and larger openings between neighbouring cages. Such structural features will play a critical role in the adsorption and diffusion of guest particles in this material, as will be shown in Chapter 6.

2.3.3 Graphene and related materials

Graphene is a single-atom-thick layer of sp^2 carbon, that can be obtained from the exfoliation of graphite, a mineral that is entirely composed of two-dimensional carbon layers. Because of its peculiar features, graphene has been proposed as a competitive candidate for many applications, ranging from electronics to molecular sensing [59]. Such properties include high in-plane electron mobility, high thermal conductivity, and remarkable mechanical properties. Graphene can also be modified with specific functional groups, or used as a building block for derived materials. This material has also been applied to the adsorption of various chemical species, such as organic and inorganic pollutants, resulting in competitive performances if compared with other commercially available materials [60].

In this thesis, graphene is presented as a potential adsorbent for methane molecules. The interaction of methane molecules with a single graphene sheet is rather weak, resulting in poor sorptive properties. However, we also tested a derived material, constituted by two parallel graphene layers, spaced with a 12 Å interlayer distance. Both the graphene unit cell and the derived material are illustrated in Fig.2.6. The choice of this spacing is optimal for both the methane-graphene and methane-methane interactions, resulting in better adsorption performance if compared to a single graphene layer. This material was inspired by layered materials with tunable interlayer distance, such as pillared graphene, a material composed by parallel graphene layers held together by carbon nanotubes [61].

For the purpose of this thesis, the study of methane adsorption in such materials offers valid benchmarks for the occupancy-based coarse-graining method. This is because such systems represent an example of adsorption on two-dimensional surfaces. As will be shown in Chapter 5, this study constitutes a step towards the generalization of our mapping scheme for the embedding of different kinds of

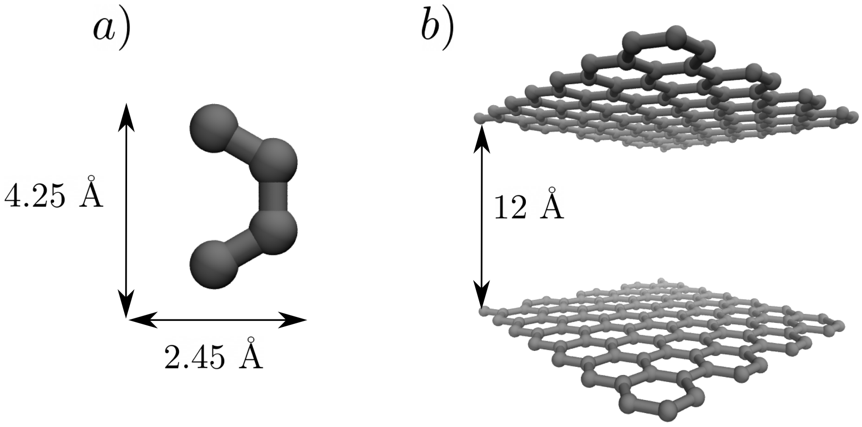

Figure 2.6: Atomistic representations of graphene and a graphene-based material. Subfigure *a* illustrates the unit cell in a ortographic projection, while the subfigure *b* depicts an hypothetical material made of graphene sheets, stacked with a certain interlayer distance. Carbon atoms are represented as grey spheres.

interactions and correlations in our lattice models.

Chapter 3

Machine learning molecular states: adsorption patterns of CO_2 in zeolites

3.1 Introduction

In this chapter, I report a method that can be used to define a set of discrete states to represent a molecular system composed by a host material and guest molecules. In this case, such roles are played by the ITQ-29 zeolite and CO_2 molecules, respectively. The method relies on the statistics drawn from atomistic MD simulations of the reference systems. The analysis of MD results is conducted with an automatic, data-based method.

The mapping of a molecular system to a discrete representation usually involves some decision-making regarding the definition of the coarse-grained states, or of the variables that should be considered relevant to the scope. Such decision is usually driven by the physicochemical understanding of the system and the phenomena involved in it. In general, the definition of coarse-grained variables depends on which kind of phenomena is under study.

Recently, techniques have been introduced to perform such mapping under a machine-learning approach. The idea is to encode the information contained in the atomistic configurations in the most general and unbiased way possible [42]. Starting from some input, a machine-learning algorithm can learn how to divide or *classify* the data in separate clusters. In this case, a discrete description of molecular states is drawn by the statistical analysis of MD trajectories. The basic assumption is that the most important configurations are the ones related to the main basins in the free-energy landscape. Roughly speaking, the discrete states will be defined on the basis of a few landmarks placed on the modes of the phase-space probability distribution. The shape of such distribution strictly depends on the choice of the kind of transformations applied to the reference configurations. For this reason, it is useful to choose a sufficiently general and *complete* descriptor to encode the information comprised in each molecular configuration.

In the following sections, I will introduce the main techniques employed in this method; then, I will show the results drawn by this analysis in terms of molecular patterns, static properties and network model construction; finally I will draw the conclusions and the future perspectives of this method.

3.2 Smooth overlap of atomic positions

The so-called Smooth overlap of atomic position (SOAP) is a kind of descriptor, which is based on an artificially-defined local atomic density [44, 46]. A detailed explanation of SOAP technicalities is beyond the scope of this thesis. However, in this section I will propose a concise description in order to provide the reader with an intuitive picture.

Every molecular configuration is split $\mathcal{S}$ in local environments $\mathbf{x}$. Each local environment is centered on a specific atom and covers the space within a certain cutoff radius r_c. For example, the environment $\mathbf{x}_i$ is the one centered on the atom i, which embeds a set of other atoms $j \in \mathbf{x}_i$ that are comprised in a sphere of radius r_c. For each environment, the atomic density is defined $\rho(\mathbf{r})$ by employing a sum of Gaussian functions g centered on the interatomic distances $\mathbf{r}_{ij}$:

$$\rho_{\mathbf{x}_i}(\mathbf{r}) = \sum_{j \in \mathbf{x}_i} g(\mathbf{r} - \mathbf{r}_{ij}) f_c(r_{ij}), \tag{3.1}$$

where $f_c(r_{ij})$ is a cutoff function which is equal to 1 at $r_{ij} = 0$ and decreases to 0 at r_c. Now, the SOAP representation can be imagined as a three-body correlation function defined on the atomic Gaussian density, instead of being defined on the original atomic positions. Taking any two points in the space covered by the environment $\mathbf{x}_i$, we can define the three-body correlation function by making use of the distance magnitudes r, r' with respect to the center, and the angle ω between the two points. The correlation function is also integrated with respect to all the possible rotations $\hat{R}$ to ensure the descriptor to be rotationally invariant.

Such correlation function is then encoded in terms of an expansion defined on an orthogonal basis of n radial functions $R_1(r), R_1(r), \ldots, R_n(r)$ and l Legendre polynomials $P_1(\omega), P_1(\omega), \ldots, P_l(\omega)$. Finally, the SOAP descriptors are represented by the power spectrum of such three-body correlation function projected on the radial and Legendre polynomial expansion.

Practically, each SOAP descriptor is obtained as a vector with an arbitrarily large number of components, which depends on the number of radial and angular functions used for the expansion. This easily yields vectors with thousands of components which can be memory-consuming and computationally heavy for training ML algorithms, if taken with the full components. For this reason, we used a technique (which is explained in the following paragraph) to project the high-dimensional data set onto a 2-dimensional space in order to run a classification algorithm.

3.3 Dimensionality reduction

There are several methods to map a set of points from a high-dimensional space with a number of dimensions D to a low-dimensional one. In general, this kind of mapping is lossy — it involves a significant loss of information — unless the removed dimensions are really redundant. Again, the choice of the coordinate system and positions in the low-dimensional space should be taken carefully.

At the moment, many of such methods focus on projecting the points in a low-dimensional space, while optimally conserving the distance or similarity between the points in the low-dimensional space [62]. In this case, we chose to use a different approach focusing more on finding the principal directions that maximize the variance in the original data set.

The principal component analysis (PCA) allows projecting an input data set onto a new coordinate system the basis of which is built upon the directions that maximise the variance on the input data (see Fig. 3.1 for a 2-d example).

The PCA transformation is linear i.e. it can be defined as a matrix T, which is constructed with the eigenvectors of the data covariance matrix $\mathbf{\Sigma}_X$. In principle, the full PCA transformation would yield a projection onto a space with the same dimensionality as the input one, but one can choose to select only the first $p < D$ eigenvectors in the definition of T. Thus, if we multiply the original vectors by T we obtain a dimensionality reduction. Not only is the PCA practically useful for lowering the dimensionality, but also provides insights regarding which features are the most distinctive among the data set.

In principle, if in the original data set two clusters can be separated by a

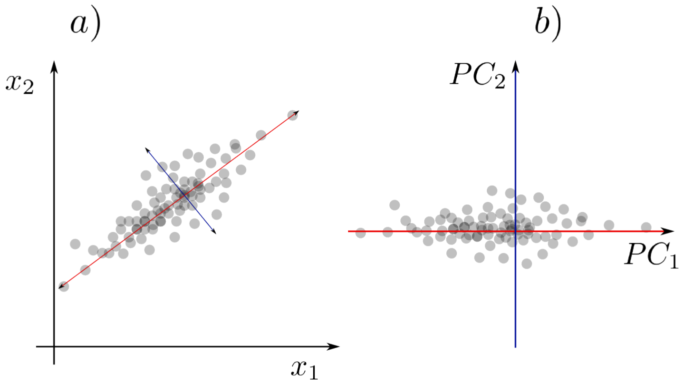

Figure 3.1: PCA on an example data set. A reference cloud of points (subfigure a) is projected onto a new coordinate system (subfigure b) defined on the basis of the two principal components, PC_1 (red line) and PC_2 (blue line).

plane, the PCA transformation should keep such separation during the projection. However, if the two clusters are separable only by employing a non-linear function, the PCA transformation would fail to keep such separation in the lower dimensional projection. To cope with this issue, more sophisticated techniques should be used to introduce non-linearities, such as the kernel-PCA method [63].

3.4 Probabilistic analysis of molecular motifs

Once the set of points, each one representing a specific environment, is projected onto a low-dimensional space it is possible to proceed with the automatic analysis of local motifs or patterns. Our choice is to use the so-called probabilistic analysis of molecular motifs (PAMM) developed by Ceriotti et al. to automatically discover recurrent local motifs in molecular configurations [64–66]. PAMM is an *unsupervised learning* algorithm, which serves to separate the input configurations in a bunch of clusters, each one representing a specific pattern.

The algorithm consists of the following main steps:

1. a subset of the reference data is selected with the *farthest-point sampling*

algorithm in order to obtain a sparse grid;

2. the density on each grid point is estimated with a *kernel density estimation* (KDE) using a gaussian kernel;

3. starting from each point of the grid, the modes of the local environment probability distribution are found as local maxima using the *quick-shift* algorithm;

4. a probabilistic model which assigns a point to a specific cluster is defined. Such model can be used to analyse new molecular trajectories of analogous systems.

The first step begins with the selection of a random point $\mathbf{y}_1 \in \{\mathbf{x}\}$, then from such point we select the next ones iteratively choosing the sample with the maximal minimum distance to the points that had already been selected. Once a grid of $M \leq N$ points have been constructed, a Voronoi tiling is constructed such that each point of the grid is the center of a tile. Then, a local probability density estimation is performed on the basis of the number of points of the data set falling in each tile. Such estimation is performed with the KDE method with a gaussian kernel

$$P(\mathbf{y}_i) \propto \sum_{j}^{N} K(\mid \mathbf{y}_i - \mathbf{x}_j \mid, \sigma_j), \tag{3.2}$$

where $K(\mid \mathbf{y}_i - \mathbf{x}_j \mid, \sigma_j)$ is a Gaussian function with standard deviation σ_j.

Once every point of the grid is assigned with its relative probability density $P(\mathbf{y}_i)$, the algorithm searches the modes of the probability distribution by looking for the local maxima by employing the quick-shift algorithm: basically, starting from each point we shift to the next point with the maximal probability within a certain range and iterate until it's no more possible to find a point with a higher probability value. With this criterion each point is assigned to the nearest local maximum, and it is assumed that each mode of the distribution represents a specific molecular pattern.

3.5 Results

3.5.1 Molecular patterns

We simulated the system composed by CO_2 and ITQ-29 with canonical MD, at different temperatures (100, 200, 300, 400 and 500 K) and CO_2 concentration regimes (1, 2, 3, 4 and 5 molecules per cavity), saving the molecular configurations every 0.5 ps, until we obtained about 50000 snapshots. In such simulations, the host material was represented by a $3 \times 3 \times 3$ supercell. The simulations were performed considering periodic boundary conditions (PBC) and the temperature was controlled through a Nosé-Hoover thermostat. The whole system was modelled with the force field parameterized for CO_2/silicates systems developed by Cygan et al.[67].

Once all the configurations were collected, each molecular snapshot was transformed into a set of SOAP vectors, according to the following criteria:

- the O atoms belonging to each CO_2 molecule are ignored, since they would bring an almost-constant signal to the local density, which may obscure other important correlations;

- each local environment we considered was centered on the C atoms, the local density was built on the basis of the central carbon atom and the O atoms belonging to the zeolite framework;

- each SOAP had a 4.0 Å cutoff in order to embed the first O coordination shell around each C atom (see Fig.3.2).

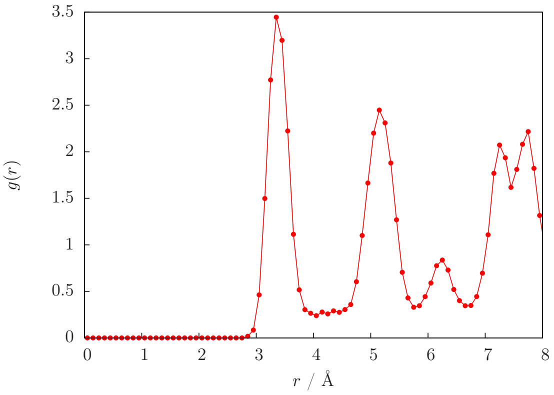

Figure 3.2: Pair correlation function centered in C atoms and correlating with zeolite O atoms at 100 K. The first peak corresponds to the coordination shell analysed in this work.

All the SOAP vectors are then used to define a common PCA transformation. Thus, we defined a projection gathering all the diversity captured during all the molecular trajectories, in order project the single MD runs in the same map with the same transformation.

The result of the PCA projection on the first two components is reported in Fig.3.3. Among thousands of molecular environments considered, PAMM grouped the data in two clusters, separated approximately along the first principal component (PC1) direction. By analysing the molecular *landmarks* — the grid points associated with the largest probability — we found that the first cluster is associated with CO_2 molecules residing in the middle of the neighboring zeolitic cages. Conversely, the second cluster is associated with the CO_2 molecules coordinated with 8 zeolitic O atoms, between two neighboring cages.

This suggests that PC1 is positively correlated with the density of O atoms in each environment. This hypothesis is also supported by the presence of three vertical branches of the distribution in the left part, which are related, from left to right, with the coordination with one, two, and three O atoms. It is still not clear what is the physical interpretation of the second principal component (PC2),

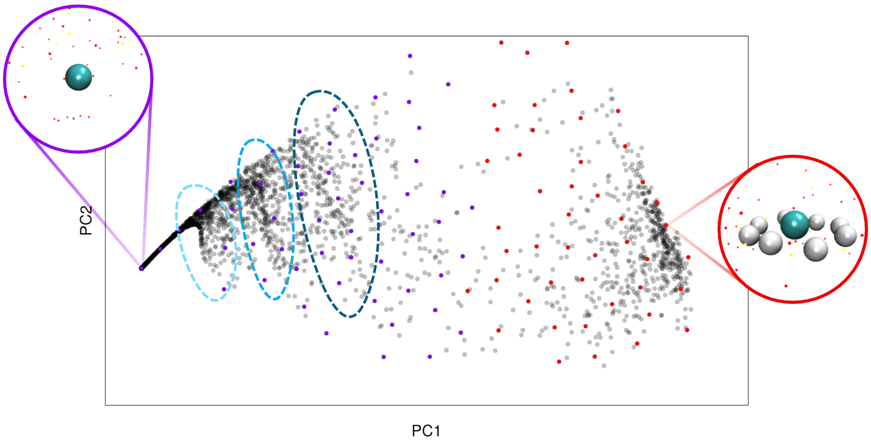

Figure 3.3: Projection of the local environments (represented as grey transparent dots) onto the first two principal components. The grid points defined by PAMM are colored according to the cluster they belong to. The two landmark molecular environments are represented in the circles with the C atoms in green and the zeolitic O atoms in white. The atoms ignored in each environments are represented by dots. On the left part, the branches of the distribution corresponding to different coordination states are highlighted by dashed lines.

since it could be correlated either with the distances from the central atom, or the angular distribution of the O atoms, or both.

The second pattern is particularly important, since it highlights the tendency of CO_2 to sit between two neighboring cages, which hinders the other molecules from jumping from one cage to the next. This is widely known as segregation effect and was observed in several microporous materials [68, 69].

3.5.2 Static properties

Once a common transformation is defined and the molecular identity of the two patterns found by PAMM was interpreted, we projected the data to discover the differences in the distribution of molecular environments under different conditions. Figure 3.4 shows a comparison of the environment probability distributions at different temperatures. It is clear that, for low temperatures the second cluster is

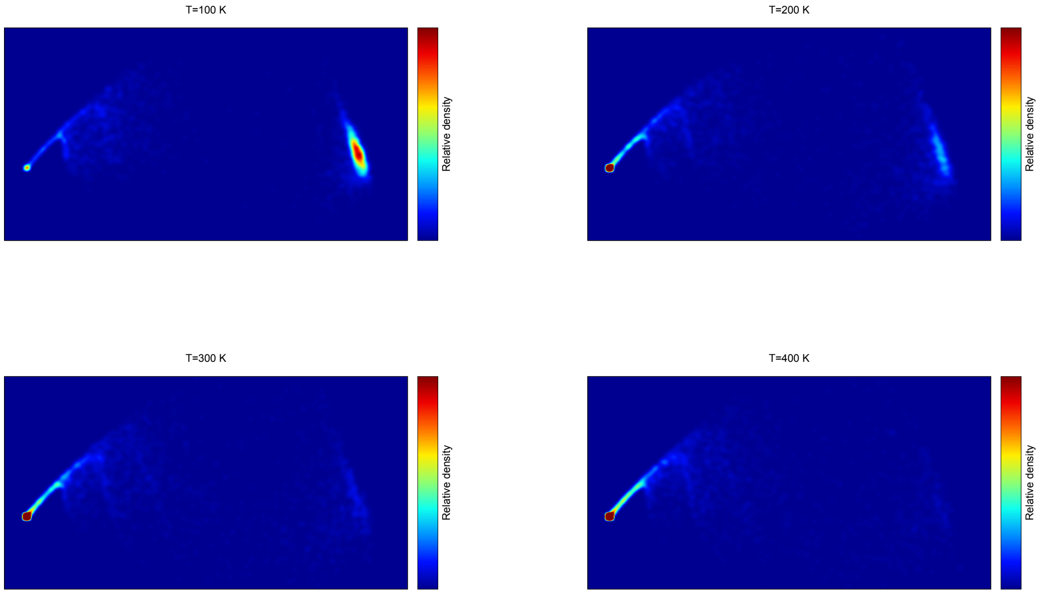

Figure 3.4: Local environment distributions as projected in the first two principal components, at different temperatures. The probability density is represented as a color gradient on each point. Red color corresponds to maximal density, while blue corresponds to the complete absence of points.

more densely populated than the first one. Basically, CO_2 molecules tend to sit in the inter-cage site rather than to occupy positions which are more any farther from the surface of the host material. As the temperature increases, the guest molecules tend to sit less often in the inter-cage site because of the more flat free-energy surface. This causes a progressive loss of density in the mode located on the right side of the projections.

It is interesting to notice that by increasing the temperature we obtained a narrower distribution, in contrast with the usual broadening effect on probability distributions related to internal or cartesian coordinates. This is a consequence of the particular mapping we chose for the analysis of the molecular environments.

We also compared the distributions coming from simulations conducted at different CO_2 loading. However, no significant change was observed within the range of 1 to 5 CO_2 molecules per cage. Further investigations are needed to explore a broader range of concentrations in order to see if stronger guest-guest correlations can modify the patterns of the distributions.

3.5.3 Backmapping and network model construction

The patterns we found can be interpreted as metastable states, based on the possible molecular environments around each CO_2 molecule in this system. Such states can be used for different purposes, e.g. to classify host-guest systems on the basis of local adsorption patterns. They can also be used for the construction of coarse-grained stochastic discrete models of the reference systems. This can be done in two main ways: by defining a random walk on pattern space — the one defined by the SOAP-PCA system, or by backmapping the CO_2 landmark positions into real space and then connecting such positions to construct a network model of the system.

In this work we started to explore the second approach. Starting from the landmarks found by PAMM, we projected the CO_2 positions into a single unit cell of the host material. Since the SOAP representation is invariant to translations, rotations and permutations, one expects to obtain an asymmetric unit of landmark positions. This is because the same crystallographic positions would be mapped into the same region of the SOAP-PCA space and would be associated to the same cluster. In order to find all the possible landmark positions within the unit cell, we applied the same symmetry operations of the host material space group.

If the landmarks sit perfectly on the Wyckoff positions, the network positions are already defined without ambiguities. In our tests, the landmarks were always displaced with respect to the Wyckoff positions and this yielded a set of ambigu-

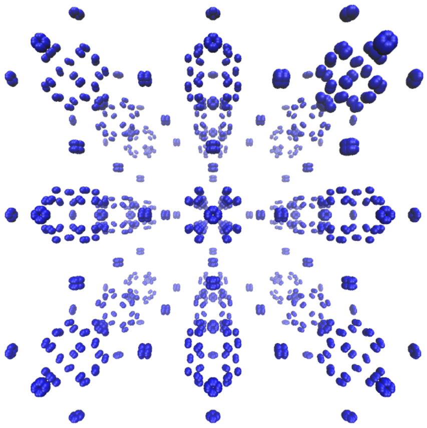

Figure 3.5: $3 \times 3 \times 3$ Supercell of the landmark sites before applying the hierarchical clustering procedure.

ous overlapping positions (see Fig.3.5). This is a consequence of the fluctuations produced during the MD simulations. In order to obtain a set of non-overlapping sites, we applied a hierarchical clustering algorithm to merge the site positions up to a certain threshold. The closest sites are iteratively merged into new sites on the basis of the mutual euclidean distance. The merging procedure is stopped when the distance between different sites exceeds 3.3 Å i.e CO_2 kinetic radius. This is done in order to ensure that CO_2 molecules are allowed to sit in neighboring sites without overlapping.

Once the sites' unit cell is refined, all we have to do is connect the different sites in order to define the links in the network model. The criterion used to determine the connections is very simple. The algorithm starts proposing a small trial distance d_t (say ~ 2 Å), it connects the sites if the mutual distance is lower than d_t and it checks if the network is *connected* — i.e. if it is possible to reach any site from every other site with a finite number of steps. If the network is not connected, then the algorithm proposes a larger distance. At each stage, the connectivity is automatically checked by applying the *depth-first search* (DFS) algorithm to the

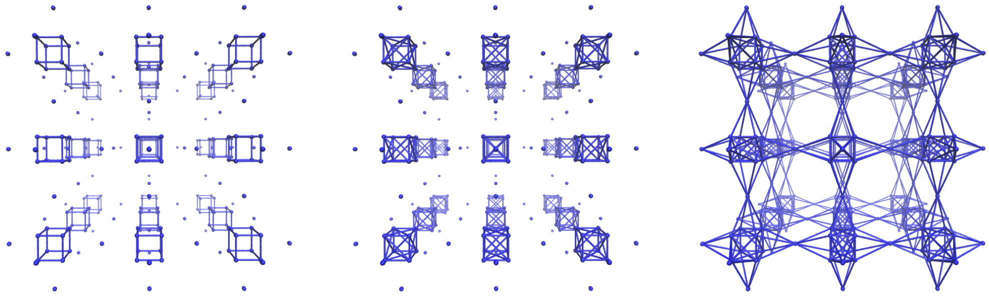

Figure 3.6: Connecting the network sites on the basis of an increasing distance threshold d_t.

current network. DFS is specifically designed to recursively search in every node of a graph structure. Basically, starting from a root node the algorithm explores all the neighbor nodes and recursively tagging the visited sites. If at a certain point all the nodes become tagged, the network is considered connected. This procedure is depicted in Fig.3.6. The final network model is represented in Fig.3.7. This model can be used as a template for discrete diffusive models. By transforming the molecular trajectories into time series according to the SOAP-PCA mapping, one can compute the marginal probabilities of finding a guest molecule in one of the two clusters and also the transition probability between the two states. Those are the basic ingredients for the parameterization of the network model that can be used to represent the adsorption and diffusive behaviour under a Markov chain approach.

3.6 Conclusions and perspectives

The SOAP-PCA-PAMM protocol succeded in sorting out the local molecular environments encountered in the MD simulations and in determining the most representative patterns. We decided to analyze only the adsorption patterns, thus

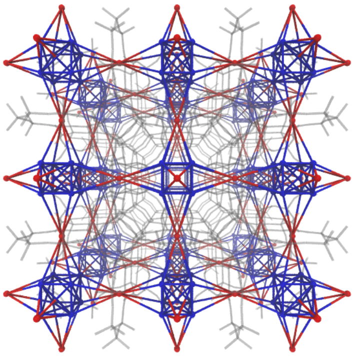

Figure 3.7: Final network model of CO_2 in the ITQ-29 zeolite, overlapped with the host material structure (in grey). The sites are coloured according to the respective cluster id (blue for cluster 1 and red for cluster 2).

ignoring the guest-guest correlations. For this reason, each SOAP vector was centered in each C atom and the environment density was constructed using only the O atoms of the host — thus deliberately ignoring all the other species. The cutoff of the SOAP descriptors was chosen to be 4 Å in order to embed all the information contained in the first coordination shell. The application of PAMM resulted in the separation of the reference data in two clusters: the first one representing the guest molecules sitting in the middle part of each pore, and the second one related to the guest molecules sitting between two neighboring pores.

This mapping can be used to compare the different local environment probability distributions of the same system under different thermodynamic conditions. In our case, this kind of analysis was conducted comparing simulations at different temperature and CO_2 concentration values.

One can interpret and use the molecular patterns as discrete states to obtain a coarse-grained representation of the reference system. This can be done in two ways: by defining a stochastic walk in pattern space (i.e. among the landmarks in the SOAP-PCA projection), or by backmapping the landmark states in real space. In this work, we constructed a simple network model through the latter approach. However, the model we obtained should be used with caution since the backmapping and the definition of the site-to-site links were determined with arbitrary assumptions. For example, different sites are connected by following a simple distance criterion rather than according to the transitions observed in the reference simulations — in general, the two networks may not necessarily be equal.

However it is remarkable that with our method, the algorithm reproduced the pore-pore connectivity correctly. This suggests that such network model could be used to compare different host-guest systems on the basis of the connectivity (as seen from the point of view of the local patterns).

Another issue that should be addressed is that the patterns found for this system are rather trivial and the whole procedure seems to be over-complicated with respect to the straightforward results. In fact, a simple algorithm based on counting the number of neighboring O atoms could have yield the same results. Also, the guest-guest correlations were completely ignored during the analysis, as they might influence the resolution of host-guest pattern classification. However, this is still an ongoing work and we will apply the same procedure to systems that may exhibit a richer variety or simply more complicated patterns. For example, more complex host materials such as hierarchical pore structures or disordered materials will be object of investigation. Also, if more than one guest species is involved in the same system, an analysis which comprises also the guest-guest patterns might highlight cooperative states or scenarios characterized by competition for the same set of adsorption sites within the host material.

Chapter 4

Local free energy approximations for the coarse-graining of adsorption phenomena: the interacting pair approximation

Adapted with permission from F. G. Pazzona, G. Pireddu, A. Gabrieli, A. M. Pintus, and P. Demontis. Local free energies for the coarse-graining of adsorption phenomena: The interacting pair approximation. *J. Chem. Phys.*, 148:194108, 2018. “Copyright 2019, American Institute of Physics.”

4.1 Introduction

Despite the increasing availability of computing power, molecular simulations with atomistic detail suffer from severe limitations in the length and time scales, even when the interaction field is classical.

To reduce the number of degrees of freedom involved in a simulation, thus allowing simulations to be carried out over wider scales, is the scope of coarse-graining techniques. In the coarse-graining of a molecular system, the original, fine-grained (FG) interaction field is mapped into an effective field that depends on a smaller number of variables, and the mapping is carried out in such a way that some selected properties of the FG system and of the coarse-grained (CG) model reasonably match. Since such properties are defined on a scale that is usually larger than the one at which the FG system evolves, this comes at the cost of a certain loss of information.

In the literature, the coarse-graining of molecular systems is approached in a variety of ways. Many of such approaches are *topological*; that is, each CG coordinate groups together several atoms of the FG system, and interacts with the other CG coordinates through effective fields that can be built from structure, [70–72] or via a force-matching procedure [13, 14, 19, 20] (the two approaches leading to the same results [73]), or through iterative Boltzmann inversion [21–23] or else through Gaussian Approximation Potentials [74] and cluster expansion techniques, [75] just to mention some—we do not mean to make an exhaustive list here. Besides topological strategies, a *spatial* coarse-graining approach also exists, which maps portions of a continuous simulation space, as well as groups of FG discrete *sites*, into a coarser lattice of *cells*. [25–32, 76] A *cell state* can be constructed out of what it contains, which could be, very naturally for molecular systems, the number of molecular centers-of-mass of each chemical species that occupy its physical space.

It is the application of the latter spatial approach to the coarse-graining of adsorption phenomena at equilibrium that we intend to discuss in this work. By keeping in mind the picture of small *guest* molecules adsorbed inside the pores of some *host* material, we will identify each cell as a pore, and the state of each one of them as the *occupancy*, which we define as the number of molecular centers-of-mass it hosts—not to be confused with the *loading*, with which we will indicate the average pore occupancy. For simplicity, we will discuss the case of only one guest chemical species in the system, but extension to multispecies models is straightforward.

Occupancy-based models of adsorption/diffusion, where a CG interaction field is defined over local occupancies in the nearness of discrete locations rather than on fine-grained atomistic configurations, are frequently encountered in the literature on host-guest systems. [77–84] According to how detailed should the CG model be, these locations may represent *adsorption sites*, which usually can be empty or occupied by one guest, or pores, which often can be occupied by more than one guest. Depending on the affinity between the host material and the guest species, adsorption sites may emerge naturally within the adsorption pores as well-defined locations that bind the guest molecules more strongly than others. This is the case for, e.g., benzene in silicalite [78], methane in the zeolite ITQ-29 (a.k.a. ZK4) [85], and benzene in zeolite Na-Y, [79, 86, 87] just to mention some. In such cases, a CG version of the grand canonical partition function can be constructed by modeling the adsorption sites as mutually exclusive lattice nodes equipped with a proper adsorption energy, while the guest-guest interactions can be represented as pairwise-additive free energies (such assumption provides a satisfactory approximation especially at low densities, where many-body contributions are proved to be relatively unimportant [20]), plus, if necessary to improve the model quality, inclusion of next neighbor interaction terms. [80] Further additional interactions, expressed in the form of dependency on some collective (but still *local*) variables, [88] may be also necessary. In any case, it is preferable to work with *local*, rather than global interaction energies because, besides a number of other general drawbacks, [89] the dependence of effective potentials on global density imposes severe limitations to transferability, e.g., to inhomogeneous systems. [90]

Identifying the *pores* of an ordered microporous material, rather than adsorption sites, as the elementary units of a discrete space domain represents an even coarser description of adsorption. A pore is usually allowed to contain more than one guest molecule, and this makes the resulting CG model a so-called 'multiparticle lattice-gas'. [34] When strong confinement holds and the density is not high, the correlation between molecules located inside different pores is often found to be weaker than inside the same pore. If that is the case, a CG interaction field can be satisfactorily formulated as a function of individual, uncorrelated pore occupancies, at least at room temperature (depending on the system, this might happen to be not true at lower temperatures). [41] Assuming such a strict locality of interactions allows for a very simple and efficient description of both the thermodynamics and the kinetics of particle pore-to-pore jumps. [91, 92] If accounting for pore-pore interactions becomes necessary, pairwise additivity can still be assumed at low densities, so that we can factorize the resulting CG grand partition function into elementary terms that, in principle, can be estimated out of a proper statistical sampling of the FG system itself.

When dealing with the calculation of approximated partition functions in general, [93] factorization is really a crucial point. Somewhat radical, oversimplifying approximations usually lead to 'friendly' CG partition functions, made of independent (or nearly independent) factors that often can be evaluated easily, but often such approximations suffer from a narrow range of applicability. On the other hand, more broadly acceptable approximations usually go along with a much more difficult evaluation of the constituting factors of the CG partition function—ironically, estimating them might end up requiring the use of *further* approximations.

Therefore, a balance needs to be found between the accuracy of the approximations on which the CG model is based and the actual computability of its parameters. In the present chapter, we discuss the formulation of a CG grand partition function for host-guest systems in which effective interactions, which are portrayed by both self- and pair-interaction terms, are defined over pore occupancies. We propose a modification of an existing CG model [83] of interactions of such kind, that significantly widens its applicability to a larger density range. In our formulation, effective pair interactions are, although still local, related to the occupancy correlations that can be observed between neighboring pores within a given range of densities.

Our discussion will proceed as follows. First, in Sec. 4.2, we will briefly summarize how the CG grand partition function is formulated, based on pore occupancies rather than molecular positions. In Sec. 4.3, we will formulate a relation between local CG interactions and occupancy distributions in the FG system, with mean-field corrections taking into account the effect of the neighborhood of any single pore and of any pore pair. In Sec. 4.4, we will compare our basic CG relations to an earlier, simpler theory were the surroundings of a pore pair is not taken account of in any way, and we will also show how, under less general circumstances, the parameterization we propose here reduces to the model we proposed in a previous work. [41] In Secs. 4.5.1 and A.2, we will apply our method to the coarse-graining of FG systems of two kinds: a lattice-gas where local free energies can be computed exactly and a Lennard-Jones system of united-atom methane molecules in the static field of zeolite ITQ-29. We will assess the validity of our coarse-graining approach by comparing the adsorption isotherms and occupancy distributions of the FG systems with their CG counterparts, and we will draw conclusions in Sec. 4.6.

4.2 Local, Coarse-grained interactions

Our general FG model of reference will be a system of small guest molecules hosted inside an ordered microporous material, which is represented as a network, $\mathcal{L} = \{\ell_1, \ldots, \ell_M\}$, of M pores with *local* connections, meaning that the molecules inside a pore, e.g., pore i, interact with the inner surface of the pore itself, with the molecules inside the same pore, and with the molecules hosted in the ν neighboring pores. Interactions with pores located beyond the first neighborhood are neglected (this is often a fair assumption, since in several microporous materials, like LTA- and FAU-type zeolites, the pore size is approximately equal or larger than 12 Å, which in most cases is near the customary cutoff radius for Lennard-Jones interactions). The system is assumed to be in contact with a thermal bath and a reservoir of molecules, so that both the temperature, T (we will indicate with β the 'inverse temperature', $\beta = 1/k_BT$, where k_B is the Boltzmann's constant), and the chemical potential, μ, are held fixed and uniform throughout the whole system, while the energy and the total number of guest molecules are allowed to vary.

For every possible configuration of guest molecules in the system, we can count how many of them fall within each pore, and then measure a global occupancy configuration, $\{n_1, \ldots, n_M\}$, indicating that pore 1 contains n_1 guests, pore 2 contains n_2 of them, etc. We assume then that

(i) every single pore, say pore i, contributes to the free energy of the entire system by an amount H_{n_i} and

(ii) the interaction between two neighboring pores, say i and j, contribute by an additional amount K_{n_i,n_j}.

The quantities Q_{n_i} and Z_{n_i,n_j} can be conveniently introduced:

$$Q_{n_i} = \exp\big(-\beta H_{n_i}\big), \tag{4.1}$$

$$Z_{n_i,n_j} = \exp\big(-\beta K_{n_i,n_j}\big). \tag{4.2}$$

H and Q are defined over properties of one single pore, and therefore, we will refer to either of them as 'self-interaction terms'. K and Z contain information about pore pairs, and therefore, we will refer to either of them as 'pair-interaction terms'. The most detailed description of the structure of the CG system is provided by the global occupancy distribution, [83] $p_\mu(n_1, \ldots, n_M)$, i.e., the probability of pore 1 having occupancy n_1, pore 2 having occupancy n_2, etc.,

$$p_\mu(n_1, \ldots, n_M) = \frac{1}{\Xi_{\text{CG}}} \prod_{i=1}^{M} e^{\beta\mu n_i} Q_{n_i} \prod_{j\in\mathcal{L}_i} \sqrt{Z_{n_i,n_j}}, \tag{4.3}$$

where $\mathcal{L}_i$ is the list of the ν neighbors of pore i. In Eq. (4.3), the normalization constant Ξ_{CG} is the *CG grand partition function*:

$$\Xi_{\text{CG}} = \sum_{n_1} \cdots \sum_{n_M} \prod_{i=1}^{M} e^{\beta\mu n_i} Q_{n_i} \prod_{j\in\mathcal{L}_i} \sqrt{Z_{n_i,n_j}}, \tag{4.4}$$

where the square root is introduced to correct for counting the pair-interaction terms twice. A simple Monte Carlo algorithm that samples the distribution in Eq. (4.3) is provided in the supplementary material of our previous work [41].

In Eq. (4.4), Q_{n_i} plays the role of the 'effective partition function of a single pore constrained to occupancy n_i'. Z_{n_i,n_j} instead plays the role of the 'contribution to the configuration integral of a pore pair constrained to occupancies n_i, n_j, due to the interaction of the n_i molecules in pore i with the n_j molecules in pore j'.

The scope of our coarse-graining approach here would be to formulate CG interaction terms such that, once used in a CG (lattice) simulation, they allow for the CG model to produce a global occupancy distribution, $p_\mu(n_1, \ldots, n_M)$, in good agreement with its FG counterpart, $P_\mu(n_1, \ldots, n_M)$ (throughout the whole paper, lowercase p's will indicate CG probabilities, whereas capital P's will refer to the FG system). We used 'would be' rather than 'is' because, in practice, the M-variated histogram $p_\mu(n_1, \ldots, n_M)$ can be estimated for none but the smallest systems. Therefore, we will seek agreement in terms of simpler (namely, uni- and bi-variated) distributions. As long as the assumed locality of interactions holds, we can reasonably expect that a good agreement in terms of local distributions will entail agreement also on a larger scale.

One important aspect we would like to remark is that we want CG interactions to be *local*; therefore we require both Q_{n_i} and Z_{n_i,n_j} *not* to depend on chemical potential; i.e., we want the same set of self- and pair-interaction terms to be portable within a whole range of densities, from infinite dilution to saturation.

Let us now discuss the meaning of the interaction terms Q_{n_i} and Z_{n_i,n_j} on a statistical-mechanical basis. Further details about the connection between FG and CG partition function and occupancy distribution are provided in Appendix 4.7. Q_{n_i} is commonly seen as the canonical partition function of the pore i when it contains exactly n_i guest molecules; i.e., $Q_{n_i} = z_{n_i}/\Lambda^{3n_i} n_i!$ where Λ is the De Broglie thermal wavelength and z_{n_i} is the following configuration integral:

$$z_{n_i} = \int_{v_i} \mathrm{d}\mathbf{r}_{i1} \cdots \int_{v_i} \mathrm{d}\mathbf{r}_{in_i} e^{-\beta U_i(\mathbf{r}_{i1}, \ldots, \mathbf{r}_{in_i})}, \tag{4.5}$$

where U_i denotes the potential energy experienced by the n_i molecules hosted inside pore i due to their interaction with the host material and with each other, given that their coordinates inside the pore are $\{\mathbf{r}_{i1}, \ldots, \mathbf{r}_{in_i}\}$ and the coordinates of each molecule are integrated over the volume ascribed to pore i. In other words, the pore described by Q_{n_i} is a small *closed* system. In principle, however, molecular configurations inside neighboring pores are correlated. Therefore, assigning Q_{n_i} a fixed value, although being very convenient, might seem quite unnatural. The pair term, Z_{n_i,n_j}, is thus introduced in order to account for such correlations.

The accepted meaning [83] of Z_{n_i,n_j} is that of the ratio between the configuration integral of two pores with occupancies n_i, n_j and the product of the individual pore configuration integrals z_{n_i} and z_{n_j},

$$\begin{aligned} Z_{n_i,n_j} \sim & \frac{1}{z_{n_i} z_{n_j}} \int_{v_i} \mathrm{d}\mathbf{r}_{i1} \cdots \int_{v_i} \mathrm{d}\mathbf{r}_{in_i} \int_{v_j} \mathrm{d}\mathbf{r}_{j1} \cdots \int_{v_j} \mathrm{d}\mathbf{r}_{jn_j} \\ & \times e^{-\beta U_{ij}(\mathbf{r}_{i1},\ldots,\mathbf{r}_{in_i},\mathbf{r}_{j1},\ldots,\mathbf{r}_{jn_j})}, \end{aligned} \tag{4.6}$$

where U_{ij} is the potential energy experienced by the molecules inside pore i and pore j due to the interaction with the host material and with each other, given that the n_i molecules in pore i are configured according to the coordinates $\{\mathbf{r}_{i1}, \ldots, \mathbf{r}_{in_i}\}$, and that the n_j molecules in pore j are configured according to the coordinates $\{\mathbf{r}_{j1}, \ldots, \mathbf{r}_{jn_j}\}$. With the symbol $\sim$ in (4.6), we remark that we prefer to assume a weaker relation than equality. This is because relation (4.6) refers to a system made of two pores, i and j, occupied by n_i and n_j guest molecules, respectively, as if it were 'extracted' from the system where it belongs and sampled separately from it, whereas in general the surroundings of any pair of neighboring pores *do* affect the correlations between them.

In a previous work, [41] we proposed an estimation of effective free energies based on a very simple reductionistic model, in which the surroundings of a given pore were taken account of, but, in order to derive an equation for the pair contributions that could be solved straightforwardly, the neighbors' occupancies were all constrained to the same value. In Sec. 4.3, we will introduce a more accurate model in which the constraint on the neighbors' occupancies is relaxed, and mean-field (occupancy dependent) correction terms are added to the free energy in the attempt to overcome the limitations of relation (4.6).

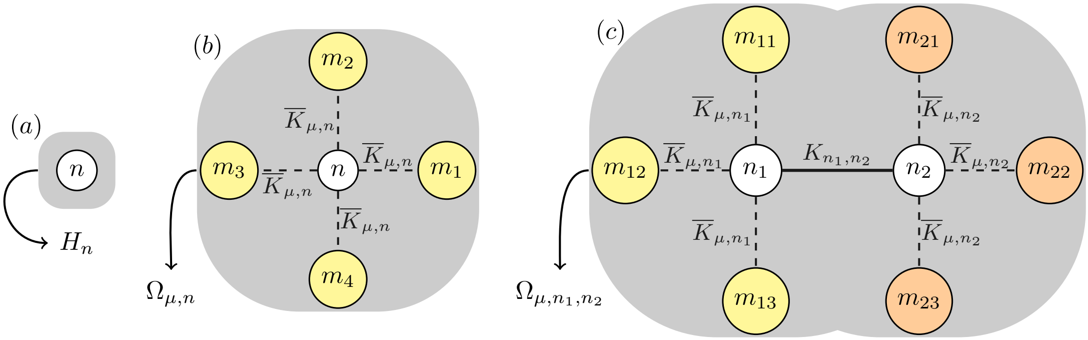

Figure 4.1: A sketch of the role played by each interaction term in the basic equations of our coarse-graining strategy. In this graphic example, the reference FG system is a square lattice of pores, in which each pore is connected to $\nu = 4$ neighbors. The pores represented by yellow and red (larger) circles are mean-field pores. White (smaller) circles represent non-mean-field pores. In (a), we consider a single, n-occupied closed pore whose equilibrium properties are related to the self-interaction term H_n. In (b) and (c), we consider the FG system as a whole, and from its equilibrium properties, we derive the pair-interaction terms: in (b), the ν neighbors of a single, n-occupied pore contribute to the CG potential, each one by adding a mean-field contribution $\overline{K}_{\mu,n}$ to the self-interaction H_n; in (c), the pores in a connected pair are assumed to interact with each other through the non-mean-field pair term K_{n_1,n_2} that adds to the self-terms H_{n_1} and H_{n_2}, and their interactions with the rest of the system are approximated by two mean-field terms, $\overline{K}_{\mu,n_1}$ and $\overline{K}_{\mu,n_2}$, each with multiplicity $\nu - 1$.

4.3 Coarse-graining under the Interacting Pair Approximation

Let us reformulate the problem in terms of simpler probability mass functions than $p_\mu(n_1, \ldots, n_M)$. Temperature and volume will be assumed constant throughout the entire discussion. For a given value of chemical potential, μ, we will consider the following distributions:

$p^o_\mu(n)$: probability of a pore to be occupied by n molecules, when interactions with *all* the other pores are neglected;

$p_\mu(n)$: probability of a pore to be occupied by n molecules, with interactions with every one of the ν pore neighbors represented as a mean-field, $\overline{K}_{\mu,n}$; and

$p_\mu(n_1, n_2)$: probability of a pore pair, made of pores 1 and 2, to show the occupancy pair n_1, n_2 with the effective interactions between the two pores given by K_{n_1,n_2}; the interactions between pore 1 and every one of its remaining $\nu - 1$ neighbors represented as a mean-field $\overline{K}_{\mu,n_1}$, and the interactions between pore 2 and every one of its remaining $\nu - 1$ neighbors represented as a mean-field $\overline{K}_{\mu,n_2}$.

The distributions $p^o_\mu(n)$, $p_\mu(n)$, and $p_\mu(n_1, n_2)$ are defined in terms of the potential functions, which we call *CG potentials*, $\Omega^o_\mu(n)$, $\Omega_\mu(n)$, and $\Omega_\mu(n_1, n_2)$, respectively, according to:

$$p^o_\mu(n) = (\zeta^o_\mu)^{-1} \exp\{-\beta\Omega^o_\mu(n)\}, \tag{4.7}$$

$$p_\mu(n) = \zeta_\mu^{-1} \exp\{-\beta\Omega_\mu(n)\}, \tag{4.8}$$

$$p_\mu(n_1, n_2) = \xi_\mu^{-1} \exp\{-\beta\Omega_\mu(n_1, n_2)\}, \tag{4.9}$$

where ζ^o_μ, ζ_μ, and ξ_μ are normalization constants, and, following the Bethe-Peierls mean-field approximation, [94, 95] the CG potentials are defined as follows:

$$\Omega^o_\mu(n) = -\,\mu n + H_n, \tag{4.10}$$

$$\Omega_\mu(n) = \Omega^o_\mu(n) + \nu\overline{K}_{\mu,n}, \tag{4.11}$$

$$\begin{aligned}\Omega_\mu(n_1, n_2) = {} & \Omega^o_\mu(n_1) + \Omega^o_\mu(n_2) + K_{n_1,n_2} \\ & + (\nu - 1)(\overline{K}_{\mu,n_1} + \overline{K}_{\mu,n_2}).\end{aligned} \tag{4.12}$$

H_n, the free energy of a *closed* n-occupied pore, and K_{n_1,n_2}, the contribution to the free energy provided by the interaction between the n_1 molecules located in pore 1 and the n_2 molecules located in pore 2, were already introduced in Eq. (4.2).

By definition, $K_{n,0} = 0$; i.e., there is no effective interaction energy between the molecules inside a pore and an empty pore.

Mean-field terms like $\overline{K}_{\mu,n}$ are used as corrections to the free energy. They can be thought as $\overline{K}_{\mu,n} \sim \sum_m p_\mu(m|n)K_{n,m}$, with $p_\mu(m|n) = p_\mu(n,m)/p_\mu(n)$, even though, as we are going to show, there is no need to compute mean-field interactions explicitly. In other words, when we consider a single pore in the system, as in Eq. (4.11), $\overline{K}_{\mu,n}$ accounts for the interaction between the n molecules inside that pore and the molecules in its ν neighbors. The number of such surrounding molecules, although it is related to μ, is not specified anywhere; therefore, such ν neighbors can be thought as mean-field pores. When a pore pair of occupancy (n_1, n_2) is considered instead, as we do in Eq. (4.12), we account for the rest of the system in terms of $2(\nu - 1)$ surrounding mean-field pores, $\nu - 1$ of which interact with cell 1 through the potential $\overline{K}_{\mu,n_1}$, while the other $\nu - 1$ ones interact with cell 2 through the potential $\overline{K}_{\mu,n_2}$. In order to obtain a solvable system of equations, we assume mean-field neighbors to not interact with each other.

The crucial point in Eqs. (4.11) and (4.12) is that, although the mean-field terms are μ-dependent, the pair interaction terms, K_{n_1,n_2}, *do not depend on* μ.

In Fig. 4.1, we sketched the role of the interaction terms used in Eqs. (4.10)–(4.12). The closed-pore equation, Eq. (4.10), does not contain any mean-field term—in some sense, it is 'exact,' meaning that if we were able to estimate with infinite accuracy the probability distribution $p^o_\mu(\cdot)$, e.g., by an infinitely long grand canonical sampling [by the grand canonical Monte Carlo (GCMC) method [5]] of a version of the FG system where only pore 1 can be occupied and all the pores in the system stay empty, we could retrieve H_n from the difference $\Omega^o_\mu(n) - \Omega^o_\mu(n')$, where $n' \neq n$, knowing that $H_0 = 0$. From Eq. (4.10), we have

$$H_n - H_{n'} = \mu(n - n') + \Omega^o_\mu(n) - \Omega^o_\mu(n'),$$

and by making use of Eqs. (4.1) and (4.7), we obtain the equivalent relation

$$\frac{Q_n}{Q_{n'}} = e^{-\beta\mu(n-n')}\frac{p^o_\mu(n)}{p^o_\mu(n')}, \tag{4.13}$$

which we can use to estimate Q_n from the probability ratios $p^o_\mu(n)/p^o_\mu(n')$, knowing that $Q(0) = 1$ can be used as starting point. Resorting to ratios like $Q_n/Q_{n'}$ rather than calculating every Q_n directly from $p^o_\mu(n) = (\zeta^o_\mu)^{-1}e^{\beta\mu n}Q_n$ [see Eqs. (4.1), (4.7), and (4.10)], is motivated by the fact that we do not know in advance the normalization constant ζ^o_μ.

The ratio in Eq. (4.13) does not depend on the chemical potential, meaning that, in principle, when carrying out the calculation of the R.H.S. of Eq. (4.13), one should recover the same result independently of the value of μ at which the probabilities were evaluated. In practice, however, numerical simulations are carried out over a finite time. Therefore, when replacing $p^o_\mu(n)$ and $p^o_\mu(n')$ with $P^o_\mu(n)$ and $P^o_\mu(n')$, i.e., the probabilities estimated from simulations of the FG system (with all the pores kept empty except for one), the R.H.S. of Eq. (4.13) will return a slightly different value for each μ, that is,

$$\frac{Q_n}{Q_{n'}} \approx e^{-\beta\mu(n-n')}\frac{P^o_\mu(n)}{P^o_\mu(n')}. \tag{4.14}$$

A proper combination of the ratios in Eq. (4.13) computed at different values of μ is the strategy we (successfully) used in our previous work [41] to obtain very reasonable results.

Once we computed the array of Q's (or H's) from GCMC on a single pore, we can proceed to the evaluation of the pair-interaction parameters K_{n_1,n_2} appearing in Eq. (4.12). By knowledge of the difference in CG potential

$$\Omega_\mu(n_1,n_2) - \Omega_\mu(n'_1,n'_2) = -\frac{1}{\beta}\ln\frac{p_\mu(n_1,n_2)}{p_\mu(n'_1,n'_2)},$$

where n'_1 and n'_2 are chosen to be not simultaneously equal to n_1 and n_2, we obtain an equation that relates them with $K_{n_1,n_2} - K_{n'_1,n'_2}$. First of all, we use Eq. (4.11) to eliminate the mean-field terms from Eq. (4.12), that then, through Eq. (4.10), becomes

$$\Omega_\mu(n_1,n_2) = \frac{1}{\nu}\Big[H_{n_1} + H_{n_2} - \mu(n_1+n_2)\Big] \\ + \left(1-\frac{1}{\nu}\right)\Big[\Omega_\mu(n_1) + \Omega_\mu(n_2)\Big] + K_{n_1,n_2}, \quad (4.15)$$

from which we obtain the relation

$$K_{n_1,n_2} - K_{n_1',n_2'} = \Omega_\mu(n_1,n_2) - \Omega_\mu(n_1',n_2') \\ + \frac{1}{\nu}\Big[\mu(n_1+n_2-n_1'-n_2') + H_{n_1'} + H_{n_2'} - H_{n_1} - H_{n_2}\Big] \\ + \left(1-\frac{1}{\nu}\right)\Big[\Omega_\mu(n_1') + \Omega_\mu(n_2') - \Omega_\mu(n_1) - \Omega_\mu(n_2)\Big]. \quad (4.16)$$

Knowing that $K_{n,0} = K_{0,n} = 0$ for any n, Eq. (4.16) can be used to calculate the matrix elements K_{n_1,n_2}. We can use Eqs. (4.1), (4.2), (4.8), and (4.9) to re-formulate Eq. (4.16) as an equation for the ratio $Z_{n_1,n_2}/Z_{n_1',n_2'}$:

$$\frac{Z_{n_1,n_2}}{Z_{n_1',n_2'}} = \left(e^{-\beta\mu(n_1+n_2-n_1'-n_2')}\frac{Q_{n_1'}Q_{n_2'}}{Q_{n_1}Q_{n_2}}\right)^{\frac{1}{\nu}} \\ \times \left(\frac{p_\mu(n_1')p_\mu(n_2')}{p_\mu(n_1)p_\mu(n_2)}\right)^{1-\frac{1}{\nu}} \frac{p_\mu(n_1,n_2)}{p_\mu(n_1',n_2')}, \quad (4.17)$$

which, while being completely equivalent to Eq. (4.16), shows very clearly the connection with occupancy probabilities. In the R.H.S. of Eq. (4.17), the mean-field interactions, appearing in Eqs. (4.11) and (4.12), are accounted for through the $1/\nu$ exponent on the first term (regarding the properties of a *lone* cell) and through the ratio involving single-cell probabilities, raised to the power of $1 - 1/\nu$.

We can write down an equation by which the physical meaning of pair-interaction terms will appear very intuitive. To do so, we first introduce the observed-to-expected (o/e) ratio,

$$C_\mu(n_1,n_2) = \frac{p_\mu(n_1,n_2)}{p_\mu(n_1)p_\mu(n_2)}, \quad (4.18)$$

whose deviation from unity is a measure of the correlations between the neighbor pore occupancies n_1, n_2, and the ratio

$$D_\mu(n) = \frac{p_\mu(n)}{p^o_\mu(n)}, \tag{4.19}$$

which measures the amount by which the mean-field neighborhood of a single pore causes its properties to deviate from the closed-pore case. Now, if we consider that the guest-guest interaction between two pores with no guests inside is null (so that $Z_{0,0} = 1 \Rightarrow K_{0,0} = 0$), then we can see that the pair terms have the following meaning:

$$\begin{aligned} K_{n_1,n_2} = &-\frac{1}{\beta}\Big[\ln C_\mu(n_1,n_2) + \frac{1}{\nu}\ln\left[D_\mu(n_1)D_\mu(n_2)\right] \\ &- \ln C_\mu(0,0) - \frac{2}{\nu}\ln\left[D_\mu(0)\right]\Big], \end{aligned} \tag{4.20}$$

where the terms $\ln C_\mu(0,0)$ and $\frac{2}{\nu}\ln D_\mu(0)$ are related to the occupancy pair $0, 0$, taken as a reference state. All terms in the R.H.S. of Eq. (4.20) depend on μ, but for each μ they change such as to return the same value. According to Eqs. (4.10)–(4.12), for a given pair of neighboring occupancies n_1, n_2, the R.H.S. of Eq. (4.17) must be the same at all chemical potentials. Therefore, one can formally remove the dependence on μ from Eq. (4.20) by integrating it over a range that goes from μ_i, corresponding to very low density, to μ_f, corresponding to very high density, close to saturation. In this way, the terms related to the reference state, i.e. the ones in which both the pores of the pair are empty, will appear as a single constant:

$$\begin{aligned} K_{n_1,n_2} = &-\frac{1}{(\mu_\mathrm{f}-\mu_\mathrm{i})\beta}\int_{\mu_\mathrm{i}}^{\mu_\mathrm{f}} \mathrm{d}\mu\Big[\ln C_\mu(n_1,n_2) \\ &+ \frac{1}{\nu}\ln\left[D_\mu(n_1)D_\mu(n_2)\right]\Big] + \text{const.} \end{aligned} \tag{4.21}$$

Although only formally, Eq. (4.21) provides us with the meaning of the CG pair interaction terms, consistent with the assumptions made in Eqs. (4.11) and (4.12); that is, *except for a constant term, contributions to the pair free energy* K_{n_1,n_2}

come from the correlation between the neighbor occupancies n_1 and n_2 and from the effect of the local surroundings on each of the two pores (divided by the pore connectivity ν), at all *the chemical potentials in the range $\mu_\mathrm{i} < \mu < \mu_\mathrm{f}$.*

Eqs. (4.17) and (4.21) cannot be used directly for the calculation of the pair-interaction terms because they require knowledge of the coarse-grained p_μ distributions, which are unknown. Therefore, we need a key assumption in order to convert our mean-field formulation of this problem into an operative coarse-graining strategy. Our proposal is to replace the unknown distribution p_μ, with the distribution obtained by numerical simulation of the FG system, P_μ. This amounts to saying that at any μ in the range $\mu_\mathrm{i} < \mu < \mu_\mathrm{f}$, the approximation

$$P_\mu(n_1, n_2) \approx p_\mu(n_1, n_2), \tag{4.22}$$

holds for every occupancy pair n_1, n_2. We will refer to the approximation (4.22), together with Eqs. (4.11) and (4.12), as *Interacting Pair Approximation* (IPA), to emphasize that we considered the pair of pores as a physical region that is not kept away from the rest of the system but rather interacts with its surroundings through mean-field correction terms. As an immediate consequence of the fact that relation (4.22) is an approximation, once we replaced the theoretical p_μ with the numerical distribution P_μ, we have that the R.H.S. of Eq. (4.17) becomes only approximately equal to the ratio $Z_{n_1,n_2}/Z_{n_1',n_2'}$:

$$\begin{aligned}\frac{Z_{n_1,n_2}}{Z_{n_1',n_2'}} \approx &\left(e^{-\beta\mu(n_1+n_2-n_1'-n_2')}\frac{Q_{n_1'}Q_{n_2'}}{Q_{n_1}Q_{n_2}}\right)^{\frac{1}{\nu}} \\ &\times\left(\frac{P_\mu(n_1')P_\mu(n_2')}{P_\mu(n_1)P_\mu(n_2)}\right)^{1-\frac{1}{\nu}}\frac{P_\mu(n_1,n_2)}{P_\mu(n_1',n_2')}.\end{aligned} \tag{4.23}$$

In other words, *in practice*, different chemical potentials will contribute differently to the estimation of the ratio $Z_{n_1,n_2}/Z_{n_1',n_2'}$. Among all such contributions, we can identify some values of μ that we want to contribute more than other ones, because they correspond to situations in which the pore occupancies n_1, n_2, n_1', and n_2' are visited frequently enough for us to reckon that our estimation of the probabilities $p_\mu(n_1)$, $p_\mu(n_2)$, $p_\mu(n_1, n_2)$, $p_\mu(n_1')$, $p_\mu(n_2')$, and $p_\mu(n_1', n_2')$ is accurate enough (e.g., if the probabilities are larger than some threshold). Conversely, we want μ values at which those pore occupancies are sampled rarely to contribute *less*, since in those cases our estimation of the probabilities is expected to be rather inaccurate. Extreme situations, i.e., values of μ at which some or all of the occupancies n_1, n_2, n_1', and n_2' are never sampled, should then give no contribution to $Z_{n_1,n_2}/Z_{n_1',n_2'}$. This might cause some Z_{n_1,n_2} to remain unknown, [41] but this does not really represent an issue, as long as the computable entries of the matrix Z ensure that the probability distribution that can be obtained by simulation of the resulting coarse-grained system and their FG counterparts reasonably match at all chemical potentials. Further details are discussed in the supplementary material. In the Appendices 4.8.1 and 4.8.2 we describe two possible routes for the estimation of the interaction terms Q_n and Z_{n_1,n_2}. In the first one, reported also in our previous work, [41] and indicated here as 'one-chemical-potential-at-a-time' (OCT), in a first stage, we make use of Eqs. (4.13) and (4.17) recursively *for each chemical potential*, thus obtaining μ-dependent CG interactions, and in a second stage, we remove the μ-dependency through a weighted average. In the second one, that we indicate as 'choose-the-best-ratio' (CBR), we select the μ for which the R.H.S. of Eq. (4.23) can be regarded as the best representative of the ratio $Z_{n_1,n_2}/Z_{n_1',n_2'}$; e.g., by using, as selection criterion, how large and how similar the probabilities $P_\mu(n_1, n_2)$ and $P_\mu(n_1', n_2')$ are, and then, we use the ratios we selected to calculate recursively the individual entries of the matrix Z. Essentially, the differences in the interaction matrix Z obtained using either of the two methods are very small, while a much more crucial role is played by the accuracy in the probability histograms evaluation from GCMC, as we briefly discuss in Appendix 4.8.3.

4.4 Comparison with previous models

It is worthwhile to compare our coarse-graining (IPA) approach with the more drastic assumption in which a pair of neighboring pores is treated as if it was uncorrelated with the rest of the FG system. [82–84] We will indicate the latter assumption as *Non-Interacting Pair Approximation* (NIPA).

NIPA relies on relation (4.6) *taken as if it were an equality.* To compare IPA and NIPA, we find it convenient to write the IPA equation for the pair interaction terms, i.e., Eq. (4.17), as follows:

$$\begin{aligned}\frac{Z_{n_1,n_2}}{Z_{n'_1,n'_2}} =& \left(\frac{p^o_\mu(n'_1)p^o_\mu(n'_2)}{p^o_\mu(n_1)p^o_\mu(n_2)}\right)^{\frac{1}{\nu}} \left(\frac{p_\mu(n'_1)p_\mu(n'_2)}{p_\mu(n_1)p_\mu(n_2)}\right)^{1-\frac{1}{\nu}}\\ &\times \frac{p_\mu(n_1,n_2)}{p_\mu(n'_1,n'_2)}, \end{aligned} \tag{4.24}$$

If relation (4.6) was an equality, we could drop the mean-field terms in Eq. (4.12), thus obtaining the NIPA equation for the pair interactions:

$$\frac{Z^*_{n_1,n_2}}{Z^*_{n'_1,n'_2}} = \frac{p^o_\mu(n'_1)p^o_\mu(n'_2)}{p^o_\mu(n_1)p^o_\mu(n_2)} \frac{p^*_\mu(n_1,n_2)}{p^*_\mu(n'_1,n'_2)}, \tag{4.25}$$

where $p^*_\mu(n_1,n_2)$ is the probability of a pair of neighboring pores *separated from the rest of the system* to show the occupancy pair n_1, n_2, given that the chemical potential is μ. The first major problem with NIPA is that the adsorption isotherm of a closed pair is, at high densities, different from the adsorption isotherm of the FG system as a whole (as shown in the Supplementary Material for the case of the Lennard-Jones system we will discuss in Sec. A.2). Therefore, in general, the NIPA and IPA occupancy distributions are expected to be also different. Moreover, we can see by comparing the NIPA Eq. (4.25) with the IPA Eq. (4.24) that, when switching from NIPA to IPA, inclusion of the mean-field corrections causes the single-pore NIPA term in the R.H.S. of Eq. (4.25),

$$\frac{p^o_\mu(n'_1)p^o_\mu(n'_2)}{p^o_\mu(n_1)p^o_\mu(n_2)},$$

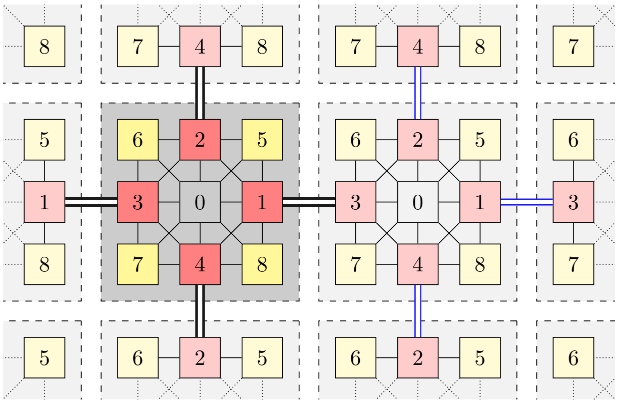

Figure 4.2: Structure of the lattice-gas we studied in this work to compare the IPA with the NIPA coarse-graining approach. Sites, which can assume either state 0 (empty) or 1 (singly occupied), are represented as small squares, which can be grouped into cells (gray shades). A number is assigned to each site within every cell to distinguish from one another. Site-site interactions are pairwise, and they take place between connected sites—connections are displayed as lines, which are thin if the connected pair entirely belongs to one cell and thicker (and doubled) if they connect two sites that belong to different cells.

to split into two factors, in the R.H.S. of Eq. (4.24),

$$\left(\frac{p^o_\mu(n'_1)p^o_\mu(n'_2)}{p^o_\mu(n_1)p^o_\mu(n_2)}\right)^{\frac{1}{\nu}}\left(\frac{p_\mu(n'_1)p_\mu(n'_2)}{p_\mu(n_1)p_\mu(n_2)}\right)^{1-\frac{1}{\nu}},$$

that is, one independent-pore contribution, raised to the power of $1/\nu$, where a single pore of occupancy n is taken as if it were a closed system, and one correlated-pore contribution, raised to the power of $1-1/\nu$ (and therefore, more important than the first one), which istead relates the properties of a single pore to its surroundings in the FG system, via mean-field correction terms. Therefore, use of the NIPA matrix Z^* will in general ensure the correct coarse-graining of only a special version of the FG system, in which only two pores are non-empty, but not of the FG system as a whole. Since the correlations between any pore and its surroundings becomes of crucial importance at high density, the IPA matrix Z is expected to provide, in general, a more accurate CG representation.

Before we proceed further with Sec. 4.5, it is worth mentioning the conditions under which the IPA strategy described here reduces to the coarse-graining strategy we proposed in a previous work, [41] where a CG equation for the ratio $Z_{n_1,n_2}/Z_{n_1-1,n_2}$ was derived by constraining the occupancies in the neighborhood of a given pore to the same value. By letting $n_1' = n_1 - 1$ and $n_2' = n_2$, we can rewrite Eq. (4.17) as

$$\frac{Z_{n_1,n_2}}{Z_{n_1-1,n_2}} = \left(e^{-\beta\mu}\frac{Q_{n_1-1}}{Q_{n_1}}\right)^{\frac{1}{\nu}} \left(\frac{p_\mu(n_2|n_1)}{p_\mu(n_2|n_1-1)}\right)^{1-\frac{1}{\nu}}, \tag{4.26}$$

where $p_\mu(n_2|n_1)$ is the conditional probability of a pore, belonging to a pair of neighboring pores, to have occupancy n_2, given that the other pore has occupancy n_1. We can see that the basic CG expression we proposed in our previous work is retrieved when the last factor in the R.H.S. of Eq. (4.26) can be neglected (i.e. when it is ~ 1). This happens under the approximation $p_\mu(n_2|n_1) \approx p_\mu(n_2|n_1 \pm 1)$, that represents a less general case where the conditional distribution $p_\mu(\cdot|n_1)$ does not vary much when the neighbor occupancy n_1 is *slightly* varied, thus implying weak (even though still non-null) lateral correlations.

4.5 Simulations and discussion

In this section, we apply both the IPA and the NIPA approaches to a lattice-gas system of interacting Boolean sites (Sec. 4.5.1) and to a Lennard-Jones system of confined particles (Sec. A.2). All the simulations were carried out by standard Metropolis GCMC. [5]

4.5.1 Lattice-gas with repulsive interactions

According to Eqs. (4.14), (4.23), and (4.25), the calculation of CG interaction parameters via IPA and NIPA relies on the previous knowledge of probability histograms. However, histograms are calculated from the outcomes of numerical simulations and are therefore unavoidably affected by accuracy issues. Extending the simulation length leads to a gain in accuracy, but in the case of off-lattice molecular simulation, this comes at a significant computational cost. Since we reckoned it to be very important to first test the IPA, in comparison with the NIPA, in an essentially inaccuracy-free environment, we did it on a computationally cheaper simulation model, i.e., a FG Boolean lattice-gas. Due to the finiteness (and discreteness) of their configuration space, simple Boolean interacting lattice-gases are an invaluable tool for comparing different coarse-graining strategies, such as IPA and NIPA.

The local contributions given by the Q array and the NIPA pair-interaction matrix Z^* can be (numerically) calculated exactly for a lattice-gas where cells are composed of a small number of $n_{\max}$ mutually exclusive sites, since in that case the integrals in Eqs. (4.5) and (4.6) reduce to summations over a large but finite number of configurations. Therefore, since in this case the estimation of Q and Z^* (NIPA) does not require numerical simulations, the application of NIPA to lattice-gases is *totally* unaffected by accuracy issues. In the IPA case, instead, evaluation of the Z pair-interaction terms cannot be performed by direct summation of Boltzmann weights, and Eq. (4.23) must be used, which is based on the knowledge of probability histograms numerically evaluated from simulations. As a consequence, application of the IPA on lattice-gases is not totally free from accuracy issues. Nevertheless, we reckon it to be very interesting to check whether the IPA would provide better results than the NIPA, even though accuracy issues affected the former more than the latter.

Our lattice-gas here is a square lattice of cells, each one made of nine sites arranged as a square as well. Every site can be either empty (occupancy 0) or occupied by one particle (occupancy 1). Neighboring sites, say i and j, interact with each other (lateral interactions) repulsively, according to the interaction energy of ϵ. With the aim of increasing the correlations, in some simulations we 'extended' the FG interactions by including an attractive interaction parameter, ψ, based on the number of occupied neighbors around each site:

$$E(\mathbf{s}) = \sum_{\langle i,j \rangle} s_i s_j \Big[\epsilon + \psi(M_i) + \psi(M_j) \Big], \tag{4.27}$$

where the sum runs over all the pairs of neighboring sites, and s_i and s_j are the occupancies of sites i and j, according to the occupancy configuration $\mathbf{s}$ of the whole FG lattice. M_i and M_j are defined as the total occupancy in the neighborhood of site i (including the occupancy of j) and of site j (including the occupancy of i), respectively, and

$$\psi(M) = \begin{cases} \phi, & M \geq M_0 \\ 0, & M < M_0 \end{cases}, \tag{4.28}$$

where $\phi < 0$. The energy $\psi(M)$ adds to the interaction between two neighboring sites if the number of occupied neighbors of each of them becomes equal or larger than some threshold value M_0, which we set at $M_0 = 4$.

In Fig. 4.2, the structure of a portion of the lattice is depicted. Interacting sites are joined by lines that are either thin or thick in the case of intra-cell and inter-cell connections, respectively. Intercell connections are represented in Fig. 4.2 as 'double' connections, but this does not imply that the interaction energy is doubled. GCMC simulations of this FG system under different setups of the interaction parameters were performed at several values of chemical potential, chosen such as to ensure that the resolution was at least of two density points between each interval $(\langle n \rangle, \langle n \rangle + 1)$ in the average cell occupancy. In Fig. 4.3, we show results for the following parameter settings: (a) $\epsilon = 4$ kJ mol^{-1} and $\phi = 0$, (b) $\epsilon = 8$ kJ mol^{-1} and $\phi = 0$, and (c) $\epsilon = 4$ kJ mol^{-1}, $\phi = -1.6$ kJ mol^{-1}, and $M_0 = 4$. For every chemical potential, two simulations were performed. In the first one,

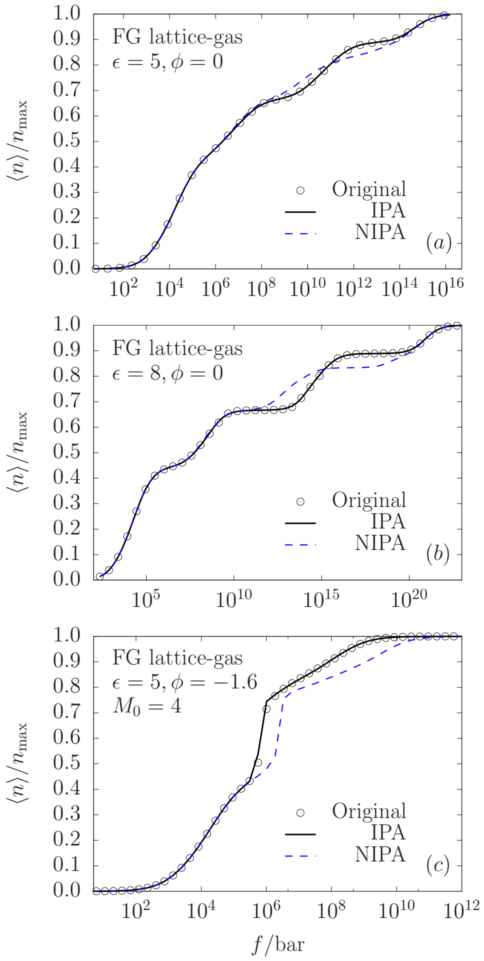

Figure 4.3: Adsorption isotherms for the lattice-gas system under different interaction setups. In (a) and (b), the site-site interaction is purely repulsive (it amounts to 5 and 8 kJ mol^{-1}, respectively). In (c), the lateral interaction is set at 5 kJ mol^{-1}, but extended attractive interactions are added. Results for the FG system are depicted as empty circles, whereas solid black lines are used for IPA and dashed blue lines for NIPA results. For the sake of readability, we reduced the density of points in the FG scatter plot to one-half of the actual dataset.

inter-cell interactions were neglected and the Q terms were evaluated from (4.14). In the second simulation, we included inter-cell interactions and evaluated the Z interaction terms through (4.23). Every simulation was carried out over a number of steps that varied from $N = 10^6$ to 10^7 moves, equally (and randomly) distributed

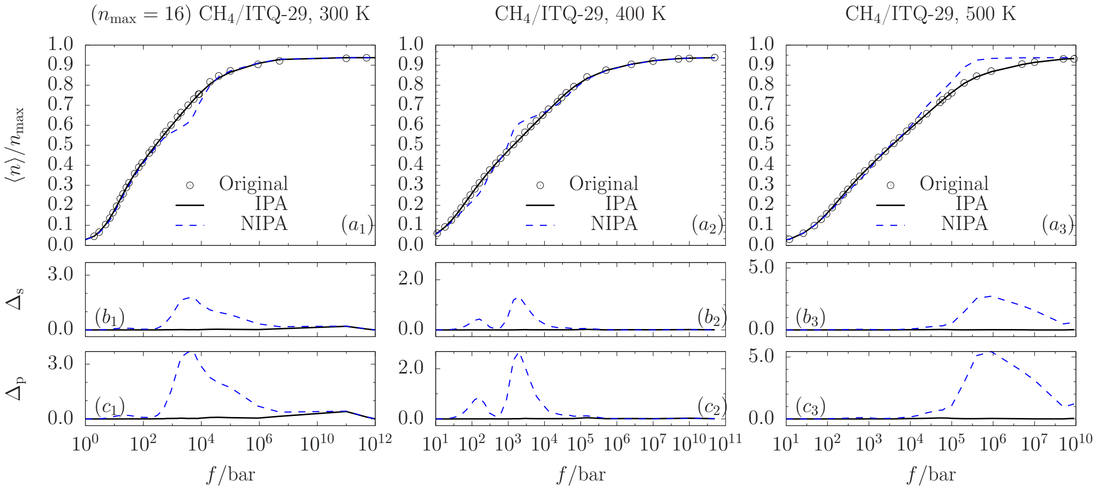

Figure 4.4: Adsorption isotherms and Kullback-Leibler divergence for a system of a Lennard-Jones methane molecules (united atom approximation) under the static field of zeolite ITQ-29 at the temperatures 300, 400, and 500 K. In subfigures a_1, a_2, and a_3, adsorption isotherms are shown (empty circles: FG lattice system; black solid lines: IPA; blue dashed lines: NIPA). In subfigures b_1, b_2, and b_3, the Kullback-Leibler divergence for the occupancy distribution of one single cell are shown [see Eq. (4.29)]. Subfigures c_1, c_2, and c_3, refer instead to the occupancy distribution one pair of neighboring cells [see Eq. (4.30)]. Black solid lines represent divergences between FG and IPA systems, blue dashed lines represent divergence between FG and NIPA systems.

among displacement, insertion, and deletion attempts. Simulations of both IPA and NIPA CG systems were performed through GCMC as well but over a smaller number of steps ($N \sim 10^5$) due to the much faster convergence to equilibrium. The results reported in this work are for lattice systems of 4 × 4 cells. Larger systems were explored (6 × 6 and 8 × 8) for a smaller number of GCMC moves and of chemical potential values, and they gave results that were indistinguishable from the ones obtained for the 4 × 4 cases.

For both the IPA and the NIPA coarse-graining, the results we reported were obtained through the CBR approach described in Sec. 4.3. However, both OCT and CBR provided nearly the same results. Adsorption isotherms, i.e., plots of the density (expressed as the average cell occupancy, $\langle n \rangle$, divided by the total number of sites per cell, n_{max}) *vs.* the fugacity (here meant as $f = f_0 e^{\beta\mu}$, where $f_0 = 1$ bar), are reported in Fig. 4.3, and they show that the NIPA approach starts failing at intermediate-high densities, where intercell correlations become important. On the other hand, IPA provides isotherms (see Fig. 4.3) and occupancy distributions (see supplementary material) in good agreement with the FG system at all densities. In particular, in the example shown in Fig. 4.3(a), at high densities, pair correlations are non-negligibly affected by the presence of the other neighbors of both cells of the pair, and this causes the adsorption isotherm of the whole FG system to exhibit curvature changes that are not well reproduced by NIPA. In Fig. 4.3(b), a more repulsive site-site interaction enhances this phenomenon, and the isotherm tends toward a step-like shape as repulsion is increased. In this case, the more quantitative agreement provided by the IPA approach is even more evident. The isotherm in Fig. 4.3(c) is related to a more extreme case, where, due to the increasingly important effect of the attractive contribution from $\psi(M)$ to the total energy, see Eqs. (A.2) and (4.28), site correlations extend to the second neighborhood. One can immediately figure out that extended interactions may cause cell pairs to be correlated very differently, depending on whether we consider every pair as if it was part of a larger portion of the system (as in the IPA approach), or as if it evolved on its own, detached from the rest of the system (as is the NIPA approach). As a consequence of the balance between repulsive and attractive interactions, a larger step appears in the isotherm at intermediate densities, and as the density approaches the step (for $\langle n \rangle / n_{\text{max}}$ between 0.4 and 0.5), the NIPA method fails. On the contrary, IPA better preserves the shape of the original system, indicating that also in this case the cell-cell correlations induced by more complicated FG interactions are well represented through the inclusion of the mean-field terms in Eqs. (4.11) and (4.12).

We remark that in the calculations above, the NIPA interaction terms were evaluated as exact sums rather than through simulations of a pair of cells, so they are not affected by any accuracy issue, whereas the IPA interaction terms were calculated straight from the distributions obtained from simulations of the FG system—therefore, contrarily to the NIPA case, IPA parameters are supposed to be not immune to noise and accuracy issues (related to the fact that low-probability occupancies are unavoidably sampled less frequently, and then less accurately, than the high-probability ones); despite everything, the IPA reveals the most accurate of the two. However, as we will see in Sec. A.2, in systems where the structure is determined by a much smoother potential energy function, the difference between IPA and NIPA, although undeniably present, appears less marked and starts becoming non-negligible at higher densities.

4.5.2 Lennard-Jones particles under the influence of an external field

Methane molecules, represented by the united atom approximation as Lennard-Jones (LJ) spheres, confined in the all-silica zeolite ITQ-29 (formerly called ZK4) have been widely used in the literature as a host-guest system to test statistical-mechanical theories, adsorption-diffusion models, methods for the calculation of free energy profiles, and coarse-graining approaches under various computational environments (like kinetic Monte Carlo and Cellular Automata). [81–85, 96–102] The ITQ-29 framework is particularly interesting because of its peculiar structure of relatively wide pores (when compared with methane size), called α-cages ($\sim$ 11.4 Å in diameter), arranged in a simple cubic network ($\nu = 6$), and interconnected through narrower eight-ringed windows ($\sim$ 4.5 Å in diameter), allowing the passage of one methane molecule at a time. We modeled guest-guest and host-guest interactions according to the force fields used by Dubbeldam *et al.* [98] with a cutoff of 12 Å, and since the zeolite flexibility does not affect significantly the sorption properties of methane (although it would be not negligible for larger molecules [103]), a pre-tabulation of the host-guest potential energy on a grid of ~ 0.2 Å of spacing allowed for a significant reduction of the process time of the simulations. [104] Our framework system consisted of a grid of $4 \times 4 \times 4$ pores, corresponding to $2 \times 2 \times 2$ unit cells (the ITQ-29 unit cell we used consisted of eight pores). GCMC simulations were carried out using the standard Metropolis acceptance-rejection method for displacements, insertions, and deletions. [5] Such MC moves were performed in equal proportions, within a total number of post-equilibration steps that varied from $\sim 10^6 N_{\mathrm{uc}}$ to $\sim 10^8 N_{\mathrm{uc}}$, with N_{uc} as the average number of molecules per unit cell. The temperatures we investigated were 100, 200, 300, 400, and 500 K. The fugacities were chosen in such a way as to explore loadings more or less uniformly (at least two points within each loading interval from $\langle n \rangle$ to $\langle n \rangle + 1$) from ~ 0.1 up to ~ 14.5 molecules per pore. In all cases, methane molecules were not allowed to enter the sodalite cages nor the double six-ringed cages. At 100 K, due to the very low acceptances at the highest loadings, simulations were carried out up to ~ 12 molecules per pore. Due to the very simple (cubic) topology of the pore network, and since methane-methane interactions across non-first neighboring pores can be safely neglected, [41] the $\mathrm{CH_4}$/ITQ-29 system is especially suited for testing the IPA coarse-graining scheme as well.

In Fig. 4.4, we compare results for the temperatures 300, 400, and 500 K. At such temperatures, the CBR approach provided slighlty better IPA representations, whereas slightly better NIPA results were obtained by using the OCT protocol. Besides adsorption isotherms, we wanted to give the reader a quick idea on how the use of IPA rather than NIPA affects the occupancy distributions of the CG model, in comparison with the distributions that emerge from the GCMC

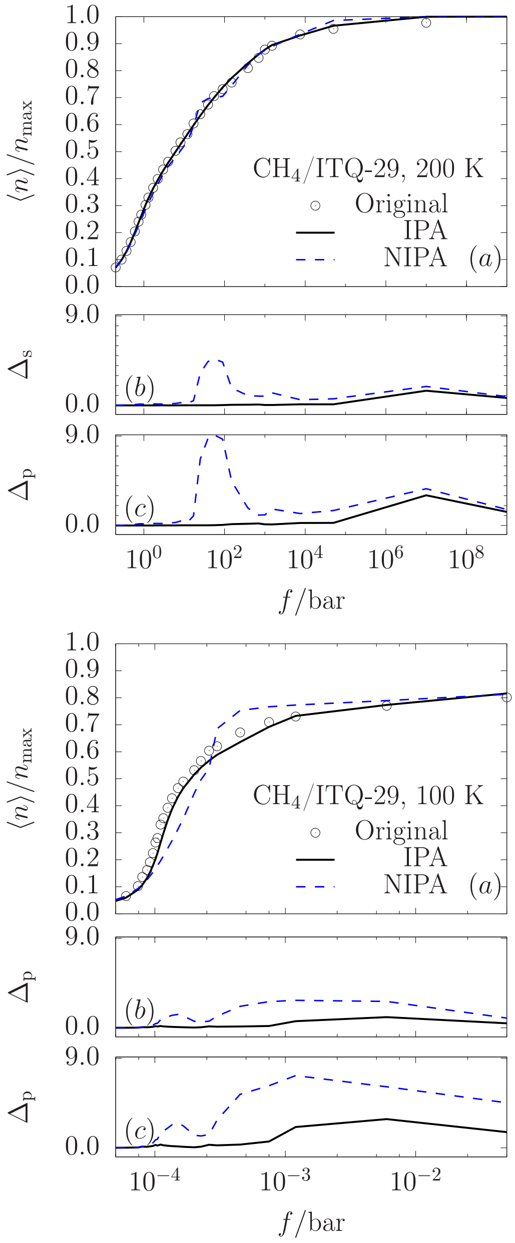

Figure 4.5: Adsorption isotherms and Kullback-Leibler divergences for a system of a Lennard-Jones methane molecules (united atom approximation) under the static field of zeolite ITQ-29 at the temperatures 100 and 200 K. The meaning of dots, line types, and line colors is the same as in Fig. 4.4.

simulations of the FG system. Since two kinds of histogram were constructed out of GCMC simulations at every chemical potential (one univariate histogram for the probability of any pore to have occupancy n and one bivariate histogram for the probability of any pore pair to show the occupancy pair n_1, n_2), in order to be able to visualize the results on a single figure per system, here we decided to compare occupancy distributions through the Kullback-Leibler (KL) divergence, that we used according to the symmetric definition given by Kullback and Leibler in their original article. [105]

We will refer to Δ_s as the KL divergence for the probability distribution of a single pore and to Δ_p as the KL divergence for the probability distribution of a pore pair:

$$\Delta_\mathrm{s} = \sum_n \Big(P_\mu(n) - p_\mu^\mathrm{cg}(n)\Big) \ln \frac{P_\mu(n)}{p_\mu^\mathrm{cg}(n)}, \tag{4.29}$$

$$\Delta_\mathrm{p} = \sum_{n_1} \sum_{n_2} \Big(P_\mu(n_1, n_2) - p_\mu^\mathrm{cg}(n_1, n_2)\Big) \ln \frac{P_\mu(n_1, n_2)}{p_\mu^\mathrm{cg}(n_1, n_2)}, \tag{4.30}$$

where P_μ and p_μ^cg refer respectively to the occupancy distribution of the FG system and of one of two possible CG systems (IPA and NIPA). Based on the resulting FG distributions, we set the maximum pore occupancy at $n_\mathrm{max} = 15$. We included more detailed comparisons of the occupancy distributions in the supplementary material.

As we anticipated at the end of Sec. 4.5.1, the discrepancies between IPA and NIPA are less evident here than in the case of lattice-gases with repulsive interactions, due to the smoothness of the LJ potentials. Nevertheless, the IPA approach shows to be the most accurate in all the cases reported, proving its robustness despite its simplicity. At low loadings, both approaches provide a reasonable agreement between CG and FG systems, but at intermediate-high loadings, non-negligible KL divergences between the NIPA and the FG distributions appear, in correspondence with discrepancies in the adsorption isotherms (as expected), and they are much more pronounced than the ones we find for the IPA case. We believe this is due to the presence of the mean-field terms in the basic equations of the IPA approach, Eqs. (4.11) and (4.12), which satisfactorily accounts for the effect of the whole neighborhood of each pore.

In Fig. 4.5, we report results at lower temperatures, namely, 200 and 100 K. The IPA parameters that produced the CG plots in Fig. 4.5 were calculated by the CBR method at 200 K, and by the OCT method at 100 K, whereas the NIPA parameters were evaluated through the OCT method at both temperatues. Also in these cases, the difference between the parameters obtained by the two methods is not so much evident, and we made our choice based on slight discrepancies.

At these temperatures, correlations between neighboring pores become more evident. Noticeably, at 200 K, while the IPA and NIPA isotherms are approximately in the same (good) agreement with the FG system, the occupancy distributions are not, and the IPA results are closer to the FG distributions, especially at low densities.

At the temperature of 100 K the NIPA approach fails to provide a reasonable agreement even at low loadings, indicating that in this case the occupancy of each pore is seriously affected by the occupancies *in the whole neighborhood.* Including only one neighbor in the statistical description of CG interactions, as NIPA prescribes, does not allow the CG model to reproduce, not even partially, the correlations observed in the FG system. The FG occupancy distributions at 100 K become highly non-central for all but the lowest loadings (this can be seen very clearly in Fig. S9 reported in the supplementary material), and we noticed that, although still providing more satisfactory results than NIPA, the agreement in the CG occupancy distributions as provided by the IPA approach becomes less striking than at higher temperatures. In particular, bimodality, which we also observed for the system at 200 K and which corresponds to states with two coexisting phases, [106–108] is not accurately reproduced. We believe this not to be an issue of the mean-field corrections as they are formulated in Eqs. (4.11) and (4.12), but rather a limitation of the pairwise nature of the CG potential model. Inclusion of other correction terms that depend on collective, but still local, variables may further improve the agreement in situations where correlations between every pore and all its neighbors are *very* large. [88] This will be the subject of forthcoming investigations.

As a further note, testing the IPA to the case of Lennard-Jones particles in ITQ-29 was indeed very convenient because in this material the pore size and structure make the six first neighbors of each pore all equivalent. Extension of IPA to the case where each pore is surrounded by non-equivalent neighbors is the scope of a future work.

4.6 Conclusions

We investigated the coarse-graining of host-guest systems of small molecules adsorbed in a regular porous material, described in terms of occupancy distributions rather than fine-grained configurations of molecular positions. In such a reductionistic representation, the interaction field is based on the free energy of every single pore, defined as a function of its occupancy (i.e., the number of molecules it hosts), plus effective contributions to the free energy coming from the interactions between neighboring pore pairs. By means of a very simple system, i.e., a lattice-gas where local free energies can be calculated exactly, we have shown that the currently accepted approximation in which the pair interaction is assumed to be the same whether the pore pair is kept within the full fine-grained system it belongs or it is made independent of its surroundings [81–84] (we referred to it as NIPA, *non-interacting pair approximation*) turns out to be inaccurate at high densities, where the interactions between every pore pair and its neighborhood induce stronger correlations. In Lennard-Jones systems, where interactions are much smoother than in lattice-gases, the inadequacy of the NIPA approach is slightly less evident but, apart from the case of high temperatures (around room temperature and above) and low sorbate density, still leads to non-negligible discrepancies between the fine-grained system and its coarse-grained counterpart. We improved the calculation of coarse-grained interactions by establishing a relation between local occupancy distributions of the fine-grained systems and the properties of a coarse-grained, occupancy-based model that we called IPA (*interacting pair approximation*), where the effect of the surroundings on both single pores and pore pairs is taken account of via mean-field terms. As a result, the pore pair interactions appear as if they were entirely related to the local pore-pore correlations, and to the discrepancy between the properties of a closed single pore and those of a pore which instead does interact with its neighbors. We remark that although in the basic IPA equations mean-field corrections depend on chemical potential (i.e., they are density-dependent), the resulting coarse-grained interactions do *not* depend on it; i.e., their local nature is preserved. We presented results for the coarse-graining of lattice-gases with repulsive interactions and for a host-guest model of methane molecules (treated as Lennard-Jones spheres) confined in zeolite ITQ-29. In every case we studied, the IPA approach provided noticeably better results than NIPA. In the majority of cases, the agreement between the properties of the coarse-grained systems obtained under the IPA approach and the properties of the original, fine-grained system was excellent.

4.7 Connection between fine-grained and coarse-grained global occupancy distribution

In order to see how the model described by Eq. (4.3) relates to its FG counterpart, we will take as an example the case of an off-lattice system of identical, indistinguishable particles hosted in a microporous material consistent with the description we provided at the beginning of Sec. 4.2, in the grand-canonical ensemble. Let us first examine the FG probability, $P_\mu(\mathbf{r}^{n_1}, \ldots, \mathbf{r}^{n_M})$, that, given a configuration of indistinguishable particles distributed in M identical pores according to the occupancies $n_1, \ldots, n_M$, the positions of the n_i particles in the i-th pore (with $i = 1, \ldots, M$ indicating each pore) are described by the set of coordinates $\mathbf{r}^{n_i} = \{\mathbf{r}_{i1}, \ldots, \mathbf{r}_{in_i}\}$, where $\mathbf{r}_{ik}$ represents the coordinates of the k-th particle in the i-th pore:

$$P_\mu(\mathbf{r}^{n_1}, \ldots, \mathbf{r}^{n_M}) = \frac{1}{\Xi_{\mathrm{FG}}} \prod_{i=1}^{M} \frac{e^{\beta\mu n_i}}{\Lambda^{3n_i} n_i!} e^{-\beta U}. \tag{4.31}$$

In Eq. (4.31), $U = U(\mathbf{r}_{11}, \ldots, \mathbf{r}_{1n_1}, \ldots, \mathbf{r}_{M1}, \ldots, \mathbf{r}_{Mn_M})$ is the potential energy of the whole FG configuration, including both mutual interactions and interactions with the medium; Λ is the deBroglie wavelength; and the normalization constant, Ξ_{FG}, is the grand partition function. Eq. (4.31) is the extension of the configuration probability density in the Small System Grand Ensemble [see Soto-Campos *et al.* [109], Eq. (2)] to the case of multiple (and identical) subvolumes in the grand-canonical ensemble. The joint probability, $P_\mu(n_1, \ldots, n_M)$, of pore 1 to contain n_1 particles, pore 2 to contain n_2 particles, and so on, is obtained by integrating the coordinates of each particle over all the possible locations within the host pore:

$$P_\mu(n_1, \ldots, n_M) = \frac{1}{\Xi_{\mathrm{FG}}} \prod_{i=1}^{M} \frac{e^{\beta\mu n_i}}{\Lambda^{3n_i} n_i!} \int_{v_i} \mathrm{d}\mathbf{r}^{n_i} e^{-\beta U}, \tag{4.32}$$

where $\int_{v_i} \mathrm{d}\mathbf{r}^{n_i} = \int_{v_i} \mathrm{d}\mathbf{r}_{i1} \cdots \int_{v_i} \mathrm{d}\mathbf{r}_{in_i}$, and $\prod_i \int_{v_i} \mathrm{d}\mathbf{r}^{n_i} = \int_{v_1} \mathrm{d}\mathbf{r}^{n_1} \cdots \int_{v_M} \mathrm{d}\mathbf{r}^{n_M}$, so that in the R.H.S. of Eq. (4.32) the integrand function, $e^{-\beta U}$, is subjected to $\sum_{i=1}^{M} n_i$

nested integrations.

The grand partition function then reads

$$\Xi_{\mathrm{FG}} = \sum_{n_1} \cdots \sum_{n_M} \prod_{i=1}^{M} \frac{e^{\beta\mu n_i}}{\Lambda^{3n_i} n_i!} \int_{v_i} \mathrm{d}\mathbf{r}^{n_i} e^{-\beta U}. \tag{4.33}$$

In principle, the nested integrals in Eqs. (4.32) and (4.33) are not factorizable. The key assumption that connects Eq. (4.32) to Eq. (4.3) is [83] to first (i) recognize the following quantity

$$A_{n_1,\ldots,n_M} = -\frac{1}{\beta} \ln \prod_{i=1}^{M} \frac{1}{\Lambda^{3n_i} n_i!} \int_{v_i} \mathrm{d}\mathbf{r}^{n_i} e^{-\beta U}, \tag{4.34}$$

as the contribution to the Helmholtz free energy coming from the occupancy configuration $n_1, \ldots, n_M$. Please note that $A_{n_1,\ldots,n_M}$ is assumed to not depend on the chemical potential. Then, (ii) we assume to retain only the interaction energy between the molecules in each pore, say pore i, and the molecules located in the pores that belong to $\mathcal{L}_i$, i.e., the neighborhood of pore i we defined right after Eq. (4.3), while we neglect the interactions between molecules farther away from each other. Finally, (iii) we assume a "pore-pairwise" additivity of $A_{n_1,\ldots,n_M}$, in such a way that it can be expressed as

$$A_{n_1,\ldots,n_M} = \sum_{i=1}^{M} H_{n_i} + \frac{1}{2} \sum_{i} \sum_{j \in \mathcal{L}_i} K_{n_i,n_j}, \tag{4.35}$$

where, consistent with the definitions in Sec. 4.2, H_{n_i} is the contribution to the free energy due to the momenta of the n_i molecules located in pore i and to the potential $U_i(\mathbf{r}^{n_i}) = U_i(\mathbf{r}_{i1}, \ldots, \mathbf{r}_{in_i})$,

$$H_{n_i} = -\frac{1}{\beta} \ln \frac{1}{\Lambda^{3n_i} n_i!} \int_{v_i} \mathrm{d}\mathbf{r}^{n_i} e^{-\beta U_i(\mathbf{r}^{n_i})}, \tag{4.36}$$

where $U_i(\mathbf{r}^{n_i})$ includes the interactions of each molecule with the host material and the mutual interactions of the n_i molecules with each other [see Eqs. (4.1) and (4.5)]. In Eq. (4.35), K_{n_i,n_j} takes into account the correlation between the neighboring pores i and j due to mutual interactions between the n_i and the n_j guest particles [see Eq. (4.2)]. By using Eqs. (4.34) and (4.35) altogether in Eqs. (4.32) and (4.33) one obtains, respectively, the CG distribution and partition function in Eqs. (4.3) and (4.4). In other words, it is upon the three assumptions we described here that that the mapping from the FG probability distribution $P_\mu(n_1,\ldots,n_M)$ to its CG counterpart $p_\mu(n_1,\ldots,n_M)$ relies.

4.8 Estimation of the interaction terms

We describe two possible routes for the estimation of the interaction terms Q_n and Z_{n_1,n_2}.

4.8.1 'One-chemical-potential-at-a-time' (OCT)

In this strategy, first we obtain μ-dependent CG interactions, and then we remove the μ-dependency through a weighted average.

In the first step, we make use of Eqs. (4.14) and (4.23) to estimate the interaction parameters recursively *for each chemical potential*, with $n' = n+1$ in (4.14) and with $n_1' = n_1 - 1 \wedge n_2' = n_2$ [contributing with a weight $\propto P_\mu(n_1-1,n_2)$] and $n_1' = n_1 \wedge n_2' = n_2 - 1$ [contributing with a weight $\propto P_\mu(n_1,n_2-1)$] in (4.23). In the second step, the chemical potential-dependence of the parameters $Q_n(\mu)$ and $Z_{n_1,n_2}(\mu)$ resulting from the first step is removed by means of weighted averages

$$Q_n = \sum_{\mu\in\{\mu\}} \omega_n(\mu) Q_n(\mu),$$

and

$$Z_{n_1,n_2} = \sum_{\mu\in\{\mu\}} \omega_{n_1,n_2}(\mu) Z_{n_1,n_2}(\mu),$$

where the weights are set as proportional to the frequency with which each occupancy (or occupancy pair) was sampled in the FG systems of reference: $\omega_n(\mu) = P^o_\mu(n)/\sum_{\mu'\in\{\mu\}} P^o_{\mu'}(n)$, and $\omega_{n_1,n_2}(\mu) = P_\mu(n_1,n_2)/\sum_{\mu'\in\{\mu\}} P_{\mu'}(n_1,n_2)$.

4.8.2 'Choose-the-best-ratio' (CBR)

In this strategy, first we compute the μ-dependent ratios $R^\Delta_n(\mu)$ [defined as the R.H.S. of Eq. (4.14)] and $R^{\Delta_1,\Delta_2}_{n_1,n_2}(\mu)$ [defined as the R.H.S. of Eq. (4.23)], with $1 \leq \Delta \leq n$, $1 \leq \Delta_1 \leq n_1$, $1 \leq \Delta_2 \leq n_2$, and $\Delta_1 \neq \Delta_2$, and then (ii) we remove the μ-dependence by simply selecting, for each doublet $\{n, \Delta\}$ and for each quadruplet, $\{n_1, \Delta_1, n_2, \Delta_2\}$, the one value, R^Δ_n, out of the candidates $R^\Delta_n(\mu_1), R^\Delta_n(\mu_2), \ldots$, and the one value, $R^{\Delta_1,\Delta_2}_{n_1,n_2}$, out of the candidates $R^{\Delta_1,\Delta_2}_{n_1,n_2}(\mu_1), R^{\Delta_1,\Delta_2}_{n_1,n_2}(\mu_2), \ldots$, which we find to be the most representative ones. Our selection criteria are the magnitude of the probabilities involved (the higher, the better) and the difference between the probabilities at the numerator and the ones at the denominator (the more similar the probabilities, the more similar the accuracies). Finally, (iii) we apply the same criteria to select the best $R^{\Delta^*}_n$ out of the set $\{R^\Delta_n\}_{\Delta=1,\ldots,n}$, and the best $R^{\Delta^*_1,\Delta^*_2}_{n_1,n_2}$ out of the set $\{R^{\Delta_1,\Delta_2}_{n_1,n_2}\}_{\Delta_1,\Delta_2}$, and calculate recursively the self-interaction terms as

$$Q_n = R^{\Delta^*}_n Q_{n-\Delta^*}, \tag{4.37}$$

and the pair-interaction terms as

$$Z_{n_1,n_2} = R^{\Delta^*_1,\Delta^*_2}_{n_1,n_2} Z_{n_1-\Delta^*_1,n_2-\Delta^*_2}. \tag{4.38}$$

4.8.3 Probability threshold

In general, one cannot expect the OCT and CBR approach to provide *exactly coincident* parameter sets. In the simulations we performed, the two methods provided nearly equal results. A much more important role than the choice of the OCT or

CBR approach was played by the magnitude of the probabilities involved in (4.14) and (4.23), which were estimated as histograms from GCMC simulations of the FG system. We carried out the calculation of the ratios in (4.14) and (4.23) only for those values of μ for which the probabilities involved were above some threshold, that was set as $tP^{(1)}_{\max,\mu}$ for (4.14) [where $P^{(1)}_{\max,\mu}$ is the maximum probability value observed in the histogram $P_\mu(\cdot)$], and $tP^{(2)}_{\max,\mu}$ for (4.23) [where $P^{(2)}_{\max,\mu}$ is the maximum probability value observed in the bivariate histogram $P_\mu(\cdot,\cdot)$]. For each FG system, the threshold t was treated as an adjustable parameter with optimal values in the range $10^{-5} < t < 10^{-3}$.

Properly adjusting t led to significantly reducing the noise in the CG simulations produced by the inaccuracy that unavoidably affects the estimation of probabilities of infrequent events. Moreover, one should notice that the approximation (4.23) is, in general, weaker than the approximation (4.14). The reason is that, while approximation (4.14) relies on Eq. (4.13), which is exact [so that the histogram $P^o_\mu(\cdot)$, as evaluated from GCMC, suffers only from the finite simulation length], approximation (4.23) instead relies on Eq. (4.17), which in turn *relies on the approximation* $P_\mu(n_1, n_2) \approx p_\mu(n_1, n_2)$ rather than on some equality. That represents an additional source of noise in the evaluation of the pair terms Z_{n_1,n_2} and makes the role of the threshold t even more important in the evaluation of CG pair interactions.

Chapter 5

Spatial Coarse-graining of Methane Adsorption in Graphene Materials

Reprinted (adapted) with permission from (G. Pireddu, F. G. Pazzona, A. M. Pintus, A. Gabrieli, and P. Demontis. Spatial coarse-graining of methane adsorption in graphene materials. *J. Phys. Chem. C*, 123(30):18355–18363, 2019). Copyright (2019) American Chemical Society

5.1 Introduction

The representation of physicochemical phenomena involving molecular systems in a variety of spatial and temporal scales has always been a challenging task. Nowadays, atomistic computational methods such as *ab-initio* molecular dynamics, offer a very detailed and accurate framework for the study of molecular systems. [2] However, the simulation of relatively large environments requires a considerable computational effort. Even with atomistic classical molecular dynamics (MD) and Monte Carlo (MC) methods, the simulation of systems at the meso- and macroscopic scales remains unfeasible.

This makes the development of coarse-graining protocols an active line of research. With a possible slight loss of accuracy, the production of less-detailed but more computationally efficient models allows switching from a fine-grained (FG) to a coarse-grained (CG) representation of the system under investigation. In this line of work we think of such CG description in terms of occupancy-based models of adsorption, where an effective interaction field is defined over the *local occupancy* (that is the number of guest molecules' centers) in the nearness of discrete locations inside the adsorbent rather than on fine-grained atomistic configurations.[20, 77, 80, 81, 84, 85, 87, 110]

Thus, the coarse-graining approach we follow is of a *spatial* rather than *topological* kind; that is, instead of building CG units out of groups of atoms through mapping operators (which is, in a *very* few words, the spirit of topological coarse-graining [13, 14, 19, 21, 72, 75]), we turn our attention to the partitioning of the system domain in non-overlapping subvolumes and the association of proper CG state variables to each of them.[25, 26, 28, 32, 38, 41, 78, 110–112] In general, the idea of representing adsorption phenomena through a real-space lattice model is at least one century-old [113], but methods are still under continuous development, due to the lack of a sufficiently general and accurate protocol. [114–116]

Local occupancies are precisely the CG state variables we are focusing on here, and we represent them as discrete stochastic variables. The subvolumes we consider are of nanometer size and above, thus making the resulting CG model a *mesoscopic* model, and we evaluate the matching between the CG and FG representation in terms of statistical properties of occupancy distributions, while neglecting any detail of the original system below that scale. Our study then is aimed to define, at constant temperature, the *effective interactions* between neighboring subvolumes in terms of local occupancies only, within a wide overall density range. Our effort points towards the development of a general procedure for performing a bottom-up spatial CG of adsorption phenomena while guaranteeing a sufficiently accurate representation of static properties.

In our previous paper [38] we worked on host-guest systems where the neighbors of each adsorption unit (e.g., every α-cage of LTA-type zeolites) were all equivalent. Here, we extend our reasoning to the case where each subvolume is surrounded by neighborhoods of *two* kinds, by making reference to two systems that can be partitioned into two-dimensional square lattices: united-atom methane adsorbed (i) on a single graphene sheet, and (ii) between two graphene sheets. The latter system is inspired by carbon-based adsorbent materials, which can exhibit interesting properties for the adsorption of chemical species such as methane, which were investigated both computationally [61, 117, 118] and experimentally. [55, 56]

In this work we will show how to use occupancy histograms to establish a correspondence between the adsorption properties of the aforementioned fine-grained molecular systems and the static equilibrium properties of a local-occupancy based coarse-grained model. We will also show that in some cases (i.e., low temperature) the coarse-grained occupancy correlations in space can be appreciably improved in accuracy by pre-processing the occupancy histograms of the fine-grained systems through a quantized gaussian distribution model.

5.2 Coarse-grained model

In Fig. 5.1 we report a picture of a portion of the simulation space of one of our FG systems of interest: a graphene layer (the host) with united-atom methane molecules adsorbed on it (the guests). As sketched in Fig. 5.1, the space is tessellated with identical, non-overlapping square subvolumes, called *cells*, of edge

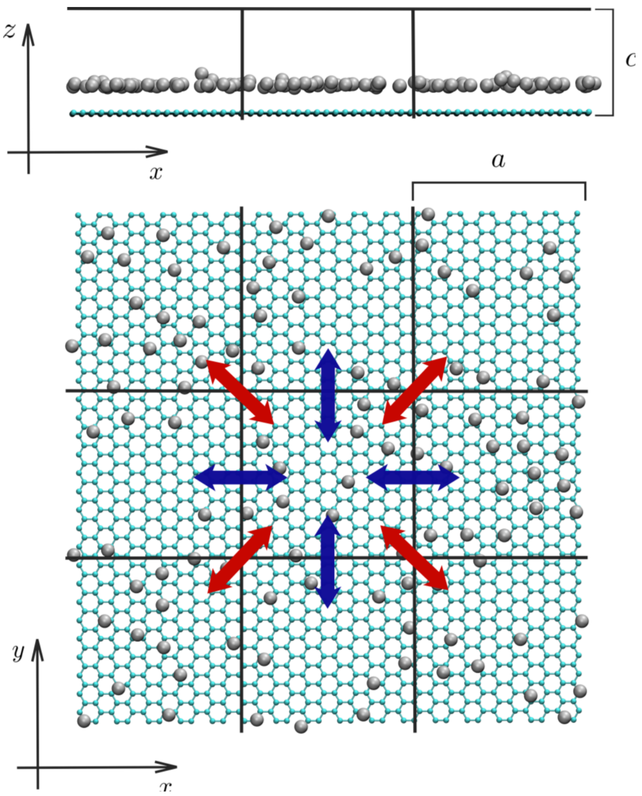

Figure 5.1: Two projections of the same snapshot from the FG simulation of the methane and single layer graphene system at 200 K. In both images the cell partitioning is represented by solid black lines. The bottom image also shows the neighboring classes for the central cell: blue arrows represent class I connections and red arrows represent class II connections.

length a. We say that two cells are neighbors of one another if they share either one edge (*class I neighbors*, center-to-center distance is equal to a) or one corner (*class II neighbors*, distance $a\sqrt{2}$). Therefore, each cell turns out to be connected to $\nu^{\mathrm{I}} = 4$ cells of class I, and $\nu^{\mathrm{II}} = 4$ cells of class II. The total number of neighbors is denoted as $\nu = \nu^{\mathrm{I}} + \nu^{\mathrm{II}} = 8$. By setting $a = r_c$, where r_c is the cutoff radius used for the potential energy evaluation in the FG simulations, we ensure that no guest molecule in any cell will interact directly with any other molecule outside the neighborhood of that cell. For any configuration of guest molecules in the space domain, we can count how many of their centers-of-mass fall within every cell; if we label the cells as $i = 1, \ldots, M$, with M as the total number of cells, the array of integer numbers that results from this counting operation is termed the *occupancy configuration* of the system, and is denoted as $\mathbf{n} = \{n_1, \ldots, n_M\}$. Effective interactions arise inside every cell and between neighboring cells, and neighboring cells of every one class contribute differently to the total effective interaction—this can be easily seen if we think of such interactions in terms of *average*, effective interactions between the n_i particles in cell i and the n_j particles in cell j: *on average*, the molecules in a cell will "feel" the molecules in the neighborhood of one kind differently from how they "feel" those in the neighborhood of another kind. We consider the system in the grand-canonical ensemble, which is the most common statistical ensemble used to represent adsorption phenomena. In this ensemble, the chemical potential, μ, of the guest species is held constant (along with the temperature T), while the overall density fluctuates around the corresponding equilibrium value. Due to guest-guest and host-guest interactions (defined on the molecular scale), any change in μ will cause the properties of the distribution of occupancies in the system to change as well; our aim is to provide our CG square cells with a set of effective, local occupancy-dependent interactions such as to produce (approximately) the same change in the distribution properties.

We define Ω, the CG potential function of the system in the grand-canonical ensemble, as a function of μ and of its occupancy configuration in the lattice:

$$\Omega_\mu(\mathbf{n}) = \sum_i \left(H_{n_i} - \mu n_i\right) + \sum_{\langle ij \rangle} K^{\chi_{ij}}_{n_i,n_j}, \tag{5.1}$$

where $\langle ij \rangle$ denotes a summation over neighboring cells, and χ is the neighboring class between cells i and j. In Eq. (6.1), H_{n_i} is the contribution to the total free energy of the system provided by the n_i guests that, according to the occupancy configuration array $\mathbf{n}$, are located in cell i, whereas $K^{\chi}_{n_i,n_j}$ is the contribution provided by the effective interaction between the n_i molecules in cell i and the n_j molecules in cell j, given that i and j are neighbors of class χ. The probability of configuration $\mathbf{n}$ to occur, $p_\mu(\mathbf{n})$, satisfies $p_\mu(\mathbf{n}) \propto \exp\{-\beta\Omega_\mu(\mathbf{n})\}$, with $\beta = 1/k_B T$, where k_B is the Boltzmann constant. It is the scope of our research to find a set of H's and K's [see Eq. (6.1)] such that the coarse-grained probability distribution $p_\mu(\mathbf{n})$ matches with the probability of configuration $\mathbf{n}$ estimated from classical GCMC simulations of the FG system; a requirement that we want H's and K's parameters to satisfy is *locality*, meaning that *they would not depend on any global variable other than temperature.*

Being H_{n_i} and $K^\chi_{n_i,n_j}$ meant as (local) free energies, the corresponding contributions to the partition function of the system are given by

$$Q_n = e^{-\beta H_n}, \qquad Z^\chi_{n_1,n_2} = e^{-\beta K^\chi_{n_1,n_2}} \tag{5.2}$$

respectively. In order to obtain the Q_n parameters, we first carry out GCMC simulations of *one single cell* of the FG system at several values of μ; for each one of them, we use the GCMC results to estimate the occupancy distribution $p^o_\mu(n)$, that is the probability that the cell we simulated contained exactly n guest molecules. For such one-cell system the CG potential is then

$$\Omega^o_\mu(n) = -\mu n + H_n \tag{5.3}$$

and its relation with the equilibrium probability of a cell to have occupancy n is $p^o_\mu(n) \propto e^{\beta\mu n} Q_n$. Therefore, for any two different occupancies n and n' we can write

$$\frac{Q_n}{Q_{n'}} = \frac{e^{-\beta\mu n}\, p^o_\mu(n)}{e^{-\beta\mu n'}\, p^o_\mu(n')}, \tag{5.4}$$

and use such relation to estimate the Q's recursively, starting from $H_0 = 0$ (or equivalently $Q_0 = 1$). As the accuracy of each bar of the $p^o_\mu(n)$ histogram we estimated from molecular GCMC would slightly vary from one chemical potential to the other, a weighting procedure such as the one described in our previous work[38] can be used to obtain the μ-independent set of Q's we are looking for.

In order to estimate the K's, i.e. the pair-interaction terms, we need to employ a different model, where additional assumptions are introduced. As different neighboring classes contribute differently to the total free energy of the system, we associate each one of them, say class χ (where $\chi = \text{I}$ or II), with its own set of probability distributions. Each element of such set is the bivariate occupancy distribution $p^{\chi}_{\mu}(n_1, n_2)$ computed at chemical potential μ. For any two specific values of n_1 and n_2, it represents the probability that two neighboring cells of neighboring class χ contain n_1 and n_2 guests, respectively, given that the chemical potential is μ. We estimated the histograms $p^{\chi}_{\mu}(n_1, n_2)$ from GCMC simulations of a 4×4-sized FG system where we neglected all the guest-guest interactions apart from (i) interactions between guests located in the same cell, and (ii) interactions between guests located in neighboring cells of neighboring class χ, and then we establish a proper connection between the bivariate occupancy histograms $p^{\chi}_{\mu}(n_1, n_2)$ and two mean-field models within the interacting pair approximation (IPA), namely one IPA model for neighborhood class I, and another one for neighborhood class II. Every such χ-IPA dedicated model is made of one pair of explicit cells ("1" and "2", respectively with occupancy n_1 and n_2; we call these cells *explicit* because n_1 and n_2 are assigned well-defined integer values) that are class χ neighbors of one another, *plus* $2\nu^{\chi} - 2$ surrounding cells with unspecified occupancy—i.e., $\nu^{\chi} - 1$ *mean-field* cells interacting with cell 1, and $\nu^{\chi} - 1$ more mean-field cells interacting with cell 2. The structure of the χ-IPA models and their role in the coarse-graining process is depicted in Fig. 5.2. The nature of such additional cells is *mean-field* in the sense that any information about their state stays hidden inside the global variable μ. We assume the guests in every such cell to interact only with the guests in either one of the two *explicit* cells of the pair (namely, cell 1 *or* cell 2); the effective interaction between an explicit cell of occupancy n and any of its $\nu^\chi - 1$ mean-field neighbors can be reasonably thought of as $\overline{K}^\chi_{\mu,n} \sim \sum_m K^\chi_{n,m} p^\chi_\mu(n,m)/p^\chi_\mu(n)$, with m as a fictitious occupancy of the mean-field cell. Such contribution is a μ-dependent mean-field term but, as we are about to show, mean-field terms will cancel out in the final formula for the pair interactions.

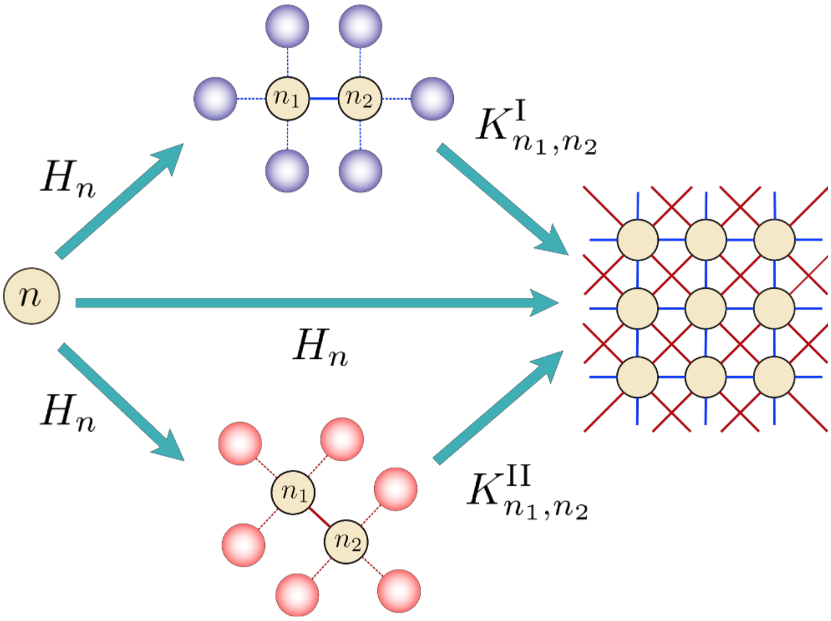

Figure 5.2: CG workflow scheme for a square lattice with classes I and II, and $\nu^{\mathrm{I}} = \nu^{\mathrm{II}} = 4$. On the left side: the closed single cell model, employed for the calculation of the H_n parameters. In the middle: I-IPA and II-IPA models for the calculation of the pair interaction parameters (I-IPA considers only neighbors of class I, while II-IPA considers only neighbors of class II). The mean field cells are indicated with a color gradient according to the respective classes: blue for class I and red for class II. On the right side: a portion of the resulting CG lattice model where each cell is connected to neighbors of both class I and class II.

Given the above considerations, for each χ-IPA model the total CG potential is

$$\begin{aligned}\Omega^\chi_\mu(n_1,n_2) =&\Omega^o_\mu(n_1) + \Omega^o_\mu(n_2) + K^\chi_{n_1,n_2}\\ &+ (\nu^\chi - 1)(\overline{K}^\chi_{\mu,n_1} + \overline{K}^\chi_{\mu,n_2}),\end{aligned} \tag{5.5}$$

where the Ω^o_μ terms are defined according to Eq. 5.3, and $\overline{K}^\chi_{\mu,n_1}$, $\overline{K}^\chi_{\mu,n_2}$ are mean-field interaction terms. Now, there are two basic assumptions we rely upon in this work: (i) the contribution from each class to the total free energy does not depend on the contribution from any other class, and (ii) each χ-IPA model is a good approximation of the reference system when *only* the interactions through the χ class and the interactions inside every cell are active. The first assumption enables us to write the CG potential for a *single* cell interacting with its ν^χ neighbors of class χ as

$$\Omega^\chi_\mu(n) = \Omega^o_\mu(n) + \nu^\chi \overline{K}^\chi_{\mu,n}, \tag{5.6}$$

whereas the second assumption establishes the proportionality between $\exp\left[-\beta\Omega^\chi_\mu(n_1, n_2)\right]$ and the $p^\chi_\mu(n_1, n_2)$, i.e. the histogram we evaluated through GCMC simulations of the FG system. If we consider another pair of occupancies (n'_1, n'_2) for two neighboring cells of class χ, we can eliminate the mean-field terms from (5.5) and (5.6), and obtain the following recurrence relation:

$$\begin{aligned}\frac{Z^\chi_{n_1,n_2}}{Z^\chi_{n'_1,n'_2}} &= \left(\frac{e^{\beta\mu n'_1} Q_{n'_1}\, e^{\beta\mu n'_2} Q_{n'_2}}{e^{\beta\mu n_1} Q_{n_1}\, e^{\beta\mu n_2} Q_{n_2}}\right)^{\frac{1}{\nu^\chi}} \\ &\times \left(\frac{p^\chi_\mu(n'_1)\, p^\chi_\mu(n'_2)}{p^\chi_\mu(n_1)\, p^\chi_\mu(n_2)}\right)^{1-\frac{1}{\nu^\chi}} \frac{p^\chi_\mu(n_1, n_2)}{p^\chi_\mu(n'_1, n'_2)},\end{aligned} \tag{5.7}$$

which starts with $Z^\chi_{0,n} = Z^\chi_{n,0} = Z^\chi_{0,0} = 1$. Eq. (5.7) becomes operative once we have knowledge of all the required probability histograms—which we gain from simulations of the FG system with the proper interaction settings. Also in this case, the weighting procedure described in our previous work[38] can be used to obtain a μ-independent set of Z's.

5.2.1 Data pre-processing at low T

.

According to Eqs. (5.4) and (5.7), the estimation of CG parameters relies on the occupancy histograms obtained from the GCMC simulation of the reference (FG) system, under a variety of conditions (i.e. by excluding some or all the interactions between molecules located in different cells). Now, GCMC simulations are finite; therefore, at each chemical potential, histogram bars in the nearness of the probability maximum will be better sampled than those far from it. At low temperatures, the noise and the irregular shape in GCMC histograms might partly compromise the accuracy of CG results in terms of occupancy correlations in space. In such cases we found out very effective to *process* the GCMC histograms *before* feeding them into the recurrence relations (5.4) and (5.7). The "processing" consists in replacing the original GCMC bivariate occupancy histograms, $p_{\mu}^{\chi}(\cdot,\cdot)$, with new distributions, $\pi_{\mu}^{\chi}(\cdot,\cdot)$, whose properties should approximate a number of selected properties (namely, marginal means, marginal variances, and covariance) of the original ones, but are "less noisy". We define these new distributions according to a bivariate quantized Gaussian distribution model:

$$\pi_{\mu}^{\chi}(n_1, n_2) \propto \exp\left[-\frac{z}{2(1-r^2)}\right], \tag{5.8}$$

where

$$z = \frac{(n_1-a_1)^2}{s_1^2} + \frac{(n_2-a_2)^2}{s_2^2} - \frac{2r(n_1-a_1)(n_2-a_2)}{s_1 s_2}. \tag{5.9}$$

In this model there are five parameters, namely a_1, a_2, s_1, s_2, and s_{12} (the parameter r is defined as $r = s_{12}/s_1 s_2$), but only three of them are independent, because $a_1 = a_2$ and $s_1 = s_2$. This is due to the fact that the occupancies n_1 and n_2 have the same nature (i.e. they are defined over two equivalent subvolumes), so that the two marginal averages are the same, and also the two marginal variances are the same. The distribution in (5.8) is a *quantized* Gaussian because variables n_1 and n_2 are integer numbers (moreover, they are defined over a finite range of non-negative values), this causing $\pi_{\mu}^{\chi}(\cdot,\cdot)$ to bear little to no resemblance with a (continuous) normal distribution. Therefore, in general, there is no correspondence neither between a_1, a_2 and the marginal means, nor between s_1^2, s_2^2 and the marginal variances, nor between s_{12} and the covariance. a_1, s_1, and s_{12} are rather free parameters that we direct-search optimize to produce $\pi_\mu^\chi(\cdot,\cdot)$ histograms that reasonably approximate the original distributions $p_\mu^\chi(\cdot,\cdot)$, in terms of marginal means, marginal variances, and covariance.

5.3 Results and discussion

We developed the present CG scheme considering two host-guest systems: united atom methane adsorbed (i) on a single layer of graphene, and (ii) between two graphene layers. In the latter system the interlayer spacing is 12 Å; results from previous computational studies on similar systems showed that a distance of 12 Å allows for an optimal methane uptake. [61] For both systems, we performed the same partitioning, consisting in a single layer tiling of tetragonal cells with $a = 17.1$ Å, and $c = 12$ Å (see Fig. 5.1). The cut-off of pair-wise interactions was also set at 12 Å. Being all the cells on the same layer, we can actually see this partitioning as a two-dimensional system of adjacent squares. Mapping such systems to the lattice model leads to a topology analogous to the King's graph, which can be imagined as the overlap of a square lattice with another square lattice rotated by a 45^o with respect to the first one and stretched by a factor $\sqrt{2}$ (see Fig. 5.2).

The host materials were represented as rigid structures, with each carbon atom modeled as a Lennard-Jones particle,[119] ($\sigma_{\mathrm{CH_4-C}} = 3.6135$ Å, $\epsilon_{\mathrm{CH_4-C}} = 0.607867$ kJ·mol^{-1}) and each methane molecule as a single Lennard-Jones bead ($\sigma_{\mathrm{CH_4-CH_4}} = 3.72$ Å, $\epsilon_{\mathrm{CH_4-CH_4}} = 1.317834$ kJ·mol^{-1}).[120]

As we mentioned in the previous section, every system consisted of 4 × 4 cells;

in order to exclude any finite-size effect on histogram evaluations, we also simulated a limited number of 6 × 6- and 5 × 5-sized versions of the same systems, without observing any meaningful difference in terms of occupancy probability distributions.

For both FG systems, classical GCMC simulations were performed in a variety of different conditions in order to separate the cell-to-cell interactions, according to the prescriptions we illustrated in the description of the model. We considered each system at three different temperatures (100, 200, and 300 K), and for each temperature we conducted a fine scan of chemical potential (or, equivalently, fugacity) values. More specifically, we investigated about 30 different chemical potential values for each system at each temperature. Now, according to the coarse-graining procedure we described in the previous section, the evaluation of CG interaction parameters requires three separated FG simulations for each chemical potential and temperature; from such three simulation we draw three different occupancy histograms, therefore the total number of histograms for each system at each temperature is about 90 (i.e. 60 bivariate *plus* 30 univariate histograms).

After calculating at each temperature the local free energy terms H_n and $K^{\chi}_{n_1,n_2}$, both with and without resorting to the pre-processing of histograms, we simulated the so obtained CG lattice models in the grand canonical ensemble with the Metropolis-Hastings scheme. Since the number of degrees of freedom of occupancy-based CG lattice models is dramatically reduced with respect to their atomistic counterparts, the computational effort for CG simulations is also significantly less than for FG simulations. Following Merrick *et al.*[121] we carried out a comparision between the efficiency of FG and CG simulations in terms of the time they required to reach equilibrium under the same conditions (i.e. same system, same size, same chemical potential, and same temperature), and found that the speedup for CG simulations is around 105—which is a mean value, i.e., on average, a CG

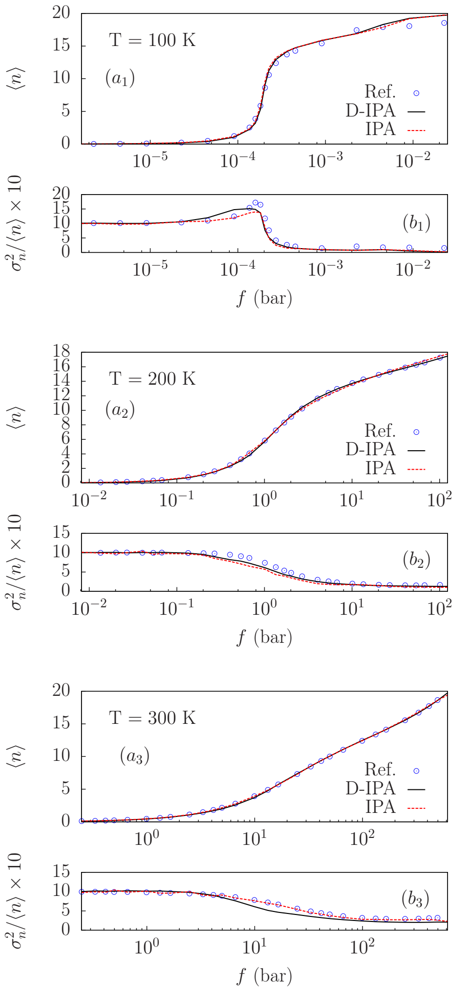

Figure 5.3: Isotherms and reduced variance for methane on single layer graphene at temperatures 100, 200 and 300 K. Isotherms are shown in subfigures labeled with letter a, reduced variances are shown in subfigures labeled with letter b. Blue empty circles represent the reference GCMC simulations with all classes active, solid black lines represent the CG simulations with data-preprocessing (D-IPA), dashed red lines represent the CG simulations without data-preprocessing.

simulation is faster than FG of a factor 105.

In this section we compare the static properties of the FG and the corresponding CG systems in terms of adsorption isotherms, occupancy fluctuations, and occupancy covariances. Adsorption isotherms are reported as the *loading* (i.e. average occupancy, $\langle n \rangle$) *vs.* fugacity f, whereas FG and CG fluctuations are compared in terms of the reduced variance, $\sigma^2_{n,Red} = \sigma^2_n / \langle n \rangle$, where σ^2_n is the occupancy variance for a single cell. [122, 123] Comparisons of spatial correlations (i.e., covariance) for each neighboring class are carried out in terms of Pearson correlation coefficients, which in the present case read $\rho^I = \sigma^I_{12}/\sigma^2_n$ and $\rho^{II} = \sigma^{II}_{12}/\sigma^2_n$ for class I and class II respectively, where σ^{χ}_{12} is the occupancy covariance of the pair occupancy distribution $p^{\chi}_{\mu}(\cdot, \cdot)$ for class χ, and σ^2_n is the marginal variance.

All results are reported in terms of fugacity, f, in units of bars, rather than chemical potential, μ, where $f = k^*_B T \exp(\mu/RT)$, with μ in units of kJ·mol^{-1}, $R = 8.3144626 \cdot 10^{-3}$ kJ·mol^{-1}·K^{-1}, and $k^*_B = 138.06488$ bar·Å^3·K^{-1}.

5.3.1 Methane on single layer graphene.

Results of numerical simulations are shown in Fig. 5.3, where “Ref” denotes results from GCMC simulations of the FG systems, while “IPA” means coarse-graining without histogram pre-processing, and “D-IPA” indicates coarse-graining *with* histogram pre-processing. From one GCMC simulation of the FG system to the next, the chemical potential (and, consequently, the fugacity) is changed by a small amount until the completion of a single layer of adsorbed methane molecules.

Increasing the temperature in the FG system yields a smoothing and straightening effect both on the isotherms and the occupancy fluctuations, this effect being due to the decrease of correlations between the host material and the guest molecules. Both the IPA and the D-IPA models perform with a comparable accuracy with respect to the FG results, which is always quantitative for the isotherms and semi-quantitative for the fluctuations. More specifically, the original isotherms are quantitatively matched at all three temperatures by both CG models, with the

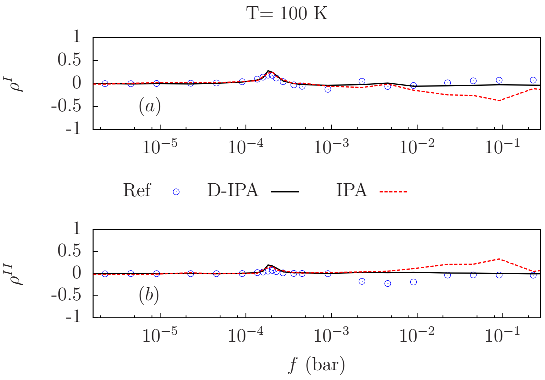

Figure 5.4: Class-wise Pearson correlation coefficients for methane on single layer graphene at 100 K. Results for class I are shown in the subfigure labeled with letter a, results for class II are shown in the subfigure labeled with letter b. Blue empty circles represent the reference GCMC simulations with all classes active, solid black lines represent the CG simulations with data-preprocessing (D-IPA), dashed red lines represent the CG simulations without data-preprocessing.

IPA case providing a nearly perfect match. The situation is the same for the re-

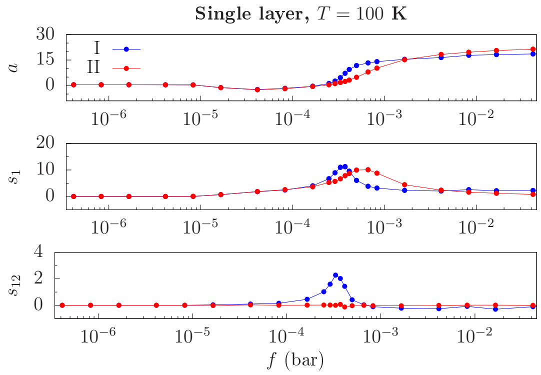

Figure 5.5: Optimized parameters (from top to bottom: a, s_1, and s_{12}) of the quantized gaussian distribution model [see Eqs. (5.8) and (5.9)] for methane on single-layer graphene at $T = 100$ K. Blue and red color refer to bivariate occupancy histograms of neighborhood class I and II, respectively.

duced variances and the covariances, except for the lowest temperature case (100

K) at high loadings, where an increase of correlations between neighboring cells is observed in the steep region of the isotherm ($f = 2.72 \cdot 10^{-4}$ bar), where we have the filling of one methane layer upon the graphene sheet (Fig. 5.4). Such increase in correlations causes GCMC histograms to assume a very "irregular" shape. Noise becomes then a relevant issue during the histogram evaluation, and the recursive nature of relations 5.4 and 5.7 for the calculation of the free-energy contributions leads to propagation of error in the estimation of occupancy histograms. Under such conditions, pre-processing the histograms proved then to be crucial, leading the CG model back to quantitative matching. In Fig. 5.5 we report the optimal values for parameters a, s_1, and s_{12} to be used in the quantized gaussian distribution model [Eqs. (5.8) and (5.9)] in order to fit the FG bivariate histograms throughout the whole set of fugacities we considered.

5.3.2 Methane between two graphene layers.

In this case, GCMC simulations of the FG system were conducted within a fugacity range which allows for the filling of a double layer of methane molecules in the interlayer space (that amounts to 12 Å). The FG and CG results (adsorption isotherms and reduced variances) for this system are shown in Fig. 5.6. The accuracy scenario of the CG representations is comparable to the one obtained for the previous system, with quantitative agreement attained in all but the lowest temperature/high loading case.

A major difference between this and the single-layer case lies in the steepness in the step in the adsorption isotherm, which for the double graphene layer case at $T = 100$ K is observed at fugacity $f = 4.15 \cdot 10^{-5}$ bar, and is definitely abrupt: a fugacity increase of about 10^{-6} bar causes the loading to sharply rise from 1.4 to 31 guest molecules per cell—correspondingly, the reduced variance shows a sharp

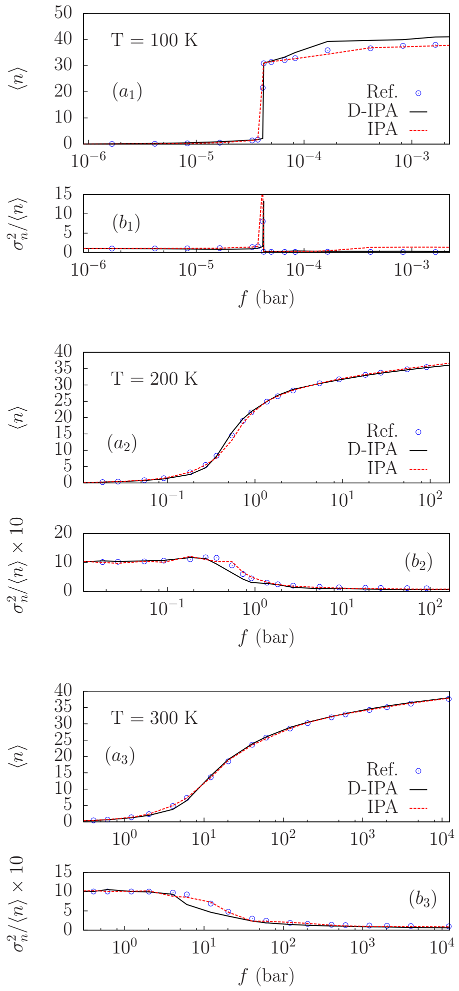

Figure 5.6: Isotherms and reduced variance for methane between two graphene layers at temperatures 100, 200 and 300 K. Isotherms are shown in subfigures labeled with letter a, reduced variances are shown in subfigures labeled with letter b. Blue empty circles represent the reference GCMC simulations with all classes active, solid black lines represent the CG simulations with data-preprocessing (D-IPA), dashed red lines represent the CG simulations without data-preprocessing.

peak (see Fig. 5.6). A detailed molecular-level analysis of this transition falls beyond the scope of this work and will be the subject of further investigations. However, in Fig. 5.7 we compare two different molecular configurations of the FG

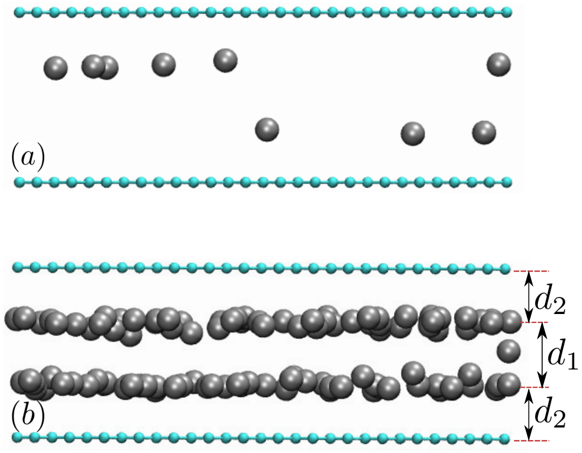

Figure 5.7: Molecular configurations of the same portion of the FG double layer system at $T = 100$ K. subfigures (a) and (b) represent two projections right before and after the adsorption step. In (b), we indicate with d_1 the methane interlayer distance, and with d_2 the distance between each methane layer and the nearest graphene sheet.

double layered system at $T = 100$ K, right before ($f = 3.70 \cdot 10^{-5}$ bar) and after ($f = 4.23 \cdot 10^{-5}$ bar) the adsorption step. Before the adsorption step, the system appears very diluted in methane and the guest molecules tend to fill uniformly two distinct layers in the graphene-graphene space. We did not observe any pattern in the position of methane molecules with respect to the carbon atoms in the graphene sheets; this is in agreement with computational studies on Coronene–CH_4 interaction energies, [124] which show nearly the same adsorption energy for all three kinds of methane-carbon adsorption sites. The position of such methane layers with respect to the graphene sheets becomes very evident at higher fugacity values, i.e. at higher methane concentrations. Under such conditions (i.e. after the adsorption step) we report a value of $d_1 \sim 4.08$ Å for the distance between these two methane layers, and of $d_2 \sim 3.96$ Å for the distance between each methane layer and the nearest graphene sheet. The observed distances d_1 and d_2 are very close to the optimal Lennard-Jones distances, $d_1^{\mathrm{opt}} = 2^{1/6}\sigma_{\mathrm{CH_4-CH_4}} = 4.18$ Å, and

$d_2^{\text{opt}} = 2^{1/6}\sigma_{\text{CH}_4\text{–C}} = 4.05$ Å. Such optimal distance values support the observed optimal methane uptake reported when the graphene-graphene spacing is about 12 Å, since the summation of the optimal distances is 12.28 Å. [61]

In this case, the pre-processing (D-IPA curves in Fig. 5.9) allowed for the production of a set of CG interaction parameters that significantly improved the agreement in terms of spatial correlations at high loadings. By looking at the D-IPA curves in Fig. 5.9 for $f > 4.15 \cdot 10^{-5}$ bar, we can see that such improvement comes along with an improvement in the single-cell reduced variance as well, but also with a slight accuracy loss in the adsorption isotherm—which could be made even slighter, but at the considerable cost of increasing the complexity of the coarse-graining model, e.g. by including a further CG equation [besides Eqs. (5.3) and (5.5)] describing three-term interactions. Therefore, we believe that the accuracy in the adsorption isotherm can be still considered very satisfactory, despite the class-independence assumption we made in order to keep the CG model definition as simple as possible.

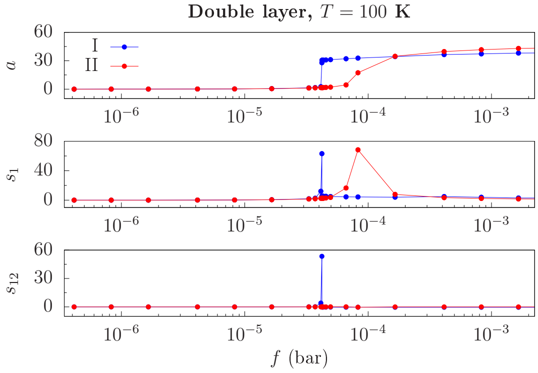

Figure 5.8: Optimized parameters (from top to bottom: a, s_1, and s_{12}) of the quantized gaussian distribution model [see Eqs. (5.8) and (5.9)] for methane on double-layer graphene at $T = 100$ K. Blue and red color refer to bivariate occupancy histograms of neighborhood class I and II, respectively.

The histogram pre-processing [optimal values for parameters a, s_1, and s_{12} to be used in the quantized gaussian distribution model, Eqs. (5.8) and (5.9), are shown in Fig. 5.5] improved the correlations in situations in which the original pair-occupancy histograms obtained through GCMC were certainly affected by non-negligible accuracy issues. In fact, at low temperature and high density (but not close to the adsorption step) the occupancy fluctuations are low; correspondingly, the occupancy distributions turn out to be sharply peaked. Now, the cell occupancy varies within a relatively small range, which goes up to about 20 and 40 molecules per cell, respectively for the case of methane in a single graphene layer and within a double graphene layer. As a consequence, occupancy histograms being sharply peaked imply good sampling of only a limited number of occupancy pairs, namely, those that are very close to the average value. Any other occupancy pair is sampled poorly. Eq. (5.5), i.e. the one that contains information about class-wise occupancy correlations, is the CG equation that is most seriously affected by such accuracy, and the diverging correlations shown in Figs. 5.4 and 5.9 are the end result. In the vicinity of the adsorption step the situation is even more complicated: the variances are *very high*, but this does not necessarily imply that the corresponding distributions are short and wide—more generally, the occupancy distributions under such conditions are no longer unimodal, and can not be considered stable (i.e., very small changes in fugacity would cause large changes in the shape of distributions). When facing such problems, the first solution that comes to mind would be to carry out much longer simulations, in order to have significantly more data to take into account while estimating the occupancy histograms. However, we wanted to find out how much the CG model could be improved with just the input data we had, without adding more data to the source set of histograms; this is the reason why we preferred to manipulate that set by means of a "histogram imitation technique", rather than to perform longer GCMC runs. Of course replacing the original distributions with ”fake, but better-behaving ones” means to coarse-grain a system that differs from the original one in some aspects. Nevertheless, if the CG model we want to build from some FG reference system aims to correctly imitate its occupancy correlations in space, such an operation appears legitimate.

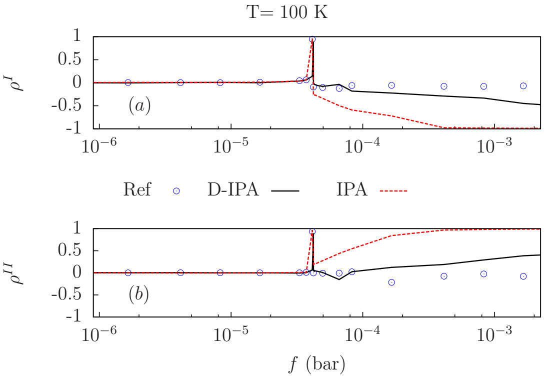

Figure 5.9: Class-wise Pearson correlation coefficients for methane and double layer graphene at 100 K. Results for class I are shown in the subfigure labeled with letter a, results for class II are shown in the subfigure labeled with letter b. Blue empty circles represent the reference GCMC simulations with all classes active, solid black lines represent the CG simulations with data-preprocessing (D-IPA), dashed red lines represent the CG simulations without data-preprocessing.

5.4 Conclusions

We performed the spatial coarse-graining of the static occupancy-related properties of two adsorption systems, namely one and two graphene sheets with methane as the adsorbate, at various temperatures. In order to accomplish this task, we extended the interacting-pair approximation (IPA) method[38], a local occupancy-based spatial partitioning approach to the coarse-graining of host-guest systems, to the case in which every subvolume of the partition is surrounded by neighboring subvolumes of two kinds. The resulting two different kinds of spatial correlations were reproduced by local, class-wise mutual interaction parameters, defined on the basis of pair-occupancy histograms evaluated from properly tailored fine-grained GCMC simulations within a wide range of fugacity—namely, from zero-loading to the complete filling of the graphene sheet(s) with adsorbate molecules. The coarse-grained (CG) potentials we obtain are functions of the local occupancies and are temperature-dependent, but do *not* depend on any other global variable (such as, e.g., overall density, or fugacity, or chemical potential); this enables us to use the same set of CG potentials at any fugacity value within the range of interest. We evaluated the quality of coarse-graining in terms of agreement between the properties of the local occupancy distributions of the coarse-grained (CG) systems, and the properties of the same distributions for the corresponding reference, fine-grained (FG) systems. The results showed a very satisfactory agreement in almost all the scenarios we investigated. Only at low temperature (100 K) and high density both systems required a pre-processing of the pair-occupancy histograms over which the CG potentials are defined, in order to allow for the production of realistic CG correlations despite the relatively poor accuracy with which they were sampled, without resorting to longer sampling runs. This pre-processing prescribed the replacement of the original GCMC histograms with quantized Gaussian distributions with similar means, variances and covariance; the improvement we obtained from it was especially relevant for the double-layer case at 100 K, where the adsorption isotherm shows an abrupt and steep loading change at intermediate loadings—a scenario where accuracy issues in the source GCMC histograms may prevent the CG parameters from producing correct occupancy correlations at high loading.

Chapter 6

Scaling-up Simulations of Diffusion in Microporous Materials

Reprinted (adapted) with permission from (Giovanni Pireddu, Federico Giovanni Pazzona, Pierfranco Demontis, and Magdalena A. Załuska-Kotur. Scaling-up simulations of diffusion in microporous materials. *Journal of Chemical Theory and Computation*, (just accepted), 2019) Copyright (2019) American Chemical Society

6.1 Introduction

Computer simulations of physical systems have widely demonstrated their usefulness in understanding complex phenomena, by both offering a direct comparison with purely theoretical approaches, and being capable of providing insightful predictions.[2, 5, 49]

Over the last three decades, multiscale modelling approaches have progressively gained interest in several disciplines and for different applications.[125, 126] In particular, bottom-up protocols allow for representing the systems of interest in increasingly large time- and length-scales, by progressively decreasing the level of detail associated with each representation through coarse-graining methods.[13, 14, 19, 21, 25, 72, 75]

Coarse-graining methods can be categorized into "topological" and "spatial". In the *topological* case, atoms are grouped into coarser beads, equipped with an effective force field, and let evolve in time through molecular dynamics (MD)-based algorithms; therefore, in this case, observables like positions, orientations and momenta are defined both in the finer and the coarser representations.

In the *spatial* case, the simulation space is partitioned into geometrically equivalent domains, thus defining a lattice representation of the system; fine-grained (FG) observables are mapped into the *internal state* of every lattice node; while the FG force fields is based on atomic positions and molecular topologies, in the coarse-grained (CG) picture effective interactions are based on internal states. Whereas FG observables evolve through MD-based algorithms, the lattice internal states of the CG representation evolve in time through specific (i.e., not general) schemes such as kinetic Monte Carlo and Cellular Automata algorithms.[34, 127–130]

Although well established protocols are currently available in the literature for topological coarse-graining,[15, 24, 131] the same cannot be said for the spatial approach, especially for what concerns the CG dynamics.

In this work, we focus on the mesoscopic representation of host-guest systems constituted by microporous materials and gas molecules. Nowadays, microporous materials are broadly employed in different scenarios and for different scopes, such as gas storage, separation of mixtures, heterogeneous catalysis, etc.[53, 132] Many of the processes involved in such applications strongly depend on the adsorption and diffusion behaviour of the guest molecules in the porous environment.[133] Thus, a general and sufficiently accurate mesoscale modelling framework for such phenomena could help to explain diffusive and sorptive properties and allow testing new systems *in-silico*, such as hypothetical sorbent materials for various applications.[134]

Lattice models of host-guest systems have demonstrated the capability of representing adsorption and diffusion phenomena with a considerably smaller computational effort compared to atomistic methods and yet allowing to reproduce the properties of interest with satisfactory accuracy.[78, 110, 135–137]

In our case, we map the reference systems into pore-scale lattice models, in which each node represents a pore or a cage of the host material and is equipped with an occupancy state n indicating the number of guest molecules present in such pore of the reference material.

Thermodynamics and mass-transfer dynamics in the CG representations are both modelled to match with the results of FG atomistic simulations — grand-canonical Monte Carlo (GCMC) to model the static properties, and MD simulations to model the transition-rates associated with the inter-cage jumps performed by the guest molecules. More specifically, our CG representation is defined on the basis of statistical data obtained from GCMC and MD simulations. In particular, in order to test the ability of our coarse-graining strategy to produce a reliable mesoscale counterpart from a minimal FG dataset, GCMC and MD simulations were intentionally run over *small* portions of the FG reference systems, and over a relatively *short* time window.

To demonstrate the capabilities of our method, we chose to represent two interesting systems constituted by two different host materials and involving methane molecules as the guest species. The first material is the all-silica ITQ-29 zeolite, which is a well-studied material for the modelling of cage-to-cage dynamics and diffusion of small molecules in microporous materials.[82, 120] The second material is the LTA-zeolite-templated carbon (which we will refer to as ZTC), recently introduced as hypothetically obtainable by carbon-templating the afore-mentioned zeolite.[57]

Zeolite-templated carbons are a relatively new class of nanoporous carbon materials, which exhibit peculiar properties when employed as methane sorbents.[55, 56] Despite being related to its zeolite precursor and having the same topology in terms of pore connectivity, the ZTC we consider is structurally different as it presents larger free-volume in each pore and significantly larger openings between neighboring cages. For this reason, it is particularly interesting to compare the static and dynamic properties of the two systems.

The remainder of this work is organized as follows: in the section Methods, we introduce our CG method and explain our experimental setup for the numerical simulations; in the section Results and discussion, we show the numerical results in terms of comparisons between the two systems in terms of both the static and dynamical properties; finally, in the last section we draw our conclusions, by highlighting the benefits and the limits of our method, and by proposing possible applications.

6.2 Methods

Our method aims to provide a lower resolution description of host/guest systems directly derived from atomistic data, and consists of three main steps.

First of all, we carry out a number of atomistic simulations (GCMC and MD) of the systems of interest, from which we draw the statistics required by the method itself, and we map the FG simulation space into the CG occupancy-based lattice model. Since we treat the host material as a rigid framework, such mapping is static—lattice cells do not change in shape, size, or position (Section 6.2.1).

Secondly, we use static properties drawn from atomistic GCMC to define occupancy-dependent effective interactions inside every cell and among neighboring cells of the lattice (Section 6.2.2). The methodology is described in detail in our previous works.[38, 39]

Finally, we use mass transfer statistics obtained from MD to model the dynamical evolution of the CG lattice model as a Markov process based on a local operator (Section 6.2.3), which we further refine in order to take into account dynamical correlations and non-Markovian effects that are observed in MD simulations (Section 6.2.4).

Before proceeding, it is useful to note that the terminology used to denote some relevant physical quantities investigated in this work (namely, the diffusivities and the reduced variance) is in agreement with a limited part of the scientific literature of the field[138]. Other authors name the same physical quantities differently or use similar names to denote different properties. For example, what here we indicate as the *center of mass diffusivity* is elsewhere named as corrected diffusivity or collective diffusivity by other authors[120, 139]. Furthermore, what we mean by *collective diffusivity* is denoted as transport diffusivity in other works[139].

6.2.1 Coarse-grained structure

Our coarse-graining procedure begins with the structural definition of the lattice models, which represent the reference host-guest systems. Fig. 6.1 depicts a representation of the structural mapping of the molecular systems, from the atomistic picture to the occupancy-based lattice model.

We ideally tessellate the host materials with identical, non-overlapping cubic subvolumes called *cells*. In our picture of the systems, each cell embeds a single pore of the reference host material. Since the ZTC and the ITQ-29 both present a simple cubic pore connectivity, the reference structures are conveniently mapped

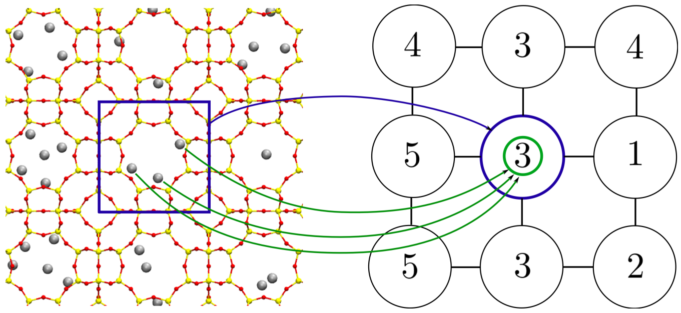

Figure 6.1: Mapping of the methane/ITQ-29 system into its corresponding occupancy-based lattice model. Methane molecules are represented as grey spheres, while framework atoms are represented by red (O species) and yellow (Si species) spheres. The left subfigure shows a 2-d projection of an exemplifying fine-grained system made of 3 × 3 × 3 zeolite cavities, which can be straighforwardly partitioned into a 3 × 3 × 3 lattice of cells (right subfigure; periodic boundary conditions are implied). Each cell of the lattice refers to a certain portion (i.e., a cavity) of the fine-grained system and bears an internal state (occupancy) equal to the number of guest molecules inside of that cavity: in the lattice representation (right subfigure), the integer numbers represent cell occupancies. In blue/green color we show, as an example, the mapping of one of such cavities into its lattice counterpart; the cavity chosen for this example is the one in the middle of the 2-d projection (the mapping of the other cavities from fine-grained to lattice representation follows the same reasoning). The blue color represents the space covered by the cavity, whereas the green color indicates the mapping of the guest position inside of it into a single-valued internal state, i.e. the cell occupancy. The links in the lattice model (black straight lines in right subfigure) represent the connections between neighboring cavities of the host material.

to cubic networks. Each i-th cell of the CG lattice is then associated with its occupancy n_i, which corresponds to the total number of molecules whose center of mass falls within the i-th pore. In this fashion, the configuration of our lattice models is completely defined as the occupancy configuration $\mathbf{n} = \{n_1, \ldots, n_M\}$, where M is the total number of cells.

6.2.2 Thermodynamics

Following the Interacting-Pair-Approximation (IPA) approach,[38, 39] we associate the occupancy configurations of the lattice models with a CG potential function Ω, which in the grand-canonical ensemble reads

$$\Omega_\mu(\mathbf{n}) = \sum_i (H_{n_i} - \mu n_i) + \sum_{\langle ij \rangle} K_{n_i,n_j}, \tag{6.1}$$

where μ is the chemical potential, H_{n_i} is the single-cell free-energy contribution of a cell with occupancy n_i, K_{n_i,n_j} represents the free-energy contribution of the mutual interactions between neighboring cells with occupancies n_i and n_j, and $\langle ij \rangle$ indicates a summation over nearest-neighboring cells.

More specifically, the self-interaction term H_{n_i} represents the contribution to the free energy of the system provided by (i) the interactions among the n_i guest molecules located in the i-th cavity, and by (ii) the interaction between the same n_i molecules and the *whole* framework (which, we remark, is kept rigid in our simulations). The pair-interaction term K_{n_i,n_j} represents the contribution to the free energy of the system coming from the interactions between the n_i guest molecules located inside cavity i and the n_j molecules located inside cavity j. For our purposes, we will assume that such parameterization will suffice to adequately represent the effective interactions at a CG level of representation.

As shown in Eq. (6.1), except from μ, the CG potential function makes only use of local, occupancy dependent free-energy contributions. The free-energy parameters are related to the respective partition function contributions via the following relations:

$$Q_n = e^{-\beta H_{n_i}}, \qquad Z_{n_i,n_j} = e^{-\beta K_{n_i,n_j}}. \tag{6.2}$$

Such contributions can be conveniently computed using the following recurrence relations

$$\frac{Q_n}{Q_{n'}} = \frac{e^{-\beta\mu n}\, p^o_\mu(n)}{e^{-\beta\mu n'}\, p^o_\mu(n')}, \tag{6.3}$$

$$\frac{Z_{n_1,n_2}}{Z_{n'_1,n'_2}} = \left(\frac{e^{\beta\mu n'_1} Q_{n'_1}\, e^{\beta\mu n'_2} Q_{n'_2}}{e^{\beta\mu n_1} Q_{n_1}\, e^{\beta\mu n_2} Q_{n_2}}\right)^{\frac{1}{\nu}}$$

$$\times \left(\frac{p_\mu(n'_1)\, p_\mu(n'_2)}{p_\mu(n_1)\, p_\mu(n_2)}\right)^{1-\frac{1}{\nu}} \frac{p_\mu(n_1,n_2)}{p_\mu(n'_1,n'_2)}, \tag{6.4}$$

where $p^o_\mu(n)$, $p_\mu(n)$ and $p_\mu(n_1, n_2)$ are respectively the univariate occupancy distribution for a single-cell closed system, the univariate occupancy distribution of a single cell inside the reference system, and the bivariate occupancy distribution of two connected cells inside the reference system. The symbol ν indicates the cell connectivity, which is 6 for a 3D cubic network.

We obtained the local occupancy distributions from atomistic grand-canonical Monte Carlo (GCMC) simulations of the reference system. Finally, the full set of free-energy contributions can be obtained by solving the recurrence relations in Eqs. (6.3) and (6.4), and using $H_0 = 0$ kJ/mol and $K_{0,n} = 0$ kJ/mol (for every possible value of occupancy n) as starting points, since empty cells do not contribute to the total free-energy of the system.

6.2.3 Elementary events for diffusion

We assume that gas diffusion in microporous materials can be treated as the composition of several elementary and strictly local events. An appropriate observation time scale τ could allows to distinguish among single migration events occurring during the dynamical evolution of the host-guest systems. The aim of our work is to provide a stochastic modelling protocol to understand and represent such events, which we identify as single molecule jumps between two connected pores. If the separability between elementary events holds, we assume that local dynamics can be represented by a local operator $W(\mathbf{m}' \mid \mathbf{m})$, which is applied to a single pair of connected cells and represents the transition probability of the transformation $(\mathbf{m}' \mapsto \mathbf{m})$ within the time interval τ, where $\mathbf{m} = (n_1, n_2)$ and $\mathbf{m}' = (n_1', n_2')$ are the pair occupancy configurations before and after the transition. For example, $W((4,6) \mid (5,5))$ represents the probability of a single molecule jump (from a cell of occupancy 5 to a neighboring cell of occupancy 5) resulting in the transformation $((5,5) \mapsto (4,6))$. By following this approach, during each elementary event the local total mass $M_{12} = n_1 + n_2$ is conserved.

We empirically determine the transition rate values W from atomistic Molecular Dynamics (MD) simulations of the reference systems, during which we saved the positional configuration of the diffusing molecules in the system (all coordinates of methane molecules) every τ seconds, thus resulting in one trajectory of positional configurations for every MD simulation.

Since the host framework is rigid, the cavity centers are known and fixed throughout the whole FG trajectories. Thus, for every instance of each MD trajectory, by following a nearest-cavity-center criterion we assign every methane molecule to the cavity it currently belongs to, so that the occupancy of each cell is simply defined as the total amount of guest molecules belonging to the corresponding cavity.

All trajectories of positional configurations obtained for the reference system are used to compute the time series of occupancy configurations; finally, the occupancies of each pair of connected pores between two consecutive occupancy configurations, say $(n_i(t), n_j(t))$ and $(n_i(t+\tau), n_j(t+\tau))$, are compared, and if the transformation from one occupancy pair to the other conserves mass [that is, if $n_i(t) + n_j(t) = n_i(t+\tau) + n_j(t+\tau)$], they are cumulated into the respective entries of W.

With this procedure, we obtain empirical values for each $W(\mathbf{m}' \mid \mathbf{m})$, where we ignore more complicated, multi-cell mass transfer mechanisms which may occur within the chosen time step, but still are much rarer than single jump events.

Since the migration of molecules from a pore to another is a thermally activated process, a common way of modelling jump rates is by introducing a temperature-dependent function of the free-energy barrier multiplied by a kinetic prefactor. The first part is a static property, which accounts for the local free-energy change associated to an inter-cell jump event, whereas the kinetic prefactor $k_{M_{12}}$ models the jump attempts frequency, and is a function of local occupancies. We model the prefactor as a function of the local occupancy summation M_{12}, which is conserved during each elementary event.

In general, k implicitly contains the Boltzmann factor related to the free-energy barrier related to each transition. Thus, such function should also explicitly contain the cell occupancies and the temperature. However, modelling k only on the basis of the local summation of occupancies is particularly convenient for two reasons: (i) a minimal amount of data is sufficient to fit k on the basis of the transitions observed during MD trajectories; (ii) it facilitates fulfilling the detailed balance condition (DB), since M_{12} is preserved during every transition. Furthermore, since the systems under study were considered at constant temperature, the temperature dependence of $k_{M_{12}}$ is kept implicit.

The functional form we propose for the jump rates is the following:

$$W(\mathbf{m}' \mid \mathbf{m}) = k_{M_{12}} e^{-\frac{\beta}{2}[\Omega_\mu(\mathbf{m}')-\Omega_\mu(\mathbf{m})]}, \tag{6.5}$$

with $\beta = 1/k_B T$, where k_B is the Boltzmann constant and T is the temperature. The factor 1/2 in the exponent on the right hand side of Eq. 6.5 stems from the DB condition imposed to a *closed* pair of cells transforming from occupancy pair $\mathbf{m}$ to occupancy pair $\mathbf{m}'$, i.e. $p_\mu(\mathbf{m})\, W(\mathbf{m}'|\mathbf{m}) = p_\mu(\mathbf{m}')\, W(\mathbf{m}|\mathbf{m}')$, where $p_\mu(\mathbf{m}) \propto \exp\{-\beta\Omega_\mu(\mathbf{m})\}$ and $k_{M_{12}}$ is symmetric with respect to the jump direction. By following the definition of CG potential function given in Eq. 6.1, the potential function for a pair of connected cells, say cell 1 and cell 2, respectively occupied by n_1 and n_2 guest molecules, reads $\Omega_\mu(n_1, n_2) = -\mu(n_1+n_2)+H_{n_1}+H_{n_2}+K_{n_1,n_2}$. By taking into account the fact that all the local transitions we consider do conserve mass, and by omitting the interaction contributions with the environment around the chosen pair of connected cells, the local change in free-energy due to the transition $\mathbf{m} = (n_1, n_2) \mapsto \mathbf{m}' = (n_1', n_2')$ is

$$\begin{aligned}\Omega_\mu(\mathbf{m}') - \Omega_\mu(\mathbf{m}) = {} & H_{n_1'} + H_{n_2'} + K_{n_1',n_2'} \\ & - \left(H_{n_1} + H_{n_2} + K_{n_1,n_2}\right).\end{aligned} \tag{6.6}$$

Although the expression we proposed for the transition rates, $W(\mathbf{m}' \mid \mathbf{m})$ (see Eq. 6.5), stems from the DB condition imposed on a closed pair of neighboring cells, the choice to not include the interactions with the neighbors around each pair hinders our operator from strictly fulfilling the DB condition on the whole system; however, this is consistent with our sampling scheme from the MD simulations, since we sample the transitions on the basis of each pair configuration only. Of course, if we wanted to ensure that DB is strictly obeyed, the jump rates W should also include information about occupancies in all the cells in the neighborhood around each pair; in other words, all such occupancies should appear as additional arguments in the conditionality of W. However, this would cost us a much heavier computational effort, that is required in order sufficiently robust statistics—this is against the spirit of our work, since, as we mentioned in the Introduction, our aim is to obtain reliable CG representations from relatively *short* and *small-scale* atomistic simulations. We also remark that modelling the full system by sampling a W based merely on local pair occupancy configurations is equivalent to implicitly assume a mean-density around each pair, since the cells are embedded in the full system. This method for the local dynamical evolution is analogue to a pair-wise stochastic evolution rule in a block cellular automaton, where we identify each block as a closed pair of connected cells.[130] We empirically found that our approximate model still yields a semi-quantitative matching of static properties in terms of occupancy histograms between the CG and MD simulations.

6.2.4 Dynamical correlations

If correlations between any two consecutive pore-to-pore jumps in the reference FG systems were negligible, then the reference systems could already be simulated by directly using the W operators for the dynamical evolution of the lattice models with a Markov chain scheme. However, in real systems dynamical time-correlations, also called memory effects, may occur and significantly influence the diffusivity [138]. In principle, a higher-order (or higher-memory) model of the dynamics could be devised in such a way as to explicitly embed memory effects, and thus yield a realistic representation of the diffusion behaviour; however, also in this case, the amount of data that would be necessary for us to base such more sophisticated kinetic model upon a reliable statistics would be enormous. Therefore, in this work we preferred to embed the higher-order effects in the transition rates W under the form of an overall scaling factor f.

Such approach is based on the analysis of the dynamics of guest molecules only (the host material is kept rigid, therefore it does not undergo a dynamical evolution). In this analysis, a very important role is played by the center of mass (c.m.) of the guest species; from now on, reference to the guest species will be implied every time we use the abbreviation “c.m.”.

We start from

the memory-expansion expression of the c.m. diffusivity $D_{c.m.}$[140]

$$D_{c.m.} = \frac{1}{2dN\tau}\left(C_0^{\delta\mathbf{R}} + 2\sum_{t=1}^{\infty} C_t^{\delta\mathbf{R}}\right), \tag{6.7}$$

where d is the dimensionality, τ is the chosen time interval, N is the total number of molecules. $C_t^{\delta\mathbf{R}}$ is the c.m. displacement autocorrelation function which reads

$$C_t^{\delta\mathbf{R}} = \langle \delta\mathbf{R}_0 \cdot \delta\mathbf{R}_t \rangle, \tag{6.8}$$

where $\delta\mathbf{R}_t = \sum_i^N (\mathbf{r}_t - \mathbf{r}_{t-1})$ is the summation of all molecular displacements between time $t-1$ and time t.

In Eq. (6.7), $C_0^{\delta\mathbf{R}}$ is the autocorrelation function $C_t^{\delta\mathbf{R}}$ calculated at $t = 0$, and corresponds to the mean-squared displacement of the c.m. after one time interval τ.

Considering a purely Markovian approximation and neglecting all the correlation effects for $t > 0$ in Eq. (6.7) yields the dynamical mean-field (DMF) expression of c.m. diffusivity[141], $D^o_{c.m.}$:

$$D^o_{c.m.} = \frac{C_0^{\delta\mathbf{R}}}{2dN\tau} = \frac{\overline{W}a^2}{2dN\tau}, \tag{6.9}$$

where $\overline{W}$ is the average jump probability and a is the lattice cell parameter. The ratio between the infinite-memory diffusivity and the DMF diffusivity, $D_{c.m.}/D^o_{c.m.}$, can be taken as a measure of how memory effects influence the overall diffusion process. If such a ratio is below 1, then the overall effect is a slowing down of diffusion induced by negative correlations in displacements; if overall correlations in displacements are positive, instead, then the ratio $D_{c.m.}/D^o_{c.m.}$ turns out to be larger than 1, this resulting in an increase of diffusivity. By computing the correction factor as $f = \left(C_0^{\delta\mathbf{R}} + 2\sum_{t=1}^{\infty} C_t^{\delta\mathbf{R}}\right)/C_0^{\delta\mathbf{R}}$ and by using Eqs. (6.7) and (6.9), we obtain

$$D_{c.m.} = fD^o_{c.m.} = \frac{f\overline{W}a^2}{2dN\tau}. \tag{6.10}$$

Our idea is then to correct the purely Markovian jump rates according to $\overline{W}^{corr} = f\overline{W}$, and then to use such corrected jump rates $\overline{W}^{corr}$ in the numerical CG simulations, rather than $\overline{W}$. In general, we expect f to be a function of the global density $\langle n \rangle$ and this would cause the evolution operator to depend on a global variable;

however, since we want f to be local as well, we can circumvent this problem by replacing the dependence on the global density $\langle n \rangle$ with a *local density guess*, i.e. a guess of $\langle n \rangle$ on the basis of local occupancies. More specifically, our choice is to use the average local pair occupancies $\overline{M}_{12} = (n_1 + n_2)/2$ rather than $\langle n \rangle$ as input for the function f. In this way, we easily correct our local operator by embedding the overall effect of time-correlations and yet we keep locality and our approximate DB condition, since the average local pair occupancy $\overline{M}_{12}$ is a conserved quantity during each elementary transition. Despite its simplicity, this method allows one to estimate the overall correlation effect directly from the analysis of the original MD trajectories, without having to perform further simulations of the reference system [142].

6.2.5 Numerical simulations

We performed the FG atomistic simulations by modelling all the atoms involved as Lennard-Jones (LJ) particles. The whole methane molecule was represented by a single LJ bead, following the widely accepted united-atom approximation [120]. The methane-carbon and methane-zeolite LJ interactions were parameterized according to our previous works [38, 39].

In all the simulations, the host materials were represented as rigid frameworks. The ZTC crystalline structure, in its *unrelaxed* version, was downloaded from `materialscloud.org`[58], while the ITQ-29 structure was taken from RASPA2's repository on `github.com`[143]. A comparison of the pores of the two host materials is presented in Fig. 6.2.

The systems we considered were simulated at the same temperature, i.e. 300 K. The reason for this choice is two-fold: we wanted to represent a realistic scenario for room-temperature applications of such systems, and at the same time this temperature was observed to yield a sufficient number of molecular inter-cage

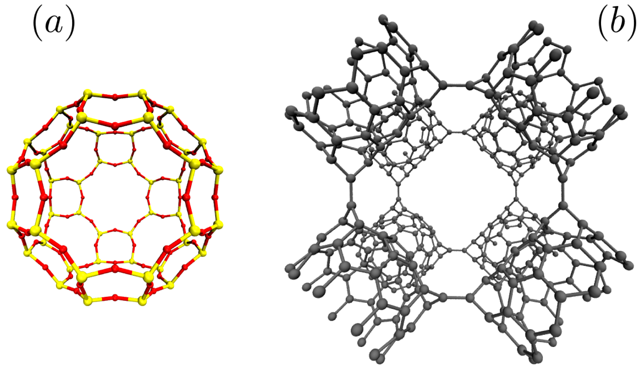

Figure 6.2: Atomistic representations of the cages corresponding to the ITQ-29 (subfigure (*a*)) and the ZTC (subfigure (*b*)) materials.

jumps in our simulations.

GCMC atomistic simulations were required for the calculation of the IPA parameters, and were performed using an in-house built code with the usual displacement, insertion and deletion trial moves [5], whereas all MD simulations were performed by using the open-source software LAMMPS.[144] We computed the MD trajectories for several methane loading values ($\langle n \rangle$ = 1,2,... up to 14 for the ITQ-29, and 15 for the ZTC) in terms of average number of guest molecules per pore, and considering a $3 \times 3 \times 3$ cells of the host materials, where every cell contains a single pore. Our choice for such maximal loading values is motivated by that fact that, in the zeolite case, we did not observe any inter-cage jump $\langle n \rangle > 14$, whereas in the ZTC case, loading values above $\langle n \rangle = 15$ resulted in the emergence of new inter-cage adsorption sites, which would require a much more complicated CG mapping. Also, since we wanted to highlight the comparison between the two materials, we chose a similar range of conditions for the two systems. In both cases, we obtained the methane trajectories by assuming periodic boundary conditions (PBCs) within the NVT ensemble; temperature was kept approximatively constant at 300 K through a Nosé-Hoover thermostat; every MD simulation started with a 0.5 ns-long equilibration stage; after equilibration, we sampled the dynamics of the methane-zeolite system for 10 ns (while saving molecular configurations every 1 ps), and the dynamics of the methane-ZTC system for 1 ns (while saving configurations every 20 fs).

In order to prove the accuracy of our method and to study the collective diffusivity in such two systems, we also performed numerical simulations of the CG models. Such simulations were conducted by applying the previously parameterized local operators to the lattice models of the reference systems and sequentially updating the states of connected pairs of cells. The evolution algorithm of our lattice models is designed as follows. Each simulation starts with initialization of the starting lattice occupancy configuration $\mathbf{n}$, then for each time-sweep the following scheme is used:

(1) we randomly extract a pair of connected cells out of all the connected pairs in the CG system (the same pair may be invoked more than once during the same time-sweep);

(2) we generate all possible outcomes $\mathbf{m}'$ and calculate the rate $W^{corr}(\mathbf{m}' \mid \mathbf{m})$;

(3) we randomly pick a new state $\mathbf{m}'$ according to the probability distribution $W^{corr}(\cdot \mid \mathbf{m})$, and then update the local occupancies;

(4) if the number of pairs invoked during the current time-sweep turns out to be equal to to the total number of connected pairs, then the current time-sweep is concluded; otherwise, we return to step (1).

Thus, the dynamical evolution of the CG system is discrete, and the time interval between one configuration and the configuration produced through one time-sweep is homogeneous and assumed equal to the time step, τ, according to which the conditional distribution W was computed from molecular-scale simulations of the reference system.

We remark that, according to the above scheme, every cell pair might happen to be invoked zero, or one, or multiple times within the same iteration. Adjacent or overlapping pairs can also be invoked at point (1): this implicitly allows the CG model for multi-molecule and multi-cell mass transfer mechanisms, still under a sequential Markovian approach.

We determined the behaviour of collective diffusivity as a function of the loading empirically, by using the Boltzmann-Matano (BM) method. This method was first introduced by Matano to study the interdiffusion of different metallic species in the proximity of the intermetallic interface [145], but it was also successfully applied to the study of collective diffusion of particles in lattice models [146, 147]. The BM analysis is conducted on the time-dependent profile of adsorbate density along a chosen direction, obtained from the spread of a step-like initial profile. The spread is numerically simulated according to the lattice CG dynamics. The relation between density profile and collective diffusion coefficient is the following:

$$D_c(\langle n\rangle) = \frac{1}{2t}\left(\frac{\partial\rho}{\partial x}\right)^{-1}_{\rho=\langle n\rangle}\int_0^{\langle n\rangle}(x-x_M)\,d\rho, \tag{6.11}$$

where ρ is the density profile (which has to be intended as the *local* average occupancy of the profile, thus ranging from 0 to the maximum occupancy n_{max}), t is the time considered for the spread of the initial profile, x is the chosen direction for the analysis and x_M is the position of the Matano plane, which is chosen to fulfil the condition $\int_0^{n_{max}}(x-x_M)d\rho = 0$. The choice of distinguishing ρ and $\langle n\rangle$ in Eq.6.11 is intentional, since $\langle n\rangle$ represents the loading, which is a global variable; while ρ represents the local value of the density profile. Hence, we imply the assumption that the profile ρ is sufficiently smooth and well-behaving such that it can be used to estimate the loading dependence of collective diffusivity.

The simulations used for the Boltzmann-Matano analysis were conducted with 200 × 5 × 5 supercells of the reference materials. We found this supercell configuration to be the optimal compromise in terms of computational effort and smoothness of density profiles. We also simulated 3 × 3 × 3 supercells of the reference materials in order to compare CG and FG relaxation behaviour in terms of occupancy correlations, and their respective static properties in terms of local occupancy histograms.

6.3 Results and discussion

6.3.1 Jump rates modelling

We started our CG procedure by calculating the local free-energy contributions in terms of single-cell H_n and mutual interaction $K_{n1,n2}$ terms for the two systems, within the IPA theoretical framework. The results of our free-energy parametrization for the two systems are shown in Fig. 6.3.

Our results show that the two systems exhibit a qualitatively similar behaviour in terms of CG thermodynamics. The H_n parameters monotonically decrease for the two systems with a progressive trend flattening at high densities. The mutual interaction parameters show an attractive regime at moderate densities i.e. $n_1 \times n_2 \leq 150$ for the ITQ-29, and $n_1 \times n_2 \leq 250$ for the ZTC system. Conversely, for higher values of loading, both systems exhibit a positive and relatively fast-growing mutual interaction contribution. Such effect reflects an overall repulsion between methane-rich cavities of the host materials. Despite qualitative similarities between the two systems, for the ZTC case we observe a deeper H_n contribution, indicating that, for a given value of density, the number of favorable configurations in the methane-ZTC system is larger than the ITQ-29 case. This is a direct consequence of the larger free volume present in the ZTC material, and of the weaker localization of the guest molecules. In fact, the presence of preferential methane adsorption sites in the ITQ-29 is well known [148], while our results yielded a more uniform distribution of methane positions within the ZTC. Also, we

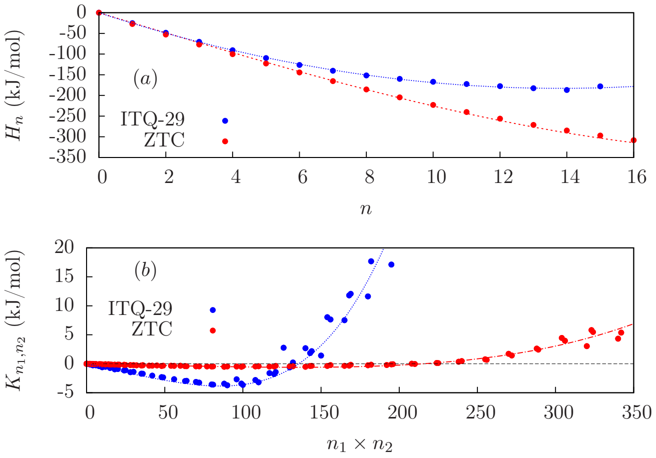

Figure 6.3: Free energy parameters in units of kJ/mol as obtained from the IPA CG of the two systems. The parameters are obtained from the analysis of the occupancy histograms drawn from the FG simulations of the reference systems which consisted of 3 × 3 × 3 supercells of the host materials. Subfigure (a) is referred to the single cell contributions H_n, while subfigure (b) shows the behaviour of mutual interaction parameters K_{n_1,n_2} as a function of the product of two local occupancies. The points represent the original data, the dashed lines represent the fitted functions used for the CG simulations.

found weaker mutual interactions in the ZTC system as compared to the ITQ-29. We believe that this is a consequence of the fact that, in ZTC, spatial correlations between methane molecules localized in neighboring cages are relatively low, due to the weaker confinement effect of the host material. We assessed the quality of the free-energy parameters by comparing the reference FG occupancy histograms with the ones obtained from CG simulations. We found a satisfactory agreement for both systems; such results are shown in the Supporting Information document.

In Fig 6.4 we show the fitting of the kinetic prefactor $k_{M_{12}}$ (we remind that M_{12} is the sum of the occupancies of the two cells of the neighboring pair considered during every inter-cell jump event) for the two systems we considered.

The behaviour of this quantity changes significantly from the zeolite to the ZTC case. In the first case, we clearly distinguish two regimes: for $M_{12} < 20$, $k_{M_{12}}$

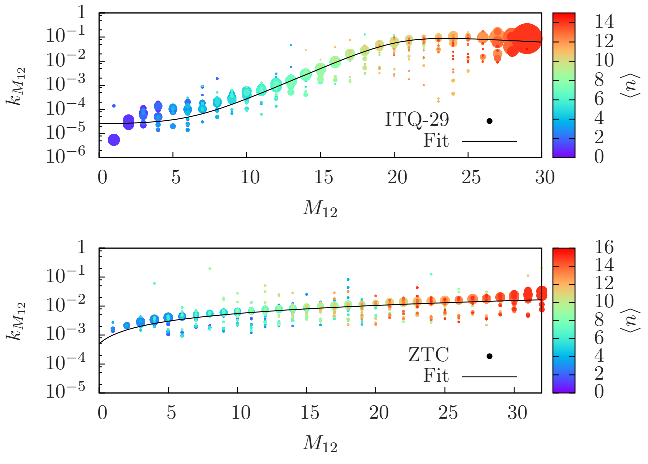

Figure 6.4: Fitting of $k_{M_{12}}$ for the ITQ-29 (upper subfigure) and ZTC (lower subfigure) systems. The y-axis represents the kinetic prefactor $k_{M_{12}}$, while the x-axis represents the summation of the local occupancies $M_{12} = n_1 + n_2$. Each point represents a transition observed during the MD simulations, sized according to the probability of the starting configuration and coloured according to the loading of the simulation where such transition occurred. The black solid lines represent the models used in the CG simulations.

grows relatively fast, following an exponential trend; above $M_{12} = 20$, the prefactor mildly decreases. Conversely, we found a simpler and more uniform behaviour in the methane-ZTC system. In this case, $k_{M_{12}}$ seems to increase linearly respect to the local occupancy. In order to perform our CG simulations, both data sets were fitted to obtain the $k_{M_{12}}$ function for the two systems; the fitting models were designed by prioritizing the most frequent events, for which we expect a better accuracy in the determination of jump probabilities. Hence, we gave priority to the transitions associated with a larger probability for the initial and the final state. A more detailed description of the models and parameters used can be found in the Supporting Information document.

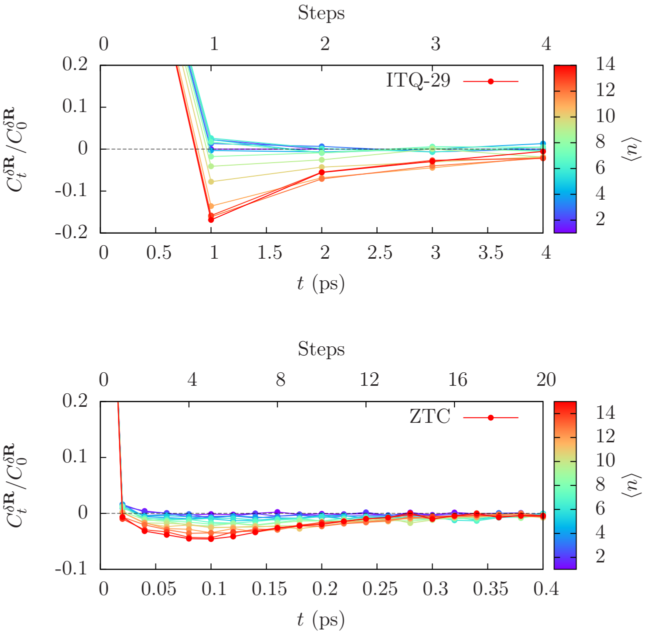

Figure 6.5: Normalized center-of-mass displacement autocorrelation as a function of time, from the MD reference simulations. The upper subfigure is referred to the ITQ-29 system, while the lower subfigure is referred to the ZTC system. The colour represents the loading associated to each MD simulation.

6.3.2 Dynamical correlations

We used the MD trajectories in order to calculate the displacement autocorrelation function $C_t^{\delta\mathbf{R}} = \langle \delta\mathbf{R}_0 \cdot \delta\mathbf{R}_t \rangle$ we previously introduced in the Methods section. For this calculation we only considered the displacements involving inter-cage jumps, in order to filter-out all the intra-cage dynamical effects. The results of this analysis are shown in Fig. 6.5.

For each system, the total number of steps was chosen in such a way as to guarantee a sufficient convergence of the autocorrelation function (Eq. 6.8) to zero. For the zeolite system, we found that the displacement autocorrelation mostly vanishes after 4 simulation steps, which corresponds to 4 ps, thus indicating that for values of time interval τ larger than 4 ps, memory effects would not be observed at all. The results also show that memory effects show up mostly as negative correlations between consecutive displacements; this indicates the importance of the backscattering effect, which is a well-known phenomenon occurring during diffusion through micropores, and takes place every time a molecule fails to thermalize after an inter-cage jump. [96, 120]; the depth and persistence of such effects change with the global density of guest molecules. In fact, we observe that larger backscattering occurs for relatively high gas densities values. This suggests that memory effects depend on the correlations between sorbate molecules. A similar effect is also observed for the ZTC system, for which we obtained a negative correlation effect that vanishes above 0.4 ps; in this case, memory effects tend to decay faster as compared to the ITQ-29 system, indicating a more efficient thermalization. However, considering that for this system we chose a time step equal to 0.02 ps (much shorter than the methane-zeolite case), memory effects vanish after 20 consecutive steps; therefore, under the viewpoint of iterations in the CG model, the backscattering effect is more persistent within the ZTC host.

We used the c.m. displacement autocorrelation functions to calculate the correlation factor for every loading; results are shown in Fig. 6.6 for both systems.

In general, we found dynamical correlations to slow down the diffusion process in both systems we considered. However, we also found significant differences between the two systems in terms of the behaviour of correlation factors as functions of the loading: for the zeolite system, we observe a sigmoid-like decay for f, while for the carbon material we obtained a simple linear decay. Such differences are due to the presence of different microscopic mechanisms contributing to the decay memory effects and to thermalization; resolving such mechanism would require detailed molecular-level investigations of dynamical correlations, which goes beyond the scope of this work (where we are focusing more on the coarse-graining than on

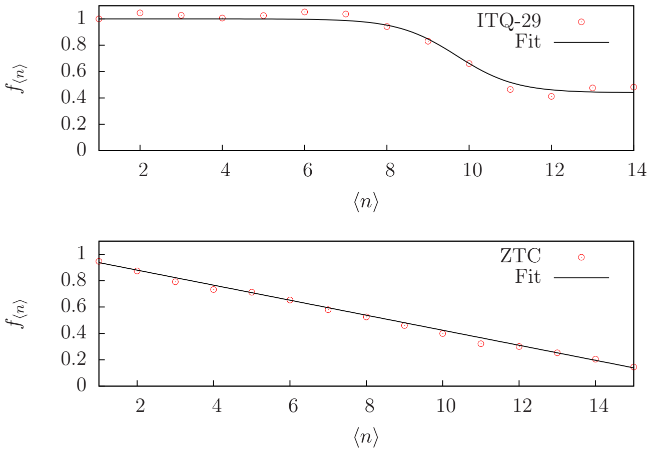

Figure 6.6: Correlation factor $f_{\langle n \rangle}$ as a function of the loading. The upper subfigure is referred to the ITQ-29 system, while the lower subfigure is referred to the ZTC system. The results from the MD simulations are represented by red circles, while our fit is represented by a solid black line.

the molecular-level analysis of the reference FG systems) and will be the object of further contributions. For our purposes, the correlation factor is as a measure of the non-Markovianity of the diffusion process; in fact, f is equal to 1 only if the diffusion is Markovian, which means that memory effects are lost between each time step. We observed such condition in the ITQ-29 system at moderate densities ($\langle n \rangle \leq 7$), and in the ZTC system at $\langle n \rangle = 1$. Our results suggest that for the systems we considered, a purely jump rates-based modelling of diffusion (i.e., if we kept $f = 1$ under all circumstances) would be accurate only for very low sorbate densities; for higher densities, ignoring the dynamical correlations would result in overestimating the jump rates and, consequently, the diffusivity as well.

6.3.3 Diffusivity

We calculated the collective diffusivity as a function of loading, through the Boltzmann-Matano analysis of CG simulations. For both systems, we simulated

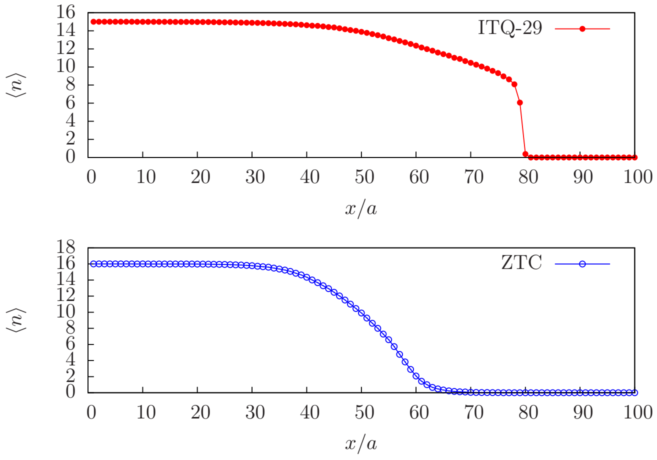

Figure 6.7: Density profiles obtained from the Boltzmann-Matano simulations of the two systems. The upper subfigure is referred to the ITQ-29 system, while the lower subfigure is referred to the ZTC system. The x axis is expressed as number of simulated cavities along x-direction; a is the lattice parameter which is equal to 11.9 Åfor the two materials. The results are obtained with $200 \times 5 \times 5$ CG supercells of the host materials. For simplicity, the plots show only a half of the actual extension of the systems along the x-axis. The ITQ-29 profile was obtained by simulating the dynamics for 7 ns, while in the ZTC case we simulated the system for 0.6 ns. In both cases, the density profiles were averaged over 100 replicas of the sytems.

the relaxation of the density profiles, according to the procedure described in the Methods section. In Fig. 6.7, we show a comparison of the density profiles for the two systems.

For the ITQ-29 system, we observed a first slow decay before a fast step-wise decay of the profile occurring at $\langle n \rangle < 8$. This is the consequence of dramatic differences between diffusivities at low and high densities. The sudden decay of the profile is particularly tricky for the numerical BM analysis, because of the lack of points for the lowest densities, an issue that leads to instabilities during the numerical calculation of the diffusion coefficient. For this reason, we split the profile relaxation experiment into three separate simulations,each starting with its own initial configurations. This procedure is explained in detail in the Supporting

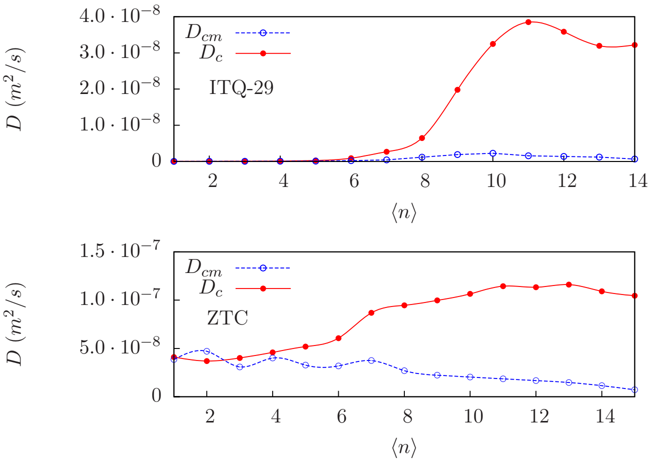

Figure 6.8: Center-of-mass diffusivity D_{cm} (blue points) and collective diffusivity D_c (red points) as a function of the loading for both systems. The upper subfigure shows the results for the ITQ-29 system and the lower subfigure is related to the ZTC system. The smooth lines serve only to help the visualization of the results.

Information document. Conversely, the BM profile for the ZTC system is more smooth and qualitatively closer to the shape of the error function, which is related to a concentration-independent diffusion coefficient.[146]

Our intuitive arguments are confirmed by the trends of collective diffusion coefficients we obtained from numerical calculation. In order to calculate the diffusivity values for all loadings, we numerically solved Eq.(6.11) by using the density profiles. The behaviours of the c.m. diffusivity D_{cm} and the collective diffusivity D_c with respect to the loading are shown , for both systems in Fig. 6.8.

The c.m. diffusivity D_{cm} was calculated from the c.m. mean-squared displacement, which we obtained from the MD trajectories. Collective diffusivities were computed, instead, straight from the BM density profiles obtained from CG simulations. For the ITQ-29 system, we observed a large increase in collective diffusivity for $\langle n \rangle > 7$, with a maximum at $\langle n \rangle = 11$ for which we report $D_c = 3.8 \times 10^{-8}$ m^2/s; this corresponds to an increase by a factor of about 10^3 with respect to the lowest D_c we measured (the lowest D_c was observed at the lowest density investigated, $\langle n \rangle = 1$). Concerning the behaviour of D_{cm}, we found similar results to the ones obtained by Dubbeldam et al., with minor differences due to the slightly different parameterization of the thermostat used in the NVT simulations.[120, 149] In our case, we observed a maximum of $D_{cm} = 2.2 \times 10^{-9}$ m^2/s for $\langle n \rangle = 10$, which is about 10^2 times higher respect to the lowest value reported at $\langle n \rangle = 1$. Results for the ZTC system show milder variations of diffusivities with respect to the loading (we report an increase of collective diffusivity up to 1.2×10^{-7} m^2/s at $\langle n \rangle = 13$), but larger collective diffusivities for the whole loading range. This difference with respect to the methane-zeolite case is mainly due to the larger free volume of the material and, in particular, to the larger windows connecting adjacent cages. Conversely, we obtained a roughly linear decay of the c.m. diffusivity with respect to the loading. In fact, at $\langle n \rangle = 16$ it reaches about half of the initial value; we found this behaviour to be surprisingly similar to the one reported by Beerdsen et al. for the LTL channel-like zeolite,[149, 150] despite ZTC and LTL being very different both in chemical composition and in framework topology, thus suggesting that the diffusive behaviour of methane in ZTC is closer to the diffusion in tube-like structures rather than in cage-like structures like ITQ-29.

The consistency of our diffusivity calculations was validated by comparing the reduced variance $\sigma_N^2/\langle N \rangle$ (which is the reciprocal of the *thermodynamic factor* [138, 139]) as computed from the ratio D_{cm}/D_c, with the same quantity as obtained through GCMC simulations of the FG reference systems. The results of this comparisons are shown in Fig. 6.9. We found a satisfactory agreement between the different data sets for both systems, especially at mid-high values of density. The FG/CG data sets for ZTC exhibit a better overlap compared to the results obtained for the ITQ-29 case. We remark that for the latter system we had a drastically lower number of observed transition in the MD simulations as

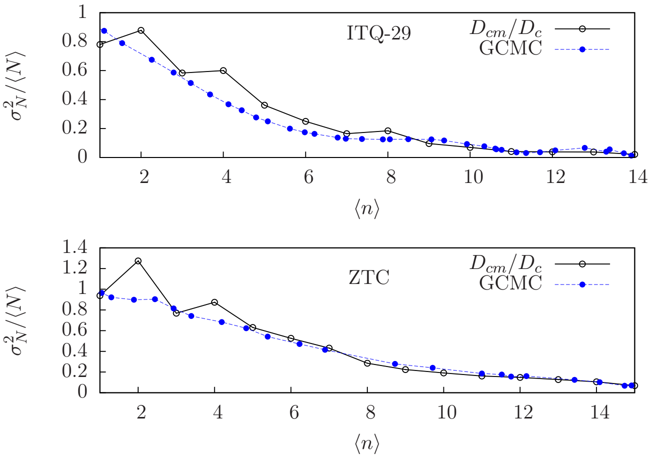

Figure 6.9: Comparison between the reduced variance of the total number of particles and the ratio D_{cm}/D_c for the two systems. The top subfigure represents the results for the ITQ-29 system, while the bottom subfigure shows the results for the ZTC system. The atomistic GCMC results are shown in blue, while the ratio between diffusivities is shown in black. The center-of-mass diffusivity D_{cm} is computed from the MD trajectories, while the D_c values are obtained via BM analysis.

compared to the ZTC case. Hence, we believe that better results could be achieved by longer MD simulations of the reference zeolite system, which would yield more accurate and more robust statistics.

6.3.4 Decay of occupancy correlations

Since our model makes use of local densities only, and since we assume periodic boundary conditions, the sorbate center-of-mass can not be tracked without introducing ambiguities. For this reason, occupancy autocorrelations should be considered as best candidates for measuring the memory decay in CG simulations, rather than correlations in center-of-mass displacements; to this aim we computed the occupancy fluctuations autocorrelation function $C_t^{\delta n} = \langle \delta n_t \cdot \delta n_0 \rangle$, where $\delta n = n - \langle n \rangle$.[151] Fig. 6.10 clearly shows that occupancy autocorrelations

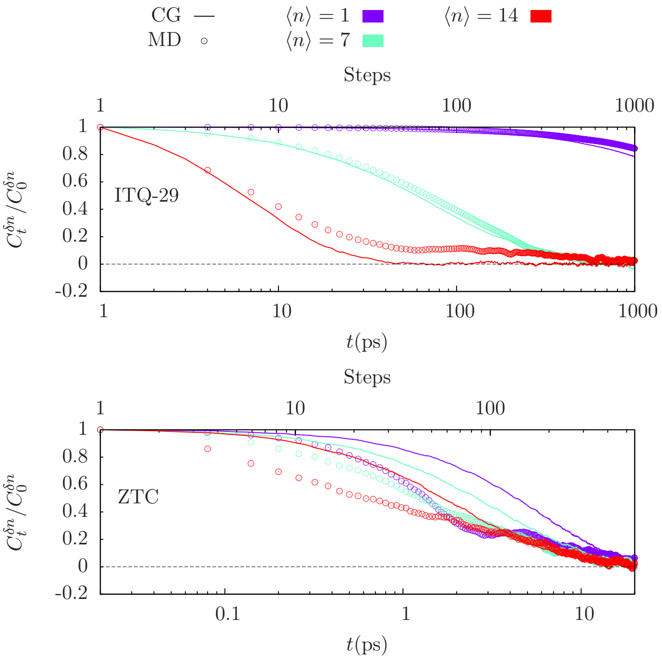

Figure 6.10: Normalized occupancy fluctuations autocorrelation functions as a function of time for the two systems at three different loadings: 1, 7, 14. The upper subfigure is referred to the ITQ-29 system, while the lower is referred to the ZTC system. The empty circles represent the results from MD simulations and the solid lines represent the results of CG simulations.

vanish more rapidly in the methane-ZTC system, mainly because of the faster mass-exchange dynamics, and that for both systems faster relaxations occur at higher density values.

This trend is more evident for the ITQ-29 system, for which we found large differences in D_c between low and high-density regimes—in fact, we observed a direct correlation between the relaxation efficiency and D_c. This general trend is also reproduced by the CG simulations. However, there are evident differences in the agreement between MD and CG results in the two systems: for the ITQ-29, we obtain a semiquantitative agreement between MD and CG relaxation behaviours, especially for lower loading values, while results for the ZTC system exhibit larger discrepancies. We believe that the poorer agreement between the CG and MD data is due to the more markedly non-Markovian nature of mass exchange processes in the carbon material, in relation with the time scale (τ) we chose for such system. In fact, for a purely Markovian process the autocorrelations are expected to decay according to an exponential behaviour;[152] in the present cases, instead, MD results suggest the presence of more complicated relaxation mechanisms, which cause deviations from simple exponential decays. Considering that our model is designed as a first-order Markov chain, our best expectation is the obtainment of an exponential approximation of the reference data. Higher order or multi-time scale transition rate models could allow for the modelling of more complex dynamics and then for a more quantitative matching of occupancy autocorrelation decays; however, as we mentioned while describing the modelling of the transition function W, in that case we would have to face the problem of obtaining statistically meaningful data, necessary to implementing a multivariate transition function, from short atomistic simulations. This will be the object of further contributions.

6.3.5 Computational speedup

Simulating the reference systems with our CG models required a considerable less effort in terms of computational resources. We quantified the efficiency gain in terms of the speedup, S, defined according to Merrick et al.[121]:

$$S = \frac{t_{MD}}{t_{CG}}, \tag{6.12}$$

where t_{MD} and t_{CG} indicate the time, in units of seconds, required to perform the same simulation with the MD and CG representations, respectively. To measure the speedups, we simulated 50 ps of the dynamical evolution of cubic supercells with different sizes (from $3 \times 3 \times 3$ to $6 \times 6 \times 6$) of the reference systems

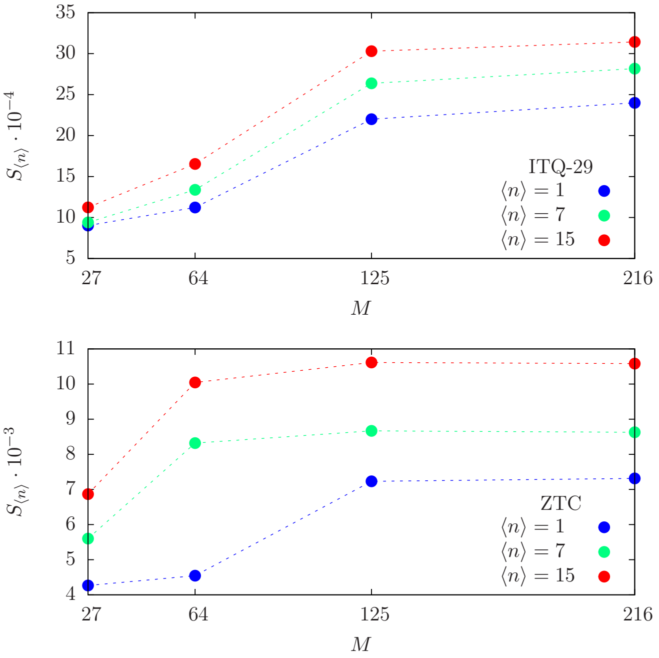

Figure 6.11: Speedup values for both the ITQ-29 and ZTC systems, calculated for different supercell sizes (from 3 × 3 × 3 to 6 × 6 × 6) and loading conditions ($\langle n \rangle = 1, 7, 15$). M is the total number of cells considered in each system.

at different densities ($\langle n \rangle = 1, 7, 15$). We performed our tests on a single CPU core. The results of the speedup calculations are reported in Fig. 6.11.

The results show that the improvement related to the ITQ-29 system is much larger compared to the one related to the ZTC system, this being due to the different time scales (τ) we considered for the transitions in the two systems: every CG iteration for the zeolite corresponds to 1 ps of dynamics, while for the carbon material one CG iteration corresponds to 20 fs. The consequence is that for the zeolite, only 50 iterations are required to cover the dynamics of the speedup tests; while for the ZTC we need to simulate the system for 2500 iterations. We also observe that the speedup is loading dependent, due to the fact that the number of degrees of freedom (DoFs) of our CG representations does not depend on the total number of molecules, but only on the number of simulated cavities of the host materials. Conversely, the computational effort of MD simulations is proportional to the loading, since the number of DoFs is proportional to the total number of guest molecules. Overall, the speedup showed similar trends as a function of the total number of simulated cavities, for both systems. In fact, S raises while the number of cells increases up to 64 or 125, after which S remains constant for bigger systems.

6.4 Conclusions

In this work, we demonstrated a successful way to map atomistic simulations of host-guest systems to occupancy-based lattice models. We focused on the problem of gas molecules confined in microporous materials. In particular, we chose to study methane gas in two different environments: the widely studied pure-silica ITQ-29 zeolite and the LTA-ZTC, a hypothetical carbon material introduced by Braun et al. obtained by the simulated carbon templating of the LTA-zeolite [57].

Our method makes use of statistical data of reference systems as drawn from the results of atomistic simulations: GCMC for static properties and MD for dynamical properties. Our lattice models are equipped with a CG potential function, representing the free-energy of the system, which depends only on local occupancies. The diffusion dynamics is thought of as a composition of several local elementary inter-cage jump events. In our CG representations, we represented such events by employing a strictly local operator, which represents the transition probability associated with each mass-preserving migration event. We modelled the local operator by taking into account the local change in free-energy associated with each transition and a purely kinetic part, which is related to the frequency of migration attempts, and we also proposed a simple way to correct the jump rates for the backscattering contribution on the basis of the displacements autocorrelations observed in the MD simulations; by this way, we allowed for CG models to take into account the non-Markovian memory effects observed in the reference FG systems, which may significantly influence the diffusion in such environments.

We assessed the accuracy of our method by comparing the CG and atomistic results from different perspectives: (i) by comparing static properties in terms of occupancy histograms; (ii) by comparing dynamical properties in terms of the ratio between the diffusion coefficients D_{cm} and D_c, and in terms of the reduced variance $\sigma_N^2/\langle N\rangle$ calculated from GCMC simulations; (iii) by comparing the relaxation behaviours in terms of the decay of autocorrelation of occupancy fluctuations. The results showed a very satisfactory agreement between atomistic and CG results, except for the occupancy relaxation behaviour in strongly non-Markovian scenarios. More sophisticated models would be able to represent such phenomena with better accuracy and will be the object of further contributions; however, we remark that the (very satisfactory) accuracy of the CG model proposed in this work was achieved from small-scale and relatively short atomistic simulations—in fact, obtaining reliable CG representations from short-scale atomistic simulations was the very purpose of our investigation.

Our results showed significant dissimilarities in the properties of the two FG systems we considered, due to the different structure and chemical composition of the two materials. In general, the larger free-volume of the ZTC material led to a weaker localization of the guest molecules resulting in faster inter-cage jump dynamics, more efficient collective diffusion, and weaker inter-cage spatial correlations. The diffusivity behaviour with respect to the loading showed the presence of a strong cage effect in the ITQ-29 material, this resulting in a large peak in diffusivity for $\langle n\rangle = 10, 11$, thus confirming the results shown in previous studies.[120] Conversely, the methane-ZTC system exhibited a mild increase in collective diffusivity and a weak decrease in D_{cm}, thus resulting in the absence of any cage effect and suggesting that this system behaves more as a channel-like material.

Finally, the use of our CG lattice models resulted in a strikingly high computational speedup comparing with the computing time required by the original MD simulations, which allowed for simulating several nanoseconds of dynamics, for very large systems constituted by thousands of the reference materials' unit cells, within a few minutes on a general-purpose computer.

In conclusion, we believe with this work to have established a theoretical framework for the representation of adsorption and diffusion in the mesoscale, starting from the atomistic representation of the reference systems. Our approach can be used to test the mesoscale behaviour of hypothetical systems in possible applications such as gas storage, separation of gas mixtures and sensors design for gaseous species.

Chapter 7

Conclusions and future perspectives

7.1 An *a posteriori* overview

The mesoscopic representation of molecular systems and related physical phenomena still remains a challenging task. In this thesis, I proposed some strategies developed by myself together with my collaborators to cope with the problem of representing adsorption and diffusion through coarse-grained discrete models. This turned out to be a great challenge, especially when trying to define a versatile and consistent methodology that could be successfully applied to a variety of systems and conditions. However, I think that some progress has been made in the development of mesoscopic models of host-guest molecular systems. This is demonstrated by the encouraging results shown in this thesis.

At first, in the third chapter of this thesis, I described a method based on machine-learning techniques which can be used to define a set of molecular states based on recurrent local atomistic patterns. Such method yielded promising results when applied to the problem of adsorption patterns for CO_2 in the ITQ-29 zeolite.

This methodology is still in its early stages and it needs to be tested on more complicated systems, in the effort to uncover the presence of metastable states which would be hardly found either by visual inspection of the MD trajectories, or through a classical kind of descriptor (such as interatomic distances, coordination number etc.).

In the fourth chapter, I introduced the interacting pair approximation (IPA) theoretical framework, as a re-adaptation of an already published article [38]. This framework lays the foundations of the parameterization of occupancy-based models such as the ones used in this thesis. In particular, it deals with the representation of reference static properties —such as the occupancy histograms— through a set of coarse-grained free-energy parameters, based on and consistently with the results of fine-grained GCMC simulations of the chosen systems.

The fifth chapter was devoted to proposing a generalization of the IPA framework, which allows for a larger variety of host-guest systems to be represented at a coarse-grained level [39]. In particular, I showed a possible use of IPA for obtaining a coarse-grained, lattice-based representation of two systems (based on methane and graphene layers), characterized by the presence of two different classes of neighbouring nodes, each of which associated to different sets of mutual interaction parameters and different spatial correlations.

The sixth chapter was focused on a possible approach for obtaining lattice models of two host-guest systems, while equipping them with a coarse-grained representation of the mass-exchange dynamics [40]. This was possible because of the definition of a local evolution operator which was parameterized *via* free-energy contributions obtained through IPA, along with transition rates calculated from MD simulations of the reference systems. In such work, we also presented a simple yet effective way to incorporate non-Markovian effects that may influence diffusion, into the local evolution operator.

Despite all the efforts, the need for a general framework for mapping atomistic host-guest systems into mesoscopic representations has not vanished yet. In particular, the methodologies introduced in this work, although successful, are based on specific sets of assumptions. An ideal protocol would allow for applying the same scheme to the widest possible scenarios. Nevertheless, many interesting questions and possible directions for future research emerged.

7.2 Future perspectives

The future perspectives that arise from the methodologies introduced in this thesis can be grouped into two main categories: refinements and improvements of the mapping methodologies, and possible interesting applications of the coarse-grained representations.

First, the machine-learning approach for the definition of molecular states needs to be tested on more complex scenarios such as mixtures of guest species, complex host materials etc. Such a framework would certainly turn out very useful for comparing different host-guest systems in terms of their local patterns, as well as for deriving coarse-grained models as simplified representations of self-diffusion and/or molecular kinetics in pattern space or in real space.

The occupancy-based modelling framework has to be consolidated in two fronts: first, the IPA methodology should be generalized to the modelling of highly-correlated systems; secondly, the definition of local operators should also be generalized to a wide variety of possible mass-exchange mechanisms. A systematic study of the coarse-grained, IPA-obtained interaction parameters could be carried out in order to reveal the dependence on the temperature; such study would be particularly helpful in clarifying whether, and to which extent, effective interaction parameters at different temperatures can be obtained through interpolation. The problem of mass-exchange phenomena decomposition into elemetary events should be addressed in a general manner; within this context, an efficient classification scheme would be extremely helpful in rationalizing the transition classes observed in FG simulations. Also, different strategies could be developed for including the non-Markovian effects in all the cases in which their contribution to transport and relaxation processes cannot be neglected. In particular, for the latter property, we should be able to define a higher-order (i.e. higher-memory) local operator, that would embed the effects of dynamical correlations more accurately. The whole methodology could also be extended to cover (i) an accurate representation of different guest mixtures within the same host environment, in order to allow for simulating phenomena such as separation, interdiffusion etc.; (ii) as well as non-equilibrium phenomena, even within a local-equilibrium assumption, which would allow simulating at a coarse-grained level the behaviour of systems in the presence of macroscopic gradients of temperature, pressure etc.

Concerning the possible applications of our methodology, I would highlight the possibility of representing the behaviour of density profiles in equilibrium and non-equilibrium scenarios, for systems of technological interest. This means also representing separation processes involving host-guest systems such as the pressure swing adsorption (PSA), which is a widely used technology. Novel materials could be tested through the multi-scale approach described in this thesis in view of a possible utilization in PSA processes; but also for other scopes like gas sensing, heterogeneous catalysis etc. The development of hybrid simulation schemes where this coarse-graining approach is combined with other simulation methods is another very attractive line of research. For example, a permeation process could be represented through molecular dynamics in the region of space located around the interface between the membrane and the gas, whereas a coarse-grained mesoscopic description could be adopted in the representation of the bulk of the host material. A similar approach can be used to represent catalysis, where the reactive centers would be represented by a reactive MD or QM/MM, while the rest of the environment would be represented at CG resolution.

Appendix A

Supplementary material for Chapter 4

A.1 Occupancy distributions for the lattice-gas system

We provide simulation data on the occupancy distributions of the lattice-gas system presented in the fourth chapter, where each cell can contain up to $n_{\max} = 9$ particles. In each figure, the following probability distributions are shown for four selected fugacity values: $p(n)$ (subfigure *a*), that indicates the probability for a single cell to host n particles, and $p(n, m)$ (subfigures *b*-*e*), the joint probability of a pair of neighboring cells to have occupancies n, m. For each of the selected fugacities, the latter probability, $p(n, m)$, is shown in the form of four stacked plots, in every one of which n is kept fixed at a value close to the occupancy at which $p(n)$ reaches its maximum. For example, in Fig. A.1-*a* the maximum in the single-cell histogram $p(\cdot)$ at fugacity $f_1 = 2.49 \cdot 10^3$ bar (black color), is reached for occupancy $n = 1$; the three most probable occupancies that are closest to $n = 1$ are $n = 0$,

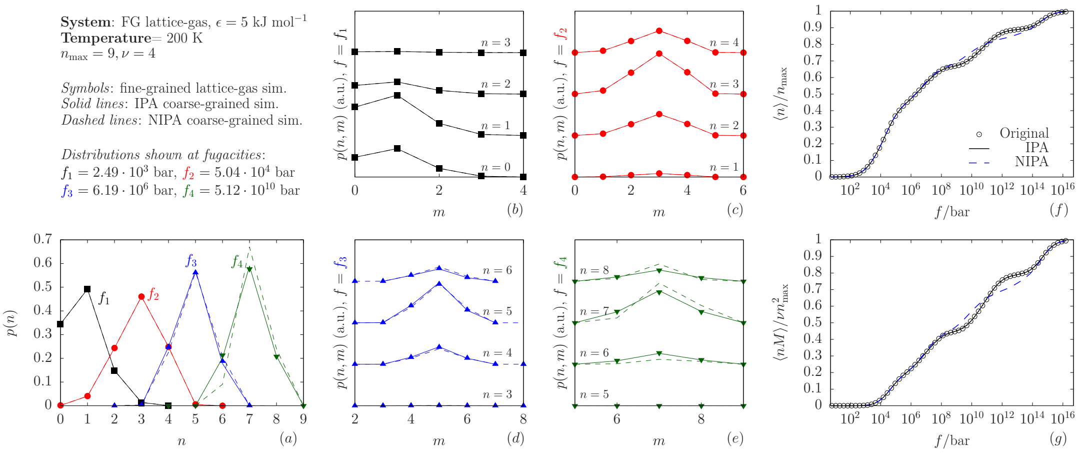

Figure A.1: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for a version of the lattice-gas system in which only plain lateral repulsions are present, and are set to the value of 5 kJ mol^{-1}.

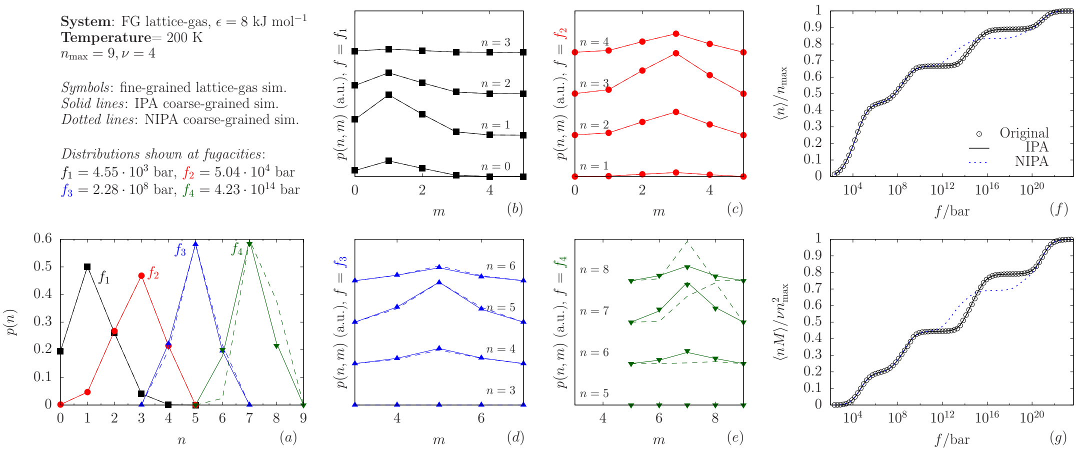

Figure A.2: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for a version of the lattice-gas system in which only plain lateral repulsions are present, and are set to the value of 8 kJ mol^{-1}.

$n = 2$, and $n = 3$, therefore Fig. A.1-*b* we show, beside the histogram of $p(1,\cdot)$, also the histograms of $p(0,\cdot)$, $p(2,\cdot)$, and $p(3,\cdot)$. As the histogram of $p(\cdot)$ was in

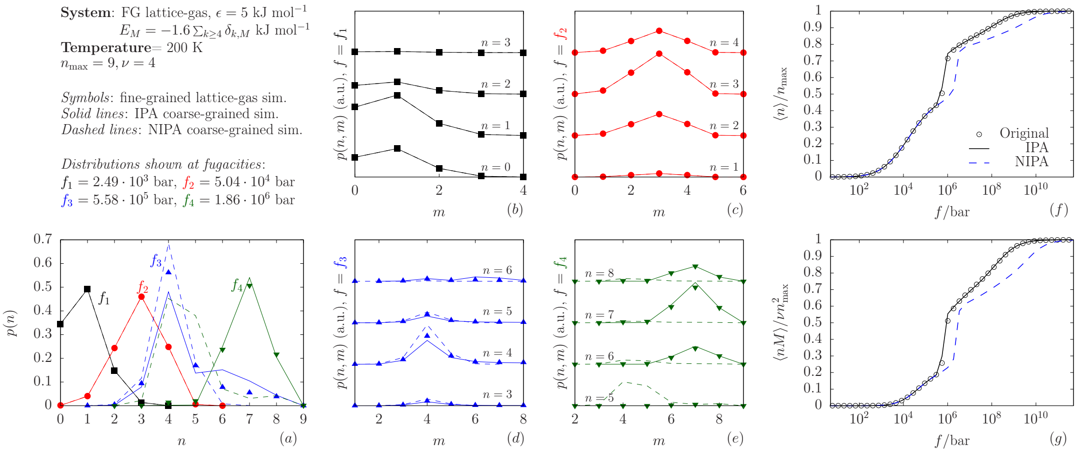

Figure A.3: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for a version of the lattice-gas system in which lateral repulsions are set to the value of 5 kJ mol^{-1}, and extended interactions are added according to parameter values $\phi = -1.6$ kJ mol^{-1} and $M_0 = 4$.

black color, also the stacked histograms $p(0,\cdot)$, $p(1,\cdot)$, $p(2,\cdot)$, and $p(3,\cdot)$ are drawn in black. Analogously, since the maximum of $p(\cdot)$ at $f_2 = 5.04 \cdot 10^4$ bar (red color) is reached at $n = 3$, and the other three most probable occupancies are $n = 1$, $n = 2$, and $n = 4$, the stacked histograms in Fig. A.1-*c* (also in red color) refer to $p(1,\cdot)$, $p(2,\cdot)$, $p(3,\cdot)$, and $p(4,\cdot)$. Obviously, the units for the stacked histograms are arbitrary (a.u.), with the lowest reported value of each histogram defining an invisible baseline of zero probability that applies only for that histogram. Both for the single-cell and the cell-pair probabilities, dots represent values estimated from the fine-grained (FG) simulation. Solid and dashed lines refer instead to the coarse-grained (CG) system, respectively simulated through the parameters obtained by the IPA and the NIPA approaches.

For every system we considered, the subfigure f reports the adsorption isotherm (reported also in the fourth chapter), i.e. the plot of the cell density, expressed as the average occupancy divided by the maximum occupancy, $\langle n \rangle / n_{\max}$, whereas in subfigure g we report the 'extended density', that we defined as the average of the product of the cell occupancy, n, times the sum of the occupancy in its whole neighborhood, M, divided by $\nu n_{\max}$, ν being the number of neighbors of each cell (for the lattice-gases we considered, $\nu = 4$). The extended density, $\langle nM \rangle / (\nu n_{\max})$, gives us information on the correlation between the density in a single cell and the density in its whole neighborhood. We chosen to show it because, since the extended density was *not* matched directly in our coarse-graining, it is a property where FG and the CG system might, in principle, show some dissimilarities. In density and extended density plots, empty circles are used to represent data from the FG system, whereas solid black lines and dashed blue lines are used respectively for IPA and NIPA results.

In all the lattice-gas setups, the temperature was set to the indicative value of $T = 200$ K.

In Fig. A.1 we report data for the lattice-gas system in which the repulsion parameter is set to $\epsilon = 5$ kJ mol^{-1}, and in Fig. A.2, we set the lateral repulsion to $\epsilon = 8$ kJ mol^{-1}. In such cases, the lattice-gas Hamiltonian is simply

$$E(\mathbf{s}) = \epsilon \sum_{\langle i,j \rangle} s_i s_j, \qquad \text{(A.1)}$$

where the sum runs over all the pairs of neighboring sites, and s_i and s_j are the occupancies of sites i and j (each of them being either 0 if empty or 1 if occupied), according to the occupancy configuration $\mathbf{s}$ of the whole FG lattice.

In Fig. A.3 we set it back to $\epsilon = 5$ kJ mol^{-1}, but we added a non-zero next-neighborhood attractive contribution (extended interactions) with parameters $\phi =$

-1.6 kJ mol^{-1} and $M_0 = 4$ In this case, the lattice-gas Hamiltonian is

$$E(\mathbf{s}) = \sum_{\langle i,j \rangle} s_i s_j \Big[\epsilon + \psi(M_i) + \psi(M_j)\Big], \qquad \psi(M) = \phi \sum_{m \geq M_0} \delta_{M,m}, \tag{A.2}$$

where M_i and M_j are defined as the total occupancy in the neighborhood, respectively, of site i, including the occupancy of j, and of site j, including the occupancy of i (see the fourth chapter for further details).

In all our lattice-gas simulations, the IPA approach provided a better match with the FG properties than the NIPA (nearly equal match at the lowest densities, where cell-cell correlations are relatively weak). In particular, the presence of extended interactions causes strong intercell correlations to emerge, in the original FG system (see Fig. A.3). This is particularly evident at intermediate-high densities, i.e. in the nearness of the step in the adsorption isotherm. Under such conditions, the NIPA approach fails, whereas the IPA coarse-graining still holds. The case of fugacity $f_3 = 5.58 \cdot 10^5$ bar (blue histogram in Fig. A.3-a) is especially interesting, since the system is approaching the step in the adsorption isotherm, and the resulting intercell correlations are very strong. Even the IPA approach encounters some difficulty in matching the probability distributions with the same excellent accuracy it granted in other conditions. Nevertheless, the CG-IPA provides an agreement with FG that is far more satisfactory than the CG-NIPA.

A.2 Lennard-Jones system

A.2.1 Coarse-graining of a lone pair of pores under the NIPA

We provide adsorption isotherms for the ‘lone-pair-of-pores’ version of the Lennard-Jones system of methane molecules (united atom approximation) under the static field of the zeolite ITQ-29 framework, where the configurations of methane molecules in *only two neighboring pores are sampled*, through standard Metropolis grand-canonical Monte Carlo (GCMC), while the rest of the system is being kept empty. Under such conditions, the *non-interacting pair approximation* (NIPA) applies [see Eq. (21) in the fourth chapter]. Adsorption isotherms for the lone pair are reported as empty circles in Fig. A.4, at the temperatures of 100, 200, 300, 400, and 500 K. The corresponding adsorption isotherms for the whole system are reported as well (small red dots), in order to highlight the different adsorption properties that emerge as an effect of considering a pair of neighboring pores as it were separated from the rest of the system. We estimated the effective interaction parameters for the coarse-grained (CG) version of the lone pore pair system under the NIPA, and used them to calculate adsorption isotherms (solid blue lines). The agreement is very good, because the NIPA refers precisely to a lone pair of pores. However, once used for CG simulations of the whole system, where pore pairs interact with their surroundings, the NIPA parameters provide a less accurate representation than the ones calculated under the *interacting pair approximation* (IPA) described in the fourth chapter.

A.2.2 Coarse-graining of the full LJ systems

We provide the same kind of data we provided in Section A.1, but here referred to the coarse-graining of the Lennard-Jones system, as a whole.

A.3 Estimation of the interaction terms — Missing entries in the CG interaction matrix

As we mentioned in the fourth chapter, depending on the features of the systems under study, it is possible that some neighboring occupancy pairs are never sampled

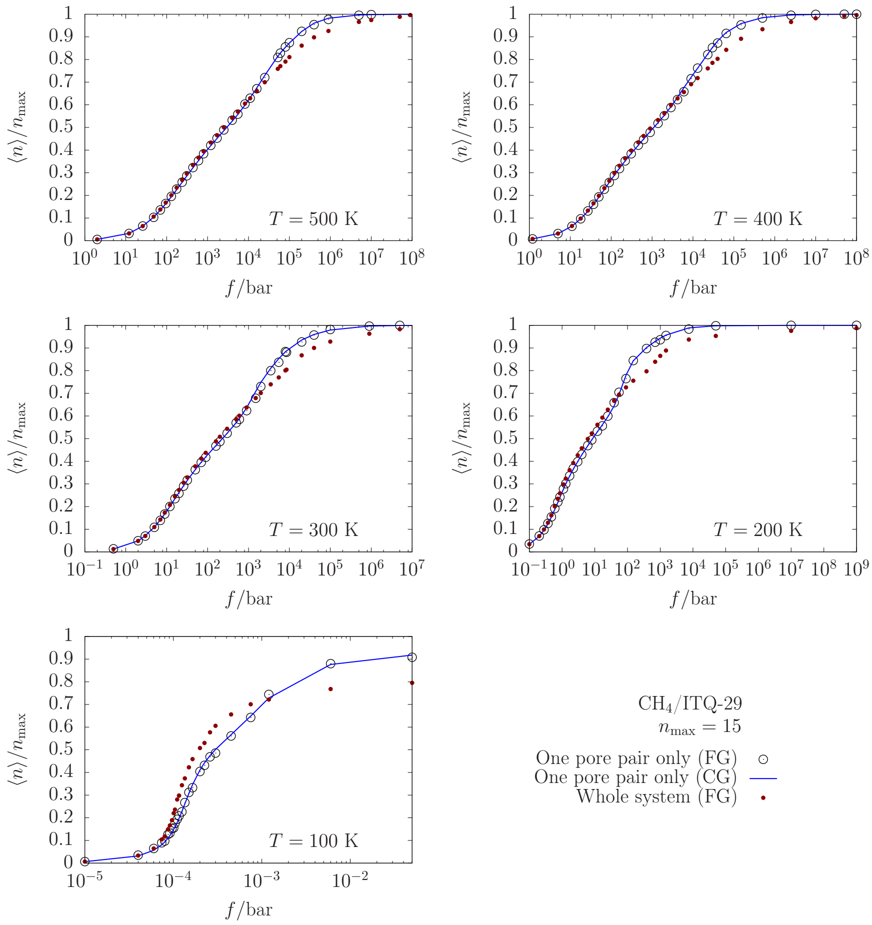

Figure A.4: Adsorption isotherms at various temperatures for a version of the Lennard-Jones system we considered in the fourth chapter, in which the adsorption of methane molecules inside of only two neighboring ITQ-29 pores is simulated through GCMC. Data from fine-grained simulations (empty circles) are shown together with data from the coarse-grained version of the same system (solid blue lines), in which we used the effective interaction parameters calculated under the NIPA. Adsorption isotherms for the full (fine-grained) system are reported as well (small red dots).

at any of the chemical potentials at which the GCMC simulations are performed, causing some entries in the interaction matrix K_{n_1,n_2} (or, equivalently, Z_{n_1,n_2}) to be missing. However, this does not really constitute a problem, since the CG system will simply not sample the occupancy pairs that neither were not sampled in the

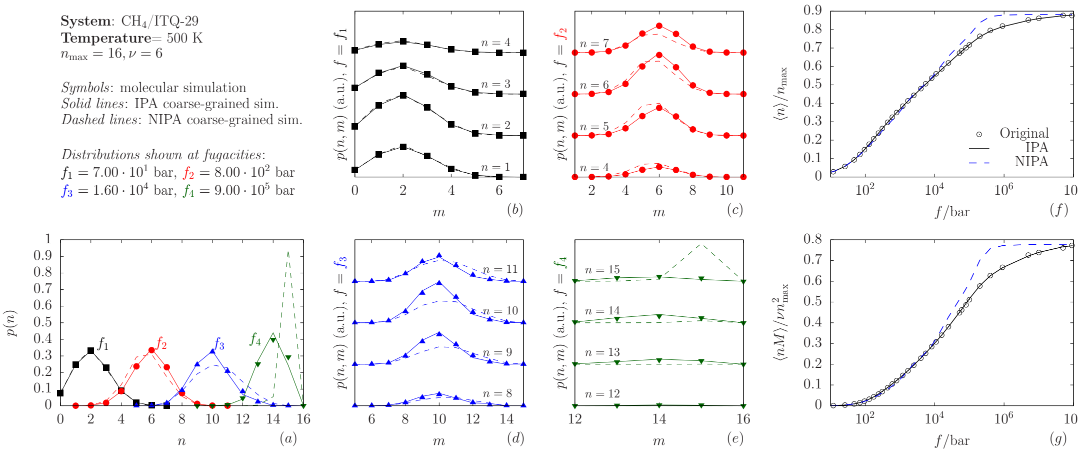

Figure A.5: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for the Lennard-Jones (united-atom-CH_4)/(static-ITQ-29) system at the temperature of 500 K.

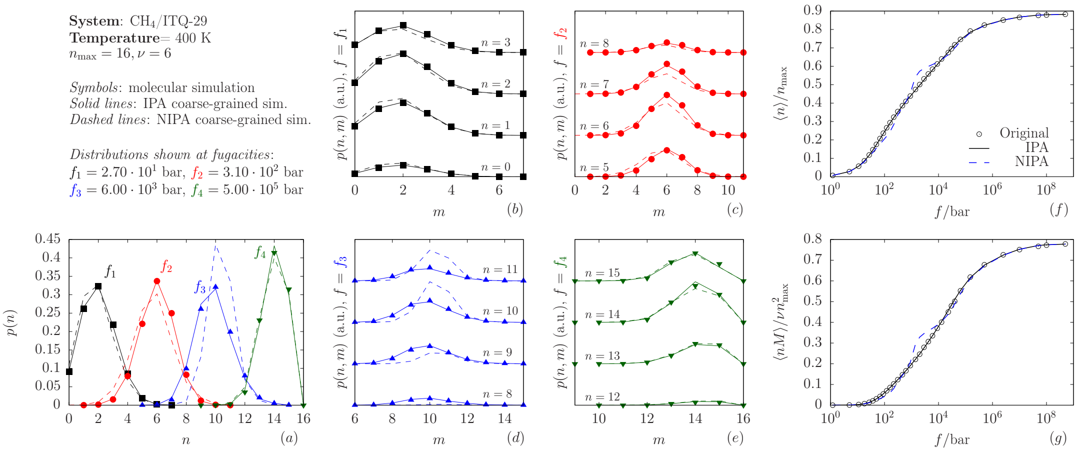

Figure A.6: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for the Lennard-Jones (united-atom-CH_4)/(static-ITQ-29) system at the temperature of 400 K.

original FG system. [41] The only thing that matters is that enough interaction terms could be evaluated, so that the CG system can correctly sample the same occupancy configurations that were sampled in the original system, saturating correctly, rather than, e.g., remaining stuck at some intermediate density because

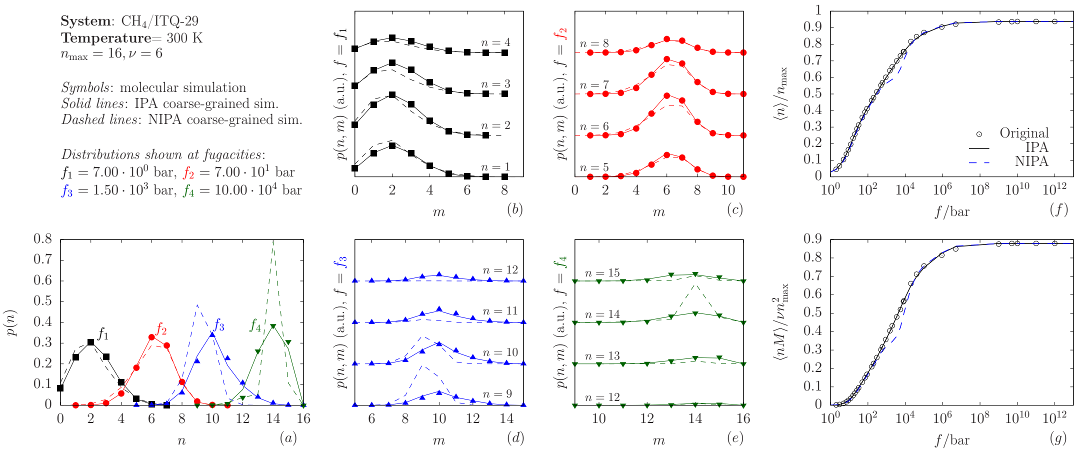

Figure A.7: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for the Lennard-Jones (united-atom-CH_4)/(static-ITQ-29) system at the temperature of 300 K.

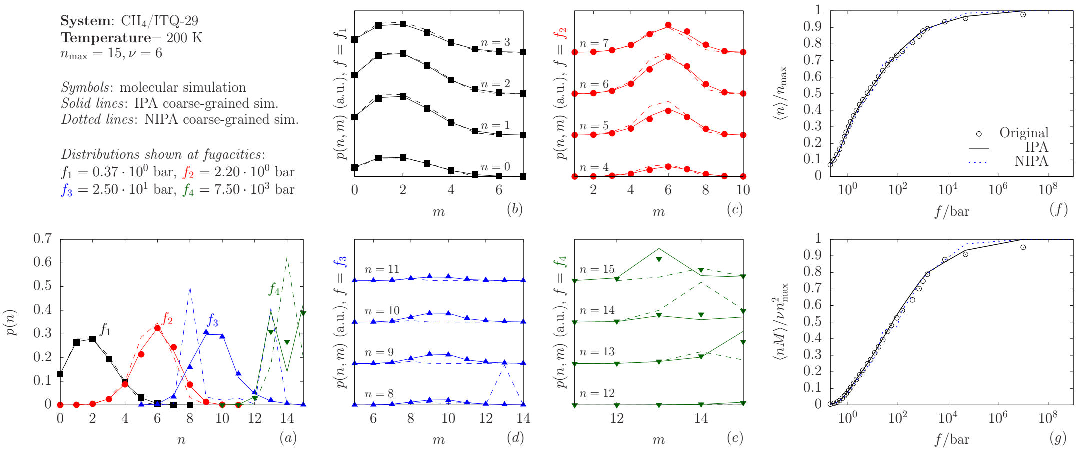

Figure A.8: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for the Lennard-Jones (united-atom-CH_4)/(static-ITQ-29) system at the temperature of 200 K.

of the lack of pair terms at high occupancies. This depends on how accurately the occupancy distributions in the FG system are determined. In earlier works, the NIPA approach was used along with expanded ensemble methods (EEM), [81–84, 153–159] which in this case would essentially prescribe, during the simulation

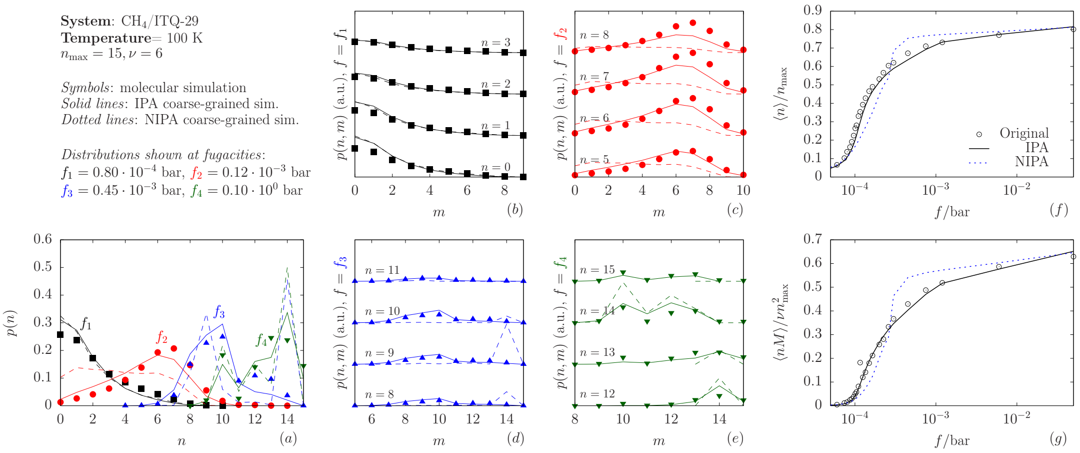

Figure A.9: Selected (a) single-pore and (b-e) pore-pair occupancy distributions, along with (f) the adsorption isotherm and (g) a plot of the extended density *vs.* the fugacity, for the Lennard-Jones (united-atom-CH_4)/(static-ITQ-29) system at the temperature of 100 K.

of the pair as separated from the rest of the system, to assign each pore a different (fictitious) chemical potential, and to find conditions allowing the sampling of all possible pair occupancies, including those that would actually never be sampled by the original FG system, in which the chemical potential is homogeneous. In all the cases we investigated, we found that the NIPA pair-interaction terms, calculated by recursively solving for $Z^*_{n_1,n_2}$ in Eq. (21) in the fourth chapter (with the same procedure we adopted in our previous work about this subject [41]), gave a perfect agreement between isotherms and occupancy distributions of the NIPA system itself, thus basically confirming that, for our purpose of producing the CG version of some FG system at thermodynamic equilibrium, the lack of sampling of pair occupancies that are never sampled in the original system does not affect the quality of the agreement between CG and FG occupancy distributions.

Appendix B

Supplementary material for Chapter 6

B.1 Sensitivity tests

Our coarse-graining method makes an extensive use of models fitted to data obtained from small-scale molecular simulations; we performed several qualitative sensitivity tests to check how the CG collective diffusivity changes in response to perturbations applied to such models. More specifically, we carried out Boltzmann-Matano CG simulations with modified versions of the transition rates; in each modification, one of the following functions was altered: (i) the single-cell free-energy H, (ii) the mutual interactions free-energy K, (iii) the kinetic prefactor k, and (iv) the dynamical correlation factor f. Each function alteration consisted in a transformation of the kind $\phi^*(\mathbf{x}) = \phi(\mathbf{x})(1 + \Delta)$, where Δ represents a percentage change ranging from -8% to $+8\%$. This analysis helped us to clarify the importance of interaction energies, frequency of jump attempts, and dynamical correlations in determining the diffusive properties.

Here, for the sake of simplicity, each BM simulation of the ITQ-29/methane

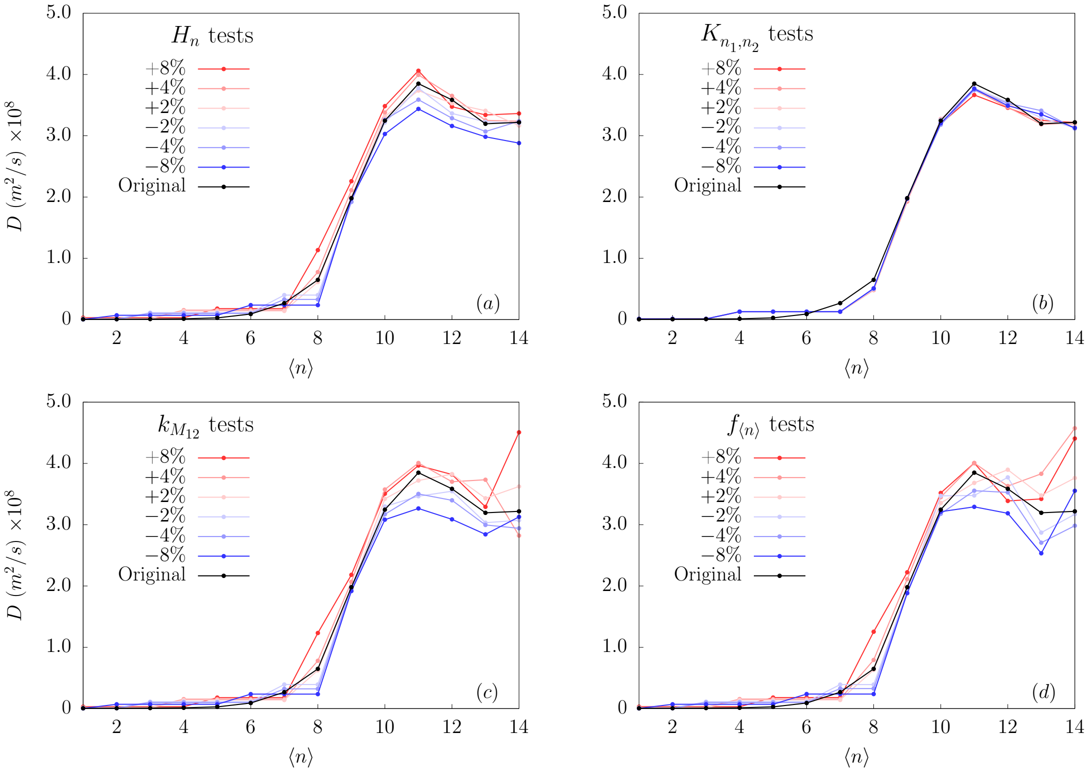

Figure B.1: Sensitivity tests for the ITQ-29/methane system. In each subfigure, the x-axis represents the loading, while the y-axis is the collective diffusivity as obtained through the Boltzmann-Matano simulations of the CG systems with the perturbed models. Each subfigure corresponds to the perturbations of each set of parameters: (a) perturbations of the single cell free-energy contributions H_n; (b) perturbations of the mutual interactions free-energy contributions K_{n_1,n_2}; (c) perturbations of the kinetic prefactor $k_{M_{12}}$; (d) perturbations of the dynamical correlation factor $f_{\langle n\rangle}$.

CG system was conducted as a single run with $n_{max} = 15$ rather than splitting it into three separate runs (which is what we did in order to obtain the results shown in Section 6.3.3) — this is the reason why, at low methane densities, diffusivity slightly differs from the original results.

The results of the sensitivity analysis are presented in Fig.B.1 and Fig.B.2. Overall, the effects of perturbations have a common consequence: the variation in

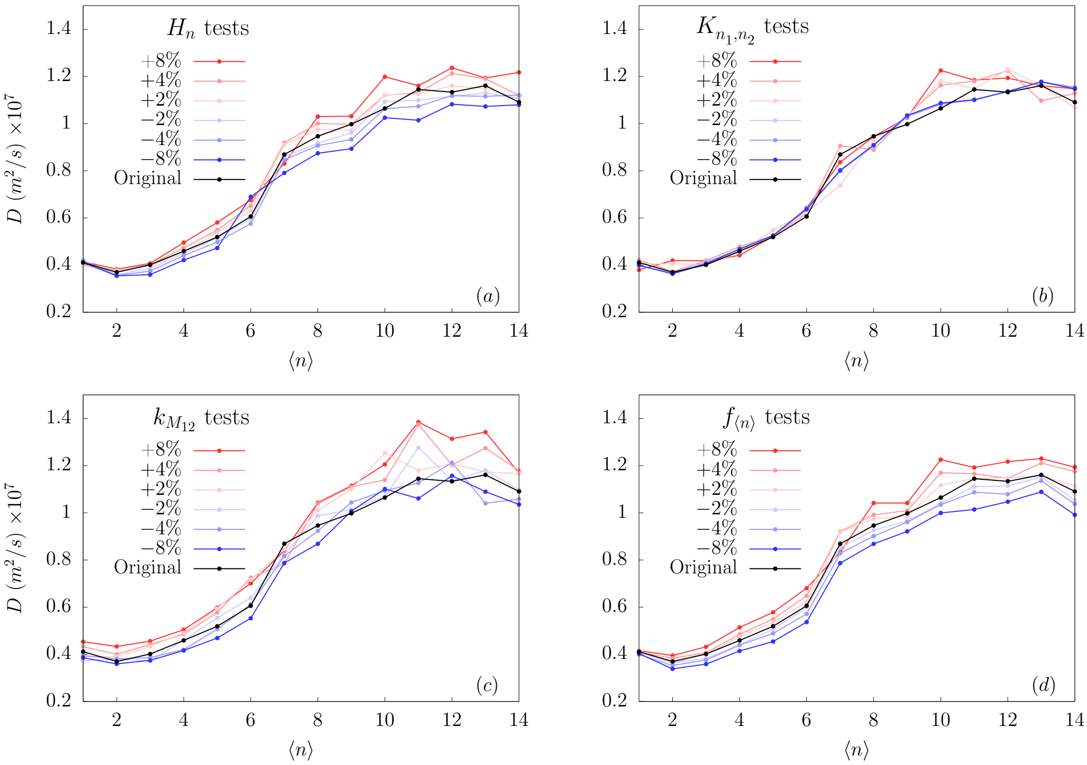

Figure B.2: Sensitivity tests for the ZTC/methane system. In each subfigure, the x-axis represents the loading, while the y-axis is the collective diffusivity as obtained through the Boltzmann-Matano simulations of the CG systems with the perturbed models. Each subfigure corresponds to the perturbations of each set of parameters: (a) perturbations of the single cell free-energy contributions H_n; (b) perturbations of the mutual interactions free-energy contributions K_{n_1,n_2}; (c) perturbations of the kinetic prefactor $k_{M_{12}}$; (d) perturbations of the dynamical correlation factor $f_{\langle n \rangle}$.

collective diffusivity is directly proportional to the variation applied to each model. However, such variations differ remarkably both in magnitude and physical meaning.

Concerning how a change in the free-energy parameters affects the diffusion profile, the variation induced by perturbing the pair-interaction matrix K_{n_1,n_2}

(Figs. B.1b and B.2b) is almost negligble if we compare it to the variation induced by perturbations in the single-cell contributions H_n (Figs. B.1a and B.2a); this comes as a direct consequence of the difference in magnitude between the two sets of parameters. Moreover, *negative* perturbations (underestimation) reflect in a lower diffusivity; this is because the free-energy contributions have a more attractive character — another consequence of this is that, during BM simulations, the desity profile spreads more slowly: attractiveness lowers the effect of concentration gradients as a driving force for collective diffusion. Conversely, *positive* perturbations (overestimation) in free-energy induce a more repulsive behaviour: as guest molecules tend to separate from each other, diffusivity increases.

The changes in diffusivity induced by modifications to $k_{M_{12}}$ and $f_{\langle n \rangle}$, instead, are comparable in magnitude. This is because k and f play similar roles in modelling the transition rates. Overestimation (positive perturbations) of $k_{M_{12}}$ implies a larger frequency of jumps attempts, and therefore results in a higher diffusivity; overestimation of $f_{\langle n \rangle}$ implies underestimation of non-Markovian backscattering effects, which also contributes to raising the diffusivity.

B.2 Fitted models for the coarse-grained simulations

In this section, we report the fitted models we adopted in our coarse-grained (CG) representations. We introduced such models to represent the local free-energy parameters H_n and K_{n_1,n_2}, the kinetic prefactors $k_{M_{12}}$, and the dynamical correlations correction factors $f_{\langle n \rangle}$.

B.2.1 ITQ-29

$$H_n = a_1 n + b_1 n^2, \tag{B.1}$$

with optimal parameter values $a_1 = -26.5$ kJ/mol and $b_1 = 0.96$ kJ/mol.

$$K_{n_1,n_2} = n_1 n_2 \left[a_2 + (n_1 n_2)^3 b_2\right], \tag{B.2}$$

with optimal parameter values $a_2 = -0.06$ kJ/mol and $b_2 = 2.4 \times 10^{-8}$ kJ/mol.

$$k_{M_{12}} = \left[\frac{a_3}{e^{b_3 M_{12}}} + \frac{c_3}{e^{d_3 M_{12}}}\right]^{-1} + e_3, \tag{B.3}$$

with optimal parameter values $a_3 = 1.915$, $b_3 = -0.071$, $c_3 = 1.603 \times 10^6$, $d_3 = 0.600$ and $e_3 = 2.5 \times 10^{-5}$.

$$f_{\langle n\rangle} = \frac{a_4}{1 + e^{b_4(\langle n\rangle + c_4)}} + 1, \tag{B.4}$$

with optimal parameter values $a_4 = -0.56$, $b_4 = -1.4$ and $c_4 = 9.7$. In numerical simulations, the density $\langle n\rangle$ was estimated based on local pair occupancies as $M_{12}/2$.

B.2.2 ZTC

$$H_n = a_1 n + b_1 n^3, \tag{B.5}$$

with optimal parameter values $a_1 = -24.4$ kJ/mol and $b_1 = 0.018$ kJ/mol.

$$K_{n_1,n_2} = n_1 n_2 \left[a_2 + (n_1 n_2)^3 b_2\right], \tag{B.6}$$

with optimal parameter values $a_2 = -0.006$ kJ/mol and $b_2 = 6.0 \times 10^{-10}$ kJ/mol.

$$k_{M_{12}} = a_3 M_{12} + b_3, \tag{B.7}$$

with optimal parameter values $a_3 = 5.0 \times 10^{-4}$ and $b_3 = 5.0 \times 10^{-4}$.

$$f_{\langle n \rangle} = a_4 \langle n \rangle + b_4, \tag{B.8}$$

with optimal parameter values $a_4 = -0.057$ and $b_4 = 0.994$. In numerical simulations the density $\langle n \rangle$ was estimated based on the local pair occupancies as $M_{12}/2$.

B.3 Static properties

In this section, some selected static properties computed from molecular dynamics (MD) are put in comparison with their CG counterparts, for both methane/ITQ-29 and methane/ZTC systems at 300 K and at different loading conditions, i.e. $\langle n \rangle = 4, 7, 10$, representing low-, mid-, and high-density regimes. The selected properties are the following:

- $P(n)$, the probability of observing n particles in a single cell of the system;
- $P(n_1 + n_2)$, the probability of observing a summation of occupancies $n_1 + n_2$ within a pair of connected cells;
- $P(n_1 \times n_2)$, the probability of observing a product of occupancies $n_1 \times n_2$ within a pair of connected cells.

The last two distributions are meant to ease the comparison between CG and MD data sets for the bivariate distributions $P(n_1, n_2)$ which, in principle, would require a comparison between different surfaces in a 3D space.

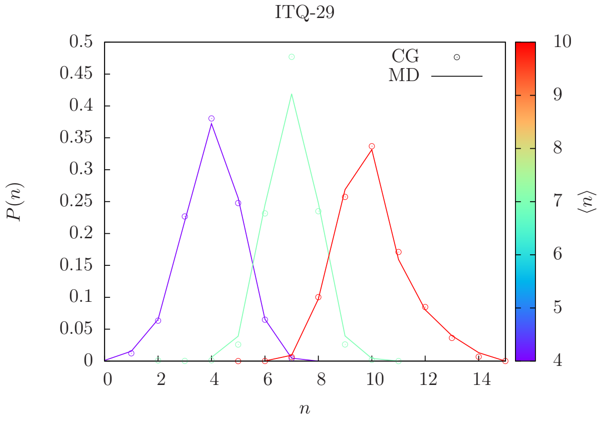

Figure B.3: Single-cell occupancy probability, $P(n)$, at different loadings ($\langle n \rangle$ = 4, 7, 10), for the ITQ-29 system. Results from CG (MD) simulations are indicated as empty circles (solid lines).

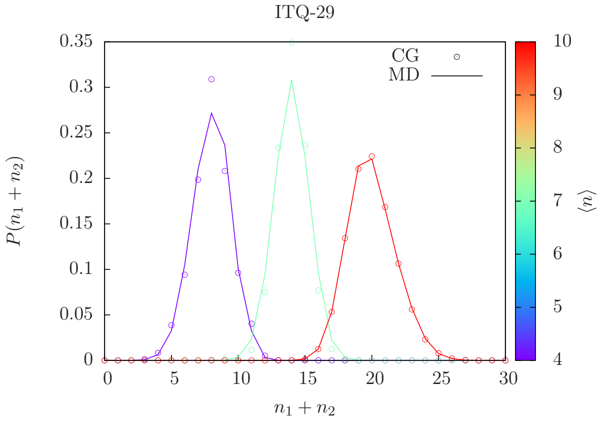

Figure B.4: Neighboring occupancies summation probability, $P(n_1 + n_2)$, at different loadings ($\langle n \rangle$ = 4, 7, 10), for the ITQ-29 system. Results from CG (MD) simulations are indicated as empty circles (solid lines).

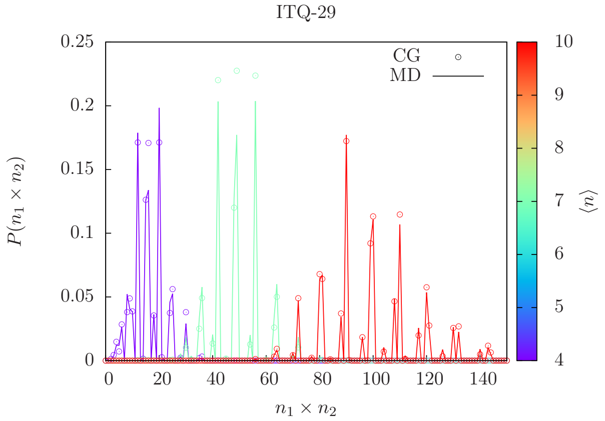

Figure B.5: Neighboring occupancies product probability, $P(n_1 \times n_2)$, at different loadings ($\langle n \rangle$ = 4, 7, 10), for the ITQ-29 system. Results from CG (MD) simulations are indicated as empty circles (solid lines).

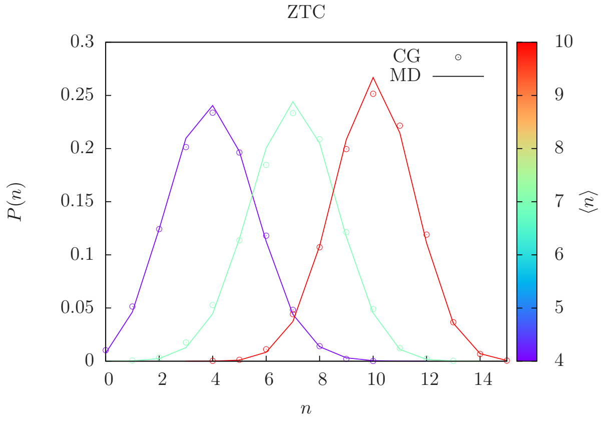

Figure B.6: Single-cell occupancy probability, $P(n)$, at different loadings ($\langle n \rangle$ = 4, 7, 10), for the ZTC system. Results from CG (MD) simulations are indicated as empty circles (solid lines).

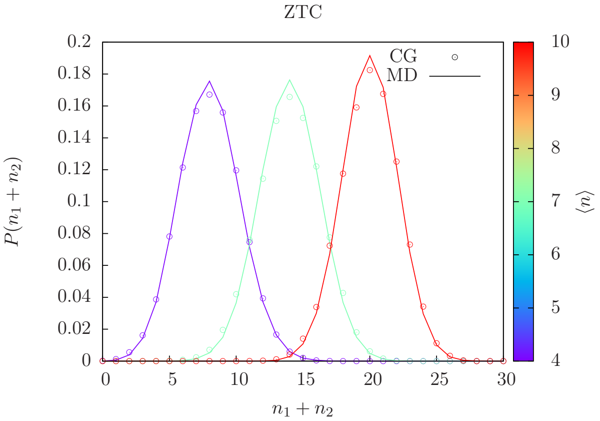

Figure B.7: Neighboring occupancies summation probability, $P(n_1+n_2)$, at different loadings ($\langle n \rangle = 4$, 7, 10), for the ZTC system. Results from CG (MD) simulations are indicated as empty circles (solid lines).

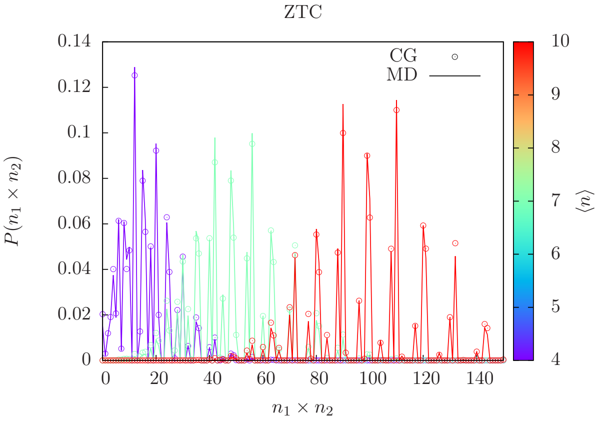

Figure B.8: Neighboring occupancies product probability, $P(n_1 \times n_2)$, at different loadings ($\langle n \rangle = 4$, 7, 10), for the ZTC system. Results from CG (MD) simulations are indicated as empty circles (solid lines).

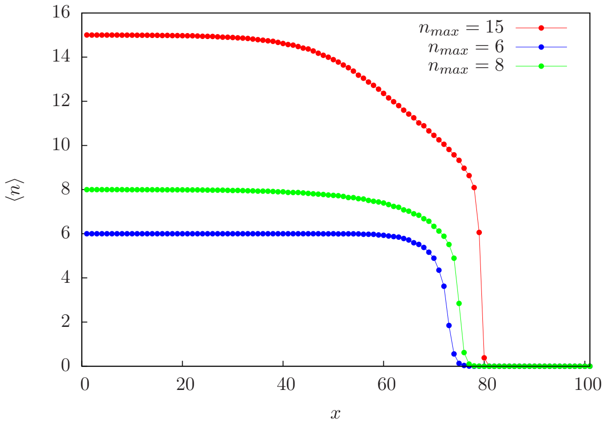

Figure B.9: Density profiles from three different BM simulations of the ITQ-29 system: n_{max} = 15 is shown in red, n_{max} = 8 is shown in green, n_{max} = 6 is shown in blue.

B.4 Boltzmann-Matano simulations of the ITQ-29 system

The large changes in collective diffusivity between low- and high-density regimes for this system caused the density profile to steeply decrease at $\rho \leq 8$. This resulted in instabilities during the numerical integration and differentiation of the density profiles. For this reason, we carried out the Boltzmann-Matano (BM) simulations of the ITQ-29 system in three different versions, each one with a different value for the maximum occupancy n_{max}. The density profiles we obtained are shown in Fig.B.9.

We empirically found that setting n_{max} to 15, 8 and 6 was a good compromise between stability and computational effort. The collective diffusivity (D_c) values for the ITQ-29 system were drawn from the different profiles in order to maximize the stability:

- the $n_{max} = 15$ profile was used to calculate D_c for $\langle n \rangle \geq 8$;
- the $n_{max} = 8$ profile was used to calculate D_c for $\langle n \rangle = 7{,}6$;

- the $n_{max} = 6$ profile was used to calculate D_c for $\langle n \rangle \leq 6$.

Since at low density the diffusivity is significantly lower than in high-density scenarios, the simulations with $n_{max} = 6$ and $n_{max} = 8$ required a larger number of simulated iterations (5 times respect to the $n_{max} = 15$) to reach the profiles shown in Fig. B.9.

Bibliography

[1] E. Gillman and M. Gillman, *Modelling Nature*, CABI, 2019.

[2] R. Car and M. Parrinello, *Phys. Rev. Lett.* **55**, 2471 (1985).

[3] D. Marx and M. Parrinello, *J. Chem. Phys.* **104**, 4077 (1996).

[4] B. J. Alder and T. E. Wainwright, *J. Chem. Phys.* **27**, 1208 (1957).

[5] B. S. D. Frenkel, *Understanding molecular simulations - From algorithms to applications*, Academic Press, 2 edition, 2002.

[6] J. F. Wendt, *Computational Fluid Dynamics*, Springer, 2009.

[7] E. R. Smith, P. E. Theodorakis, R. V. Craster, and O. K. Matar, *Langmuir* **34**, 12501 (2018).

[8] J. Hardy, Y. Pomeau, and O. de Pazzis, *Phys. Rev. Lett.* **31**, 276 (1973).

[9] D. A. Wolf-Gladrow, *Lattice-Gas Cellular Automata and Lattice Boltzmann Models*, Springer, 2000.

[10] E. Ising, *Zeitschrift für Physik* **31**, 253 (1925).

[11] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, *Proceedings of the National Academy of Sciences* **106**, 19011 (2009).

[12] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schatte, and F. Noé, *J. Chem. Phys.* **134**, 174105 (2011).

[13] S. Izvekov and G. A. Voth, *J. Phys. Chem. B* **109**, 2469 (2005).

[14] W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, *J. Chem. Phys.* **128**, 244114 (2008).

[15] M. A. Webb, J.-Y. Delannoy, and J. J. de Pablo, *J. Chem. Theory Comput.* **15**, 1199 (2019).

[16] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, and N. F. A. van der Vegt, *Soft Matter* **9**, 2108 (2013).

[17] A. J. Rzepiela, M. Louhivuori, C. Peter, and S. J. Marrink, *Phys. Chem. Chem. Phys.* **13**, 10437 (2011).

[18] A. J. Pak, T. Dannenhoffer-Lafage, J. J. Madsen, and G. A. Voth, *J. of Chem. Theo. and Comp.* **15**, 2087 (2019).

[19] S. Izvekov and G. A. Voth, *J. Chem. Phys.* **123**, 134105 (2005).

[20] A. Das and H. C. Andersen, *J. Chem. Phys.* **131**, 034102 (2009).

[21] D. Reith, M. Pütz, and F. Müller-Plathe, *J. Comput. Chem.* **24**, 1624 (2003).

[22] K. R. Hadley and C. McCabe, *J. Chem. Phys.* **132**, 134505 (2010).

[23] T. C. Moore, C. R. Iacovella, and C. McCabe, *J. Chem. Phys.* **140**, 224104 (2014).

[24] Y. Han, J. F. Dama, and G. A. Voth, *J. Chem. Phys.* **149**, 044104 (2018).

[25] S. Ma, *Phys. Rev. Lett* **37**, 461 (1976).

[26] M. Katsoulakis, A. J. Majda, and D. G. Vlachos, *Proc. Natl. Acad. Sci. USA* **100**, 782 (2003).

[27] A. Chatterjee, D. G. Vlachos, and M. Katsoulakis, *J. Chem. Phys.* **121**, 11420 (2004).

[28] J. Dai, W. D. Seider, and T. Sinno, *J. Chem. Phys.* **128**, 194705 (2008).

[29] X. Liu, W. D. Seider, and T. Sinno, *Phys. Rev. E* **86**, 026708 (2012).

[30] M. Katsoulakis and P. Plecháč, *Math. Comp.* **83**, 1757 (2014).

[31] J. J. de la Torre and P. Español, *J. Chem. Phys.* **135**, 114103 (2011).

[32] N. Israeli and N. Goldenfeld, *Phys. Rev. E* **73**, 026203 (2006).

[33] A. F. Voter, INTRODUCTION TO THE KINETIC MONTE CARLO METHOD, in *Radiation Effects in Solids*, edited by K. E. Sickafus, E. A. Kotomin, and B. P. Uberuaga, pp. 1–23, Dordrecht, 2007, Springer Netherlands.

[34] B. Chopard and M. Droz, *Cellular Automata Modeling of Physical Systems*, Collection Alea-Saclay: Monographs and Texts in Statistical Physics, Cambridge University Press, 1998.

[35] G. Milano and T. Kawakatsu, *The Journal of Chemical Physics* **130**, 214106 (2009).

[36] S. Qi, H. Behringer, and F. Schmid, *New Journal of Physics* **15**, 125009 (2013).

[37] A. P. Sgouros, G. G. Vogiatzis, G. Megariotis, C. Tzoumanekas, and D. N. Theodorou, *Macromolecules* **52**, 7503 (2019).

[38] F. G. Pazzona, G. Pireddu, A. Gabrieli, A. M. Pintus, and P. Demontis, *J. Chem. Phys.* **148**, 194108 (2018).

[39] G. Pireddu, F. G. Pazzona, A. M. Pintus, A. Gabrieli, and P. Demontis, *J. Phys. Chem. C* **123**, 18355 (2019).

[40] G. Pireddu, F. G. Pazzona, P. Demontis, and M. A. Załuska-Kotur, *J. of Chem. Theo. and Comp.* (2019).

[41] F. G. Pazzona, P. Demontis, and G. B. Suffritti, *J. Phys. Chem. C* **118**, 28711 (2014).

[42] M. Ceriotti, *J. Chem. Phys.* **150**, 150901 (2019).

[43] J. Behler, *J. Chem. Phys.* **134**, 074106 (2011).

[44] A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).

[45] J. Behler, *International Journal of Quantum Chemistry* **115**, 1032 (2015).

[46] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).

[47] B. A. Helfrecht, R. Semino, G. Pireddu, S. M. Auerbach, and M. Ceriotti, *The Journal of Chemical Physics* **151**, 154112 (2019).

[48] D. A. McQuarrie, *Statistical Mechanics*, Harper and Row, New York, first edition, 1976.

[49] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, Inc., New York, NY, USA, 2nd edition, 2017.

[50] S. Nosé, *J. Chem. Phys.* **81**, 511 (1984).

[51] W. G. Hoover, *Phys. Rev. A* **31**, 1695 (1985).

[52] A. Huang, C. Weidenthaler, and J. Caro, *Microporous and Mesoporous Materials* **130**, 352 (2010).

[53] S. M. Auerbach, K. A. Carrado, and P. K. Dutta, *Handbook of Zeolite Science and Technology*, CRC Press, 2003.

[54] H. Nishihara and T. Kyotani, *Chem. Commun.* **54**, 5648 (2018).

[55] N. P. Stadie, M. Murialdo, C. C. Ahn, and B. Fultz, *J. Phys. Chem. C* **119**, 26409 (2015).

[56] N. P. Stadie, M. Murialdo, C. C. Ahn, and B. Fultz, *J. Am. Chem. Soc.* **135**, 990 (2013).

[57] E. Braun, Y. Lee, S. M. Moosavi, S. Barthel, R. Mercado, I. A. Baburin, D. M. Proserpio, and B. Smit, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E8116 (2018).

[58] `https://archive.materialscloud.org/file/2018.0013/v1/structure-files.tar.gz`.

[59] M. J. Allen, V. C. Tung, and R. B. Kaner, *Chemical Reviews* **110**, 132 (2010).

[60] G. Ersan, O. G. Apul, F. Perreault, and T. Karanfil, *Water Research* **126**, 385 (2017).

[61] A. Hassani, M. T. H. Mosavian, A. Ahmadpour, and N. Farhadian, *J. Chem. Phys.* **142**, 234704 (2015).

[62] M. Ceriotti, G. A. Tribello, and M. Parrinello, *Proceedings of the National Academy of Sciences* **108**, 13023 (2011).

[63] B. Scholkopf, A. Smola, and K.-R. Müller, *Neural Computation* **10**, 1299 (1998).

[64] P. Gasparotto and M. Ceriotti, *J. Chem. Phys.* **141**, 174110 (2014).

[65] P. Gasparotto, R. H. Meissner, and M. Ceriotti, *J. of Chem. Theo. and Comp.* **14**, 486 (2018).

[66] `https://github.com/cosmo-epfl/pamm`.

[67] R. T. Cygan, V. N. Romanov, and E. M. Myshakin, *J. Phys. Chem. C* **116**, 13079 (2012).

[68] R. Krishna and J. van Baten, *Separation and Purification Technology* **61**, 414 (2008).

[69] L. A. Clark, A. Gupta, and R. Q. Snurr, *J. Phys. Chem. B* **102**, 6720 (1998).

[70] A. K. Soper, *Chem. Phys.* **202**, 295 (1996).

[71] M. S. Shell, *J. Chem. Phys.* **129**, 144108 (2008).

[72] I. Bilionis and N. Zabaras, *J. Chem. Phys.* **138**, 044313 (2013).

[73] J. F. Rudzinski and W. G. Noid, *J. Chem. Phys.* **135**, 214101 (2011).

[74] S. T. John and G. Csányi, *J. Phys. Chem. B* **121**, 10934 (2017).

[75] A. Tsourtis, V. Harmandaris, and D. Tsagkarogiannis, *Entropy* **19**, 395 (2017).

[76] M. Katsoulakis and D. G. Vlachos, Mathematical strategies for the coarse-graining of microscopic models, in *Handbook of Materials Modeling*, edited by S. Yip, pp. 1477–1490, Springer, Dordrecht, 2005.

[77] P. R. Van Tassel, H. T. Davis, and A. V. McCormick, *J. Chem. Phys.* **98**, 8919 (1993).

[78] R. Q. Snurr, A. T. Bell, and D. N. Theodorou, *J. Phys. Chem.* **98**, 5111 (1994).

[79] C. Saravanan, F. Jousse, and S. M. Auerbach, *Phys. Rev. Lett.* **80**, 5754 (1998).

[80] K. F. Czaplewski and R. Q. Snurr, *AIChE J.* **45**, 2223 (1999).

[81] C. Tunca and D. Ford, *J. Chem. Phys.* **111**, 2751 (1999).

[82] C. Tunca and D. Ford, *J. Phys. Chem. B* **106**, 10982 (2002).

[83] C. Tunca and D. Ford, *Chem. Eng. Sci.* **58**, 3373 (2003).

[84] C. Tunca and D. Ford, *J. Chem. Phys.* **120**, 10763 (2004).

[85] P. Demontis and G. B. Suffritti, *J. Phys. Chem. B* **101**, 5789 (1997).

[86] S. M. Auerbach, *Int. Rev. Phys. Chem.* **19**, 155 (2000).

[87] S. M. Auerbach, F. Jousse, and D. P. Vercauteren, Dynamics of sorbed molecules in zeolites, in *Computer modelling of microporous and*

mesoporous materials, edited by C. R. A. Catlow, R. A. van Santen, and B. Smit, pp. 49–108, Elsevier, Amsterdam, 2004.

[88] J. W. Wagner, T. Dannenhoffer-Lafage, J. Jin, and G. A. Voth, *J. Chem. Phys.* **147**, 044113 (2017).

[89] A. A. Louis, *J. Phys.: Condens. Matter* **14**, 9187 (2002).

[90] M. R. DeLyser and W. G. Noid, *J. Chem. Phys.* **147**, 134111 (2017).

[91] F. G. Pazzona, A. Gabrieli, A. Pintus, P. Demontis, and G. B. Suffritti, *J. Chem. Phys.* **134**, 184109 (2011).

[92] T. Becker, K. Nelissen, B. Cleuren, B. Partoens, and C. Van der Broeck, *Phys. Rev. Lett.* **111**, 110601 (2013).

[93] M. Vieth, A. Kolinski, and J. Skolnick, *J. Chem. Phys.* **102**, 6189 (1995).

[94] H. A. Bethe, *J. Comput. Chem.* **24**, 1624 (2003).

[95] K. Huang, *Statistical mechanics*, Wiley, New York, 2000.

[96] P. Demontis, L. Fenu, and G. B. Suffritti, *J. Phys. Chem. B* **109**, 18081 (2005).

[97] E. Beerdsen, D. Dubbeldam, and B. Smit, *Phys. Rev. Lett.* **93**, 248301 (2004).

[98] D. Dubbeldam, E. Beerdsen, T. J. H. Vlugt, and B. Smit, *J. Chem. Phys.* **122**, 224712 (2005).

[99] E. Beerdsen, D. Dubbeldam, and B. Smit, *J. Phys. Chem. B* **110**, 22754 (2006).

[100] P. Demontis, F. G. Pazzona, and G. B. Suffritti, *J. Phys. Chem. B* **112**, 12444 (2008).

[101] A. M. Pintus, F. G. Pazzona, P. Demontis, and G. B. Suffritti, *J. Chem. Phys.* **135**, 124110 (2011).

[102] A. M. Pintus, F. G. Pazzona, P. Demontis, and G. B. Suffritti, *J. Chem. Phys.* **145**, 184115 (2015).

[103] T. J. H. Vlugt and M. Schenk, *J. Phys. Chem. B* **106**, 12757 (2002).

[104] R. L. June, A. T. Bell, and D. N. Theodorou, *J. Phys. Chem.* **94**, 1508 (1990).

[105] S. Kullback and R. A. Leibler, *Ann. Math. Statist.* **22**, 79 (1951).

[106] J. Güémez, S. Velasco, and A. Calvo Hernändez, *Physica A* **152**, 226 (1988).

[107] T. T. P. Cheung, *J. Phys. Chem.* **97**, 8993 (1993).

[108] M. Müller and J. J. de Pablo, Simulation techniques for calculating free energies, in *Computer simulations in condensed matter systems: From materials to chemical biology, volume 1*, edited by M. Ferrario, G. Ciccotti, and K. Binder, pp. 67–126, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[109] G. Soto-Campos, D. S. Corti, and H. Reiss, *J. Chem. Phys.* **108**, 2563 (1998).

[110] C. Saravanan, F. Jousse, and S. M. Auerbach, *Phys. Rev. Lett.* **80**, 5754–5757 (1998).

[111] K. G. Ayappa, C. R. Kamala, and T. A. Abinandanan, *J. Chem. Phys.* **110**, 8714 (1999).

[112] F. G. Pazzona, P. Demontis, and G. B. Suffritti, *J. Phys. Chem. C* **117**, 349 (2013).

[113] I. Langmuir, *J. Am. Chem. Soc.* **40**, 1361–1403 (1918).

[114] L. Guo, L. Xiao, X. Shan, and X. Zhang, *Sci. Rep.* **6** (2016).

[115] A. Tarasenko, *Surf. Sci.* **679**, 284 (2019).

[116] M. Sudibandriyo, S. A. Mohammad, R. L. R. Jr., and K. A. M. Gasem, *Fl. Ph. Eq.* **299**, 238 (2010).

[117] N. Yang, D. Yang, G. Zhang, L. Chen, D. Liu, M. Cai, and X. Fan, *Sensors* **18**, 422 (2018).

[118] A. Pedrielli, S. Taioli, G. Garberoglio, and N. M. Pugno, *Microp. and Mesop. Mat.* **257**, 222 (2015).

[119] M. Abbaspour, H. Akbarzadeh, S. Salemi, and M. Abroodi, *Phys. A* **462**, 1075 (2016).

[120] D. Dubbeldam, E. Beerdsen, T. J. H. Vlugt, and B. Smit, *J. Chem. Phys.* **122**, 224712 (2005).

[121] M. Merrick and K. A. Fichthorn, *Physical Review E* **75**, 011606 (2007).

[122] J. P. Hansen and I. R. McDonald, *Theory of simple liquids*, Academic Press, London, 2 edition, 1986.

[123] T. M. Truskett, S. Torquato, and P. G. Debenedetti, *Phys. Rev. E* **58**, 7369 (1998).

[124] J. Vekeman, I. G. Cuesta, N. Faginas-Lago, J. Wilson, J. Sánchez-Marína, and A. S. de Merás, *Physical Chemistry Chemical Physics* **20**, 25518 (2018).

[125] M. Levitt and A. Warshel, *Nature* **253**, 694 (1975).

[126] S. Chen and G. D. Doolen, *Ann. Rev. of Fluid Mech.* **30**, 329 (1998).

[127] A. Bortz, M. Kalos, and J. Lebowitz, *J. Comp. Phys.* **17**, 10 (1975).

[128] K. A. Fichthorn and W. H. Weinberg, *J. Chem. Phys.* **95**, 1090 (1991).

[129] A. Gabrieli, P. Demontis, F. G. Pazzona, and G. B. Suffritti, *Phys. Rev. E* **83**, 056705 (2011).

[130] T. Toffoli and N. H. Margolus, *Phys. D* **45**, 229 (1990).

[131] T. T. Foley, M. S. Shell, and W. G. Noid, *J. Chem. Phys.* **143**, 243104 (2015).

[132] H. Furukawa, K. E. Cordova, M. O'Keeffe, and O. M. Yaghi, **341** (2013).

[133] J. Kärger and D. M. Ruthven, *Diffusion in zeolites and other microporous solids*, Wiley, 1992.

[134] M. W. Deem, R. Pophale, P. A. Cheeseman, and D. J. Earl, *J. Phys. Chem. C* **113**, 21353 (2009).

[135] O. Toktarbaiuly, V. Usov, C. Ó Coileáin, K. Siewierska, S. Krasnikov, E. Norton, S. I. Bozhko, V. N. Semenov, A. N. Chaika, B. E. Murphy, O. Lübben, F. Krzyżewski, M. A. Załuska-Kotur, A. Krasteva, H. Popova, V. Tonchev, and I. V. Shvets, *Phys. Rev. B* **97**, 035436 (2018).

[136] F. G. Pazzona, P. Demontis, and G. B. Suffritti, *J. Chem. Phys.* **131**, 234703 (2009).

[137] F. G. Pazzona, P. Demontis, and G. B. Suffritti, *J. Chem. Phys.* **131**, 234704 (2009).

[138] T. Ala-Nissila, R. Ferrando, and S. C. Ying, *Adv. Phys.* **51**, 949 (2002).

[139] J. Kärger, D. Ruthven, and D. N. Theodorou, *Diffusion in Nanoporous Materials*, Wiley, 2012.

[140] S. C. Ying, I. Vattulainen, J. Merikoski, T. Hjelt, and T. Ala-Nissila, *Phys. Rev. B* **58**, 2170 (1998).

[141] D. A. Reed and G. Ehrlich, *Surf. Sci.* **102**, 588 (1981).

[142] E. Beerdsen, B. Smit, and D. Dubbeldam, *Phys. Rev. Lett.* **93**, 248301 (2004).

[143] `https://github.com/numat/RASPA2/blob/master/structures/zeolites/cif/ITQ-29.cif`.

[144] S. Plimpton, *J. Comp. Phys.* **117**, 1 (1995).

[145] C. Matano, *Jap. J. Phys.* **8**, 109 (1933).

[146] M. A. Załuska-Kotur, S. Krukowski, Z. Romanowski, and L. A. Turski, *Surf. Sci.* **457**, 357 (2000).

[147] M. A. Załuska-Kotur, A. Łusakowski, S. Krukowski, Z. Romanowski, and L. A. Turski, *Vacuum* **63**, 127 (2001).

[148] P. Demontis and G. B. Suffritti, *J. Phys. Chem. B* **101**, 5789 (1997).

[149] E. Beerdsen, D. Dubbeldam, and B. Smit, *Phys. Rev. Lett.* **96**, 044501 (2006).

[150] E. Beerdsen, D. Dubbeldam, and B. Smit, *J. Phys. Chem. B* **110**, 22754 (2006).

[151] R. Gomer, *Rep. Prog. Phys.* **53**, 917 (1990).

[152] W. Janke, Statistical Analysis of Simulations: Data Correlations and Error Estimation, in *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, volume 10 of *NIC*, 2002.

[153] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov, *J. Chem. Phys.* **96**, 1776 (1992).

[154] P. Attard, *J. Chem. Phys.* **98**, 2225 (1993).

[155] A. P. Lyubartsev, A. Laaksonen, and P. N. Vorontsov-Velyaminov, *Mol. Phys.* **82**, 455 (1994).

[156] R. D. Kaminsky, *J. Chem. Phys.* **101**, 4986 (1994).

[157] N. B. Wilding and M. Müller, *J. Chem. Phys.* **101**, 4324 (1994).

[158] F. A. Escobedo and J. J. de Pablo, *J. Chem. Phys.* **103**, 2703 (1995).

[159] A. A. Khare and G. C. Rutledge, *J. Chem. Phys.* **110**, 3063 (1999).