INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

CONDENSED MATTER THEORY SECTOR

Molecular crystal structure prediction with evolutionary algorithm

Thesis submmitted for the degree of Doctor Philosophiæ

Academic Year 2014/2015

CANDIDATE

Cong Huy Pham

SUPERVISOR

Prof. Stefano de Gironcoli

October 2015

SISSA - Via Bonomea 265, 34136 Trieste - ITALY

Contents

Introduction

Molecular polymorphism, the observation of different crystal structures made up of the same molecules, has been a central problem standing in the way of affordable and reliable crystal structure prediction which would greatly accelerate the development of new materials for applications in solid state chemistry, material science and pharmaceutical science [1]. In summary, the key challenges for *ab initio* crystal structure prediction of molecular crystals include **i)** the computational cost of thermodynamical exploration of a rich polymorphic phase space; **ii)** the accuracy needed to resolve the similarly-low energies among polymorphs [2]; and **iii)** the fact that the crystallization procedure is controlled by kinetic factors rather than thermodynamic ones [3].

The last decade has witnessed these challenges being tackled by the scientific community and the progress can be followed through the blind tests organized yearly by the Cambridge Crystallographic Data Centre [4, 5, 6, 7, 8]. The exponential growth in the hardware performance and new, efficient algorithms tailored for molecular crystals have allowed a wider region of the phase space to be explored. The increased computational performance also enabled a transition from empirical interatomic potentials to more accurate but time consuming quantum mechanical techniques, mainly Density Functional Theory (DFT). However, this transition did not guarantee an increase in the predictive power in all cases [9]: the standard DFT functionals do not describe properly van der Waals (vdW) interactions. This fact forces crystal structure prediction studies to employ the approximate semi-empirical corrections. These approximations to the vdW interactions strongly affect the energy ordering of the explored structures, which is a core information in predicting polymorphism. Hence, to render crystal structure prediction reliable, a fully *ab initio* method that is able to obtain an accurate lattice energy including the vdW interactions has been highly desirable.

Recently a breakthrough in the description of vdW interactions in DFT has been made: many new non-local functionals that accurately describe the dispersion

interactions have been proposed and demonstrated unprecedented success in a wide range of systems from molecules, molecular crystals to layered materials, with a computational cost comparable to that of standard functionals [10, 11]. Even in difficult cases such as glycine crystals, where polymorphs show energy differences as little as 1 kcal/mol, new non-local functionals can yield the correct stability ordering as well as the accurate pressure evolution [12].

Encouraged by these results we combine this critical progress in DFT with recent developments in evolutionary crystal structure prediction [13], specifically adapted for molecular structure search [9], and perform a fully *ab initio* crystal structure prediction search on glycine crystals, without semi empirical corrections in the energy description, using neither information on cell geometry nor the symmetry of the experimentally observed polymorphs. Thus it gives us the right to assess whether state-of-the-art *ab initio* crystal structure prediction can pass the challenging blind test of exploring the phase space of polymorphic glycine. In this study, we obtain, without prior empirical input, all known phases of glycine, as well as the structure of the previously unresolved ζ phase after a decade of its experimental observation [14]. The search for the well-established α phase instead reveals the remaining challenges in exploring a polymorphic landscape. We also propose several low-energy structures of glycine, some of which are more stable than the experimental structures.

The study of crystal structure prediction for cholesterol is motivated by a medical application: a solid state nuclear magnetic resonance experiment reports that different pathologies of human gallbladder result in cholesterol gallstones with different polymorphs [15]. These polymorphs show distinct nuclear magnetic resonance (NMR) spectra and a phase which associates to gallbladder cancer, has not been structurally determined yet. Important information on the growth of gallstones associated to different diseases can be revealed if the crystal structure of the unknown phase is identified. In this work, we use evolutionary algorithm in the case of molecular crystal in the USPEX code [9] combined with a classical force-field. This force-field was designed specifically for cholesterol and its success in determining the crystal structure of cholesterol polymorphs has been reported [16]. Later in our study we discuss the validation of the classical force field as a primary selection

method. After using accurate DFT calculations within non-local van der Waals functional the lowest energy predicted structure is identified with the experimental one based on their good agreements for crystal structure parameters as well as NMR spectra. We also propose a few low-energy structures of cholesterol and characterize them by their NMR spectra.

The layout of the thesis is as follow: In Chapter 1, we present the theoretical background of DFT, Projector-Augmented-Wave (PAW) and Gauge-Including Projector-Augmented-Wave (GIPAW) methods. In Chapter 2, we introduce the crystal structure prediction problem and present evolutionary algorithms as one solution to perform crystal structure search for molecular crystals. Chapter 3 and Chapter 4 are dedicated to the detailed results when using evolutionary algorithm in crystal structure search for the studies of glycine and cholesterol respectively.

Chapter 1

Ab-initio Electronic Structure Calculation

From the theoretical point of view, in order to understand the properties of any system, all one needs to know is its wavefunction. It does not represent any physical quantity but at any given time, its square modulus is interpreted as the probability density of finding a particle at a given point in space. If the wavefunction is known, other observable quantities are easily calculated and the system is well-understood. This wavefunction is the solution of the well-known Schrödinger equation with the general form of the Hamiltonian

$$H = -\sum_i \frac{\hbar^2}{2m}\nabla_i^2 - \sum_\alpha \frac{\hbar^2}{2M_\alpha}\nabla_\alpha^2 + \frac{1}{2}\sum_{i\neq j}\frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} + \frac{1}{2}\sum_{\alpha\neq\beta}\frac{Z_\alpha Z_\beta e^2}{|\mathbf{R}_\alpha - \mathbf{R}_\beta|} - \sum_{i,\alpha}\frac{Z_\alpha e^2}{|\mathbf{r}_i - \mathbf{R}_\alpha|}, \tag{1.1}$$

where $\{\mathbf{r}_i\}$ and $\{\mathbf{R}_\alpha\}$ are the Cartesian coordinates of electrons and nuclei. The first (second) term is the kinetic energy of the electrons (nuclei) while the third and fourth terms are the Coulomb interactions between electrons and nuclei, respectively. The final term corresponds to the electron-nucleus attraction.

In general, because of the large number of independent variables (growing with the number of electrons and nuclei), the exact solution to this equation can not be found. Fortunately, since electrons and nuclei are very different, one can treat them separately in the Born-Oppenheimer (or adiabatic) approximation [17, 18]. The physical basis of this approximation comes from the fact that electron mass is much smaller than the nuclear one. Therefore the electronic time scale is much shorter than the nuclear one and a good approximation can be obtained by keeping the nuclei fixed and determining the ground state of electrons as a function of

static nuclear positions. Within this approximation, one just needs to solve the Schrödinger equation for the electrons with the Hamiltonian that reduces to the three terms

$$H = -\sum_i \frac{\hbar^2}{2m_i} \bigtriangledown_i^2 + \frac{1}{2}\sum_{i \neq j} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_{i,\alpha} \frac{Z_\alpha e^2}{|\mathbf{r}_i - \mathbf{R}_\alpha|}. \tag{1.2}$$

However, even with this simplification, the problem for electrons remains too complicated and still is a many-body problem. Solving it is the main task of computational electronic structure.

In 1964, a breakthrough was made by Hohenberg and Kohn [19] in which the problem is simplified using density functional theory (DFT). The main advantage of DFT with respect to wavefunction methods is the use of electronic charge density as a fundamental variable, therefore greatly reducing the number of variables in the calculation.

1.1 Density Functional Theory

Hohenberg and Kohn have proven two theorems that are the basis for DFT [18, 19]. The first theorem states, "For any system of interacting particles in an external potential $V_{\text{ext}}(\mathbf{r})$, the potential is determined uniquely, except for a constant, by the ground state particle density $n_0(\mathbf{r})$." Since the ground state density determines the external potential $V_{\text{ext}}(\mathbf{r})$ and so the Hamiltonian for a given system, it follows that the wavefunctions of all states (ground-states and excited-ones) are determined. Thus in general, all properties of the system are completely determined given only the ground state density. Mathematically, any property of a many-body interacting particles can be viewed as a functional of the ground state density. The total energy now reads

$$E[n(\mathbf{r})] = T[n(\mathbf{r})] + U_{ee}[n(\mathbf{r})] + U_{\text{ext}}[n(\mathbf{r})] = F[n(\mathbf{r})] + \int V_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r}, \tag{1.3}$$

where $U_{\text{ext}}[n(\mathbf{r})] = \int V_{\text{ext}}(\mathbf{r}) n(\mathbf{r}) d\mathbf{r}$ is the interaction energy of the electrons with an external field $V_{\text{ext}}(\mathbf{r})$, $T[n(\mathbf{r})]$ and $U_{ee}[n(\mathbf{r})]$ are the kinetic energy and the electron-electron interaction energy respectively. Since $F[n(\mathbf{r})] = T[n(\mathbf{r})] + U_{ee}[n(\mathbf{r})]$ requires

no explicit knowledge of $V_{\text{ext}}(\mathbf{r})$, it is an *universal functional* of the density (only in the meaning that it does not depend on the external field). The second theorem states that: the ground state energy is a functional of $n_0(\mathbf{r})$ and satisfies a variational principle, i.e $n_0(\mathbf{r})$ minimizes the total energy $E[n(\mathbf{r})]$. The consequences of the Hohenberg-Kohn theorems lead to a considerable conceptual advance. If the universal functional $F[n]$ is known, one can find the ground state electronic density for any system. Unfortunately those theorems give no information about this functional.

In 1965, Kohn and Sham developed a good approximation for the functional $F[n]$ by introducing an auxiliary non-interacting electron system which has the same ground state density as the interacting one [18, 20]. With this assumption, the interacting electron problem is mapped onto an equivalent non-interacting one which is easier to solve. In this way, the functional $F[n]$ can be written as

$$F[n] = T_s[n] + \frac{1}{2}\int \frac{n(\mathbf{r})n(\mathbf{r'})}{|\mathbf{r}-\mathbf{r'}|}d\mathbf{r}d\mathbf{r'} + E_{xc}[n], \tag{1.4}$$

where the first term $T_s[n]$ is the independent particle kinetic energy, the second one is the classical Coulomb interaction energy of the electron density $n(\mathbf{r})$ (the Hartree energy). The last one is the *exchange-correlation energy* E_{xc} that can be written as

$$E_{xc}[n] = T[n] - T_s[n] + U_{ee}[n] - \frac{1}{2}\int \frac{n(\mathbf{r})n(\mathbf{r'})}{|\mathbf{r}-\mathbf{r'}|}d\mathbf{r}d\mathbf{r'}. \tag{1.5}$$

In this form, one can see that all many-body effects are included in E_{xc}. Remember that one needs to require the integration of the density giving the correct number of electrons $\int n(\mathbf{r})d\mathbf{r} = N$. Minimizing the energy functional $E[n]$, given in eq. (1.3), with this constraint is equivalent to solve a set of self-consistent Kohn-Sham equations

$$\begin{aligned}
\left\{-\frac{1}{2}\bigtriangledown^2 + V_{KS}(r) - \varepsilon_i\right\}\psi_i(\mathbf{r}) &= 0, \\
V_{KS}(\mathbf{r}) &= V_{\text{ext}}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}(\mathbf{r}), \\
V_H(\mathbf{r}) &= \int \frac{n(\mathbf{r'})}{|\mathbf{r}-\mathbf{r'}|}d\mathbf{r'}, \qquad V_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[n(\mathbf{r})]}{\delta n(\mathbf{r})}, \\
n(\mathbf{r}) &= \sum_{i\in occ} |\psi_i(\mathbf{r})|^2.
\end{aligned} \tag{1.6}$$

This is a nonlinear self-consistent system of equations because the Kohn-Sham potential $V_{KS}(\mathbf{r})$ depends on the solutions $\{\psi_i\}$. If the functional $E_{xc}[n]$ is explicitly defined, the Kohn-Sham equations can be solved self-consistently by numerical methods.

So far DFT has been presented as a formal mathematical framework for viewing electronic structure from the perspective of the electron density. The major problem is that the exact functionals for exchange and correlation are unknown, but approximations can be introduced to permit quite accurate calculations. These approximations have to reflect the physics of electronic structure and do not come from the mathematical properties of DFT.

The simplest approximation for the exchange-correlation functional is the Local Density Approximation (LDA) which has a quasilocal form

$$E_{xc}^{\mathrm{LDA}}[n(\mathbf{r})] = \int \varepsilon_{xc}(\mathbf{r}; [n(\tilde{\mathbf{r}})]) n(\mathbf{r}) d\mathbf{r}, \tag{1.7}$$

where $\varepsilon_{xc}(\mathbf{r}; [n(\tilde{\mathbf{r}})])$ represents the exchange-correlation energy per particle at point $\mathbf{r}$. The idea of LDA is that for regions where the charge density varies slowly, the exchange correlation energy can be considered the same as that of a locally uniform electron gas of the same charge density. In practical use of LDA, the exchange term takes the simple analytic form for the homogeneous electron gas while the correlation energy is parameterized from accurate Quantum Monte Carlo data [21]. A typical limitation of LDA is that it underestimates ionisation energies but overestimates the binding energies. LDA also gives inaccurate descriptions for strongly correlated systems.

The Generalized Gradient Approximations (GGAs) are semilocal and take into account the gradient of the density at the same point

$$E_{xc}^{\mathrm{GGA}}[n(\mathbf{r}), \nabla n(\mathbf{r})] = \int \varepsilon_{xc}(n(\mathbf{r}), \nabla n(\mathbf{r})) n(\mathbf{r}) d\mathbf{r}, \tag{1.8}$$

which in many cases improve on LDA, especially for cohesive and dissociation energies. The first GGA functional was introduced by Perdew and Wang in 1992 [22]. Many different forms of GGA functionals were proposed and the most widely use till now is named PBE from Perdew, Burke and Ernzerhof [23].

It is well known that the lack of dispersion force is one of the biggest challenges of standard approximate DFT. The van der Waals interaction energy depends on atoms separation R as $E_{\mathrm{vdW}} \propto R^{-6}$ [24] and neither LDA nor GGA can describe this behavior correctly. In LDA, long-range interactions are described by an exponential decay whereas GGA makes corrections to the local approximation of LDA but is still unable to describe the van der Waals interaction behavior.

In recent years, many approaches have been proposed to deal with van der Waals interactions. Calculations using exact-exchange (EXX) and random-phase approximation (RPA)-correlation energies within the adiabatic connection fluctuation-dissipation theorem formalism have demonstrated to describe correctly the long-range van der Waals interaction [25, 26, 27]. Unfortunately, its high computational cost limits its practical use. A simple idea to account for van der Waals interactions is to add an empirical dispersion-energy term to the total DFT energy. Several methods using this idea have been introduced, for example DFT-D2 [28], DFT-D3 [29] of Grimme *et al.*, and vdW-TS of Tkatchenko and Scheffler [30]. These methods need some predetermined input parameters to calculate the van der Waals interaction. Their limitation is that the complex many-body interactions are treated outside the DFT framework and therefore the ground state wavefunction and ground state density do not contain non-local correlation effects.

In other approaches, termed non-local correlation functionals, the dispersion interaction is obtained directly from the electron density. In these approaches, a non-local correlation term

$$E_{\mathrm{c}}^{\mathrm{nl}} = \int\int d\mathbf{r_1} d\mathbf{r_2} n(\mathbf{r_1}) n(\mathbf{r_2}) \phi(n(\mathbf{r_1}), n(\mathbf{r_2}), |\mathbf{r_1} - \mathbf{r_2}|), \tag{1.9}$$

is included in the correlation functionals [10]. In eq. (1.9), $n(\mathbf{r})$ is the electron density and $\phi(n(\mathbf{r_1}), n(\mathbf{r_2}), |\mathbf{r_1} - \mathbf{r_2}|)$ is the integration kernel with a $1/|\mathbf{r_1} - \mathbf{r_2}|^6$ asymptotic behavior. The computational cost of the original formulation is quite demanding, especially in a plane-wave approach. Only after Román-Pérez and Soler [31] proposed an interpolation scheme to improve the scaling, these non-local functionals are widely used. Several non-local functionals were suggested and promising results have been obtained. In the original vdW-DF functional, revPBE functional

is applied for the exchange part [32]. Some other functionals are the same as vdW-DF except that other exchange functionals are employed instead of the revPBE, for example optB88-vdW and optB86b-vdW [33], c09-vdW [34], cx-vdW [35]. A second version of vdW-DF functional, named vdW-DF2 [36], with modified exchange and non-local correlation was suggested to improve the descriptions of the binding energy and equilibrium spacings between weakly-bound complexes. The VV10 functional [11] proposes a different functional form for the kernel $\phi(n(\mathbf{r_1}), n(\mathbf{r_2}), |\mathbf{r_1} - \mathbf{r_2}|)$ and can depict accurate interaction energies of van der Waals system, not only near the minima but also far from equilibrium. It also gives accurate covalent bond lengths and atomization energies. Its revised version, known as rVV10 [37], allows the use of the Román-Pérez and Soler scheme and therefore makes the computation less expensive while keeping the outstanding precision of the original VV10. Recently, it was found that changing the b parameter of the rVV10 functional (from its original value 6.3 to 9.3) gives a better description of structural properties in first-principles molecular dynamics of liquid water [38].

In this thesis, we use the vdW-DF functional in studies of glycine and cholesterol. In the case of glycine polymorphs at ambient condition, the performances of rVV10 and rVV10-b9.3 functionals are also given for comparison.

1.2 GIPAW Method

The main concern of this thesis is crystal structure prediction. Moreover, in the study of cholesterol, we also characterize the predicted structures by their NMR spectra using the GIPAW method. In this section, we introduce very briefly the Projector-Augmented-Wave (PAW) and Gauge-Including Projector-Augmented-Wave (GIPAW) methods. The details of these methods can be found in the original papers [39, 40, 41].

1.2.1 Projector-Augmented-Wave

One problem of DFT is the behavior of the Kohn-Sham one electron wavefunction. This wavefunction is fairly smooth in the bonding regions far from the nuclei whereas

close to the nuclei, because of the large electron-nuclear Coulomb attraction, it has rapid oscillations. For solving Kohn-Sham equation in a plane-wave basis set, this behavior requires a very large set to describe the wavefunction accurately.

The use of pseudopotentials is one way to overcome this problem [42]. Since the shape of the wavefunction in the vicinity of the nuclei does not affect much the electronic properties of the material, the rapidly oscillating wavefunction can be replaced by a smoother function. However, because of the way the pseudopotential is constructed, the information about the full wavefunction close to the nuclei is lost. As a result, the pseudopotential method cannot describe correctly properties that depend directly on the core electrons (such as X-ray photoelectron spectroscopy) or the electron density near the nuclei (such as NMR shielding and coupling constant).

In another approach, namely augmented plane-wave (APW) method [43, 44], one also divides the space into two regions. Inside the atom-centered augmentation spheres, the wavefunctions are expanded in atom-like partial waves. In the bonding region outside these spheres, some smooth envelope functions are used. At the boundary of the spheres, the partial waves and envelope functions are matched.

The projector-augmented wave (PAW) method [39] is a generalization of the pseudopotential and APW ones. Its idea is that all-electron Koln-Sham single particle wavefunctions $|\psi\rangle$ can be built from the fictitious pseudo wavefunctions $|\tilde{\psi}\rangle$ through a linear transformation operator $\mathcal{T}$. Since $|\tilde{\psi}\rangle$ and $|\psi\rangle$ differ only in the regions close to the nuclei, $\mathcal{T}$ can be written as $\mathcal{T} = 1 + \sum_{\mathbf{R}} \hat{\mathcal{T}}_{\mathbf{R}}$, where $\hat{\mathcal{T}}_{\mathbf{R}}$ is non-zero only within some spherical augmentation region Ω_R enclosing the atoms. The transformation operator $\mathcal{T}$ can be defined as

$$\mathcal{T} = 1 + \sum_{\mathbf{R}n} \left(|\phi_{\mathbf{R}n}\rangle - |\tilde{\phi}_{\mathbf{R}n}\rangle \right) \langle p_{\mathbf{R}n}|, \tag{1.10}$$

where n refers to atomic-state quantum numbers; $|\phi_{\mathbf{R}n}\rangle$ and $|\tilde{\phi}_{\mathbf{R}n}\rangle$ are the all-electron partial waves and pseudo partial waves, respectively; $|p_{\mathbf{R}n}\rangle$ are projector functions (which are localized in the region Ω_R) satisfying the orthonormal conditions: $\langle p_{\mathbf{R}n}|\tilde{\phi}_{\mathbf{R}'m}\rangle = \delta_{\mathbf{R}\mathbf{R}'}\delta_{mn}$. In this representation, all-electron operators are transformed into new operators, named "pseudo-operators", $\tilde{A} = \mathcal{T}^{\dagger}\hat{A}\mathcal{T}$. For quasilo-

cal operators, the pseudo-operators can be written as

$$\tilde{A} = \hat{A} + \sum_{\mathbf{R},n,m} |p_{\mathbf{R}n}\rangle \left(\langle \phi_{\mathbf{R}n} | \hat{A} | \phi_{\mathbf{R}m} \rangle - \langle \tilde{\phi}_{\mathbf{R}n} | \hat{A} | \tilde{\phi}_{\mathbf{R}m} \rangle \right) \langle p_{\mathbf{R}m} |. \tag{1.11}$$

The advantage of PAW method is that all-electron observable quantities can be calculated using the pseudo wavefunctions without the need to explicitly store all-electron wavefunctions in memory. This is important for the calculation of properties that directly depend on the core electrons or the electron density near the nuclei such as NMR.

1.2.2 Gauge-Including Projector-Augmented-Wave

It is known that nuclei that have spin one-half (for example ^{1}H, ^{13}C ...) can have two possible spin states (which are degenerate) with magnetic quantum numbers $m_S = \pm 1/2$. If the nucleus is placed in an external magnetic field $\mathbf{B}$, the interaction between the nuclear magnetic moment and the external field determines an energy difference between the two states $\Delta E = \gamma \hbar |\mathbf{B}|$, where γ is the gyromagnetic ratio that depends on the nuclear mass and charge. Since the two states no longer have the same energy, transition between spin states can be induced. This phenomenon is known as Larmor precession with the so-called Larmor frequency being $\omega_0 = \Delta E/\hbar = \gamma |\mathbf{B}|$.

Consequently, one can expect that all nuclei of the same nuclear mass (which have the same γ) would have very similar NMR frequency since the frequency depends only on γ and the magnetic field $\mathbf{B}$. However, this is not the case. In fact, the core and valence electrons also interact with the magnetic field and contribute a "shielding" effect to the NMR frequency. Therefore one observes different frequencies for different nuclei depending on their chemical environment. This shift in the NMR frequency is called NMR chemical shift.

The chemical shielding tensor is determined as

$$\mathbf{B}_{\text{in}} = -\overrightarrow{\sigma}(\mathbf{r})\mathbf{B}, \tag{1.12}$$

where $\mathbf{B}_{\text{in}}$ is the induced magnetic field. This magnetic field is generated by the electronic currents that are induced in the material by the external field. $\mathbf{B}_{\text{in}}$ can

be calculated as

$$\mathbf{B}_{\text{in}} = \frac{1}{c}\int d\mathbf{r}'\ \mathbf{j}^{(1)}(\mathbf{r}') \times \frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}, \tag{1.13}$$

where $\mathbf{j}^{(1)}$ is the first-order induced electric current. From this formula, it is clear that the calculation of the induced magnetic field requires an accurate description of the induced current close to the nucleus. Therefore the PAW formalism is needed to reconstruct the wavefunction behavior close to the nuclei.

It is known that, in an uniform magnetic field, a field-dependent phase factor is created in any wavefunction when translating the whole system by a vector $\mathbf{t}$

$$\langle \mathbf{r}|\psi_n'\rangle = \exp[(i/2c)\mathbf{r}.\mathbf{t}\times\mathbf{B}]\langle \mathbf{r}-\mathbf{t}|\psi_n\rangle. \tag{1.14}$$

A difficulty of the PAW method, when applied to systems in magnetic field, is that a huge number of projectors would be needed to describe this wavefunction correctly. The use of a field-dependent transformation operator

$$\mathcal{T}_{\mathbf{B}} = 1 + \sum_{\mathbf{R}n} \exp[(i/2c)\mathbf{r}.\mathbf{R}\times\mathbf{B}]\ \left(|\phi_{\mathbf{R}n}\rangle - |\tilde{\phi}_{\mathbf{R}n}\rangle\right)\langle p_{\mathbf{R}n}|\ \exp[-(i/2c)\mathbf{r}.\mathbf{R}\times\mathbf{B}] \tag{1.15}$$

can solve this problem. This transformation operator defines a new method which is called gauge-including projector-augmented-wave (GIPAW).

The details of GIPAW method can be found in the original papers [40, 41] and can be schematized as follows

$$|\psi^{(0)}\rangle \rightarrow |\psi^{(1)}\rangle \rightarrow \mathbf{j}^{(1)}(\mathbf{r}') \rightarrow \mathbf{B}_{\text{in}}. \tag{1.16}$$

First one determines the eigenstate $|\psi^{(0)}\rangle$ of the unperturbed Hamiltonian in the absence of the magnetic field then calculates its linear variation $|\psi^{(1)}\rangle$. The first-order induced electric current can be written as the sum of three terms, a bare contribution, a paramagnetic operator and a diamagnetic one. Next step is to obtain the induced magnetic field $\mathbf{B}_{\text{in}}$ using eq. (1.13) and the shielding tensor $\overrightarrow{\sigma}$ arcording to eq. (1.12). Finally the isotropic chemical shielding is given by its trace: $\sigma_{\text{iso}} = \text{Tr}[\overrightarrow{\sigma}]/3$.

Chapter 2

Molecular crystal structure prediction with evolutionary algorithms

---

In this chapter, we will briefly introduce the crystal structure prediction problem and describe the challenges in crystal structure prediction of molecular crystals. We then present evolutionary algorithms as one solution to perform crystal structure search for molecular crystals. The detailed applications of evolutionary algorithm and their results in the search for low-energy structures of glycine and cholesterol are given in Chapter 3 and Chapter 4 respectively.

2.1 Introduction to atomic crystal structure prediction

Crystal structure is one of the most important information about a system since it determines material properties. Different arrangements of even the same atoms can result in completely different properties of the material. For example, graphite and diamond are both made of carbon atoms but with different structures. In graphite, the atoms stay in layers with a hexagonal lattice while diamond has a tetrahedral form. In term of properties, graphite is a dark, soft material and is a good electrical conductor while diamond is transparent, the hardest material known up to now, and is an insulator. Thanks to the development of computational simulations, one can now determine a huge number of interesting properties of materials, assuming that their crystal structure is known. But if the crystal structure is unknown, one can gain very little information about the material.

Crystal structure is usually determined using experimental data from single-crystal X-ray Diffraction (XRD) and/or NMR spectra. Unfortunately, in many cases, the quality of the experimental data is poor, for example at some experimental conditions of high pressures and/or high temperatures; and the positions of hydrogen atoms are not measurable in standard XRD experiments. In these cases, crystal structure prediction has a leading role in the determination of the crystal structure of the material. Crystal structure prediction is also a good approach to investigate materials at extreme conditions (e.g ultrahigh pressure) that cannot or are difficult to be studied with today's experimental techniques. In searching for new materials, computational crystal structure prediction is usually much easier and cheaper than experiments.

There is no doubt that crystal structure prediction is extremely challenging and it was even stated that crystal structures are unpredictable [45, 46]. The main task in crystal structure prediction is to find the global minimum in a multi-dimensional space. For a structure with N atoms in the unit-cell, the dimensionality of the space is $3N + 3$ ($3N - 3$ for the atomic coordinates plus 6 for the unit-cell parameters). This space is huge and noisy with the number of possible structures increasing exponentially with the number of atoms in the unit-cell. Until now there is no method that can guarantee to find the global minimum successfully.

Several methods have been proposed to overcome the difficulty of searching a complex energy-landscape: data mining [47], metadynamic [48, 49], basin hopping [50], simulated annealing [51, 52] minima hopping [53] and evolutionary algorithms [13]. All of them work based on the assumption that one needs to explore only the most promising regions in a huge energy-landscape. The idea of data mining is simple: one has databases that contain a huge number of known crystal structures; in order to find the structure of a new material, one can search in the databases the structures which have chemical composition similar to the one of the system that one wants to study. Then ionic substitutions are used and the structures are optimized with the expectation that the lowest energy structure found will be the ground state one. In metadynamic, basin hopping, simulated annealing and minima hopping methods, one starts from an initial structure which can be chosen in a

"good" region of the energy-landscape. and then modifies it by some perturbations in the cell parameters and/or atom coordinates. The optimization is performed with the hope that the structure will move to a new local minima. The scheme of evolutionary algorithms will be presented in the Section 2.4.

The so-called random sampling method in principle can explore many regions of the energy-landscape. The idea of this method is that candidate structures are created randomly and local optimization are performed for these structures. The search is then terminated when the lowest energy structure has been found several times. This method has been used for predicting crystal structure of some simple systems and the results were very promising. However this method can work efficiently for small systems up to around 12 atoms only [54].

2.2 Introduction to molecular crystal structure prediction

Prediction of the crystal structure at the molecular level is essential since molecular crystal can be used in many fields such as pharmaceuticals, optoelectronic materials, pigments, explosives, molecular electronics and metal-organic frameworks [3, 55]. In general, different polymorphic forms of a certain compound have significantly different properties. Therefore in order to understand the properties of molecular crystals for industrial applications, knowing the molecular crystal structures and the condition at which they are stable is compulsory. To obtain the phase diagram of a molecular crystal, it requires the calculation of finite-temperature thermodynamic contributions and the understanding of the competition between thermodynamics and kinetics in packing determination. These issues are beyond the aim of this thesis. In this study, we only focus on finding the most thermodynamically stable molecular arrangements at 0 K.

The traditional way to do crystal structure prediction for molecular crystal is to first rank the energies of all generated structures using classical force-field. Those structures are created randomly from a given molecular conformations. The number of generated structures may be as huge as 10^7 and the force-field rules out a

significant number of the worse structures. Then accurate DFT calculations are used to estimate the stability of all the short-listed candidate structures. There are several global optimization algorithms that can be extended from simple atomic crystal structure prediction to search for molecular crystal structure. Promising results have been obtained by several studies using, for example metadynamic [56, 57], minima hopping [58, 59] and evolutionary algorithms [9, 60, 56].

2.3 Challenges in crystal structure prediction of molecular crystal

When doing crystal structure prediction for molecular crystals, an important aspect must be noticed. In general, the molecular crystal of certain molecule is thermodynamically less stable than the individual simple molecules that can be obtained from the chemical decomposition of the molecule of interest. For example, glycine ($C_2H_5NO_2$) polymorphs maybe less stable than the combinations of H_2O, CO_2, CH_4, NH_3, H_2, N_2, NO_2... Therefore in order to search for the crystal structure of molecular crystals, it is a good idea to fix the intramolecular connectivity of the molecule. It is not a limitation but in fact an advantage that reduces the search space and the number of degrees of freedom in the system.

Even when the whole molecule is used as the unit, the crystal structure prediction of molecular crystals remains difficult. The key challenges for *ab initio* crystal structure prediction of molecular crystals can be summarized as: **i)** the search space is huge; **ii)** addressing the relative stability of polymorphs is difficult.

2.3.1 The search space is huge

In molecular crystals, each molecule in the unit-cell is characterized by the coordinates of the molecule center and the angles of its orientation. In principle, for a certain number of a given molecule, there is an infinite number of possible structures that can be formed. Among them, only the ones with lowest energies (or enthalpies if non-zero pressure is considered) are likely to be found experimentally. Predicting the lowest energy structures in such a huge energy-landscape is a big challenge.

For flexible molecules, there are usually more than one inequivalent molecules in the unit-cell. In this case, the number of degrees of freedom becomes even larger; and it is a good idea to use several types of molecules to describe possible different conformations [61, 62]. We will do so in the study of cholesterol molecule which has a flexible hydrocarbon tail.

The development of crystal structure prediction methods is aimed at making the search in the energy-landscape more efficient. Although several searching algorithms have been proposed with encouraging results [58, 59, 9, 60], currently there is no single method that can prove to address the general global minimum search successfully. For the lowest energy structure found by any algorithms, there is no way to know if it really is the ground state structure or just a local minima. What one can do is to compare the lowest energy structures with the available experimental data.

2.3.2 Addressing the relative stability of polymorphs is difficult

All structures created by crystal structure prediction methods need to be optimized to estimate their energies (or enthalpy if at non-zero pressure). The lowest energy structure found is assumed to be the ground state one. There is no doubt that predicting the right energy ordering is a prerequisite to find the global minimum successfully.

In comparison with the atomic crystal, molecular crystal structure prediction is more challenging because of the accuracy needed to resolve similarly low energies among polymorphs. In molecular crystals the molecules are held together by van der Waals interactions and the energy differences between structures are small. In some cases these energy differences are even smaller than the typical numerical error of the simulation. Therefore a good treatment of long-range dispersion interactions is deeply needed.

Many crystal structure prediction studies for molecular crystal use classical force-field as the local optimization due to its favorable computational cost compared with DFT calculations in quantum chemistry. Unfortunately, empirical potential energy-landscape often gives unphysical minima [63, 64]. Inaccurate force-field parameterization can lead to significant distortion of the molecule and wrong energy ranking

of the candidate structures. Flexible molecules not only add more complexity in the energy-landscape but also make the accuracy of classical force-field become more uncertain. It is known that there is usually considerable noise that may give rise to many local minima in the hypersurface of the force-field. Structures that are optimized with this force-field may be trapped in these local minima [61].

2.4 Evolutionary algorithm in the case of molecular crystal

Evolutionary algorithm is one of the methods that are well suited for structure prediction. It requires very little information as input. For molecular crystal structure prediction, only the molecular geometry and the number of molecules in the unit-cell are needed. The evolutionary algorithm works based on the idea of natural competition and the survival of the fittest. Its scheme, as implemented in the USPEX code for molecular crystal, is sketched in Fig.2.1(**a**) and summarized in the following

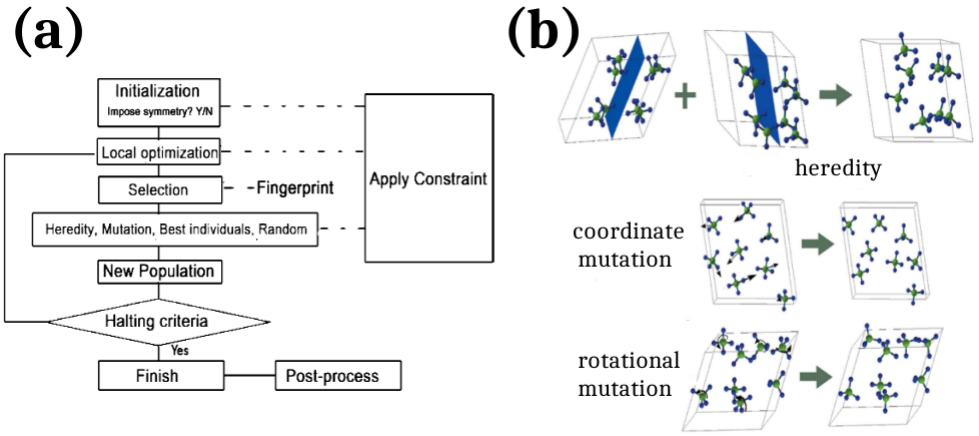

Figure 2.1: **(a)** *The scheme of evolutionary algorithm for predicting crystal structure of molecular crystal with the sequence of the steps that are performed;* **(b)** *sketch of how heredity, coordinate mutation and rotational mutation operators create a child structure: heredity combines slices of two parent structures in a new structure; coordinate mutation moves some molecules in a random direction; rotational mutation rotates the molecules by a random angle. This picture is reproduced from Ref. [9].*

2.4.1 Initialization

First, the code starts with some initial structures that are usually created randomly from an unit which is the whole molecule, instead of individual atoms as in the case of atomic crystal. The diversity of this generation guarantees the success of the search. In the USPEX code, instead of putting molecules at completely random positions, the structures are built from a space group that is chosen randomly. Experience shows that by doing so, the first generation is proved to get higher diversity [9].

For large systems with a huge volume and a huge number of molecules in the unit-cell, the randomly generated structures are very similar to a disordered system and the "unit-cell splitting" technique [65] could be used. The large cell is split into subcells which contain smaller number of molecules. In the subcell, the molecules are placed randomly. This technique was shown to produce more diverse structures for large systems [65].

Some constraints can be applied at this initial step. One may need to set a minimum distance between molecules to avoid molecules overlap. If any experimental information (cell parameters or crystal structure space group symmetry) is known, it can be used to make the search more efficient. If any structure is already known (from the studied compound or other related materials) it can also be included as template in the USPEX code (the Seed technique).

The use of space group information can be of great help in crystal structure prediction [66]. From the experimental organic crystal structure database, known crystals mostly belong to $P2_1/c$, $P\bar{1}$, $P2_12_12_1$, $C2/c$, $P2_1$ and $Pbca$ space groups. In the study of glycine, we will show that the use of this information allows to explore the energy-landscape more efficiently.

2.4.2 Local optimization

At any generation, the created structures need to be optimized. The optimization process not only gives an estimate of the energy of the candidate structures in order to find the lowest ones but also reduces the dimensionality of the energy-landscape. In some cases, structures that look very different at the beginning become the same

after relaxing with differences just due to numerical noise.

Using accurate DFT in the local optimization would be a very desirable option. The advantage of *ab initio* calculations over the empirical force-field ones is that *ab initio* can give accurate results in all regions of the energy-landscape while the results of force-field are good only near the regions where force-field parameters were fitted. Therefore *ab initio* calculations can be used in the study of new materials where no experimental data are known. However insufficiently dense k-point sampling and/or bad pseudopotentials can produce inaccurate minima [54] and since molecules interact with each others by van der Waals interactions, the functional used must include this long-range interaction. In the study of glycine, we use DFT calculation in the local optimization and the vdW-DF functional [10, 32], which predicts correct energy ordering for simple organic crystals [12]. The k-point sampling and pseudopotentials are carefully checked.

When the number of atoms in the unit-cell is large, the use of DFT in the local optimization step is too demanding and the use of classical force-field is needed. In this case a good force-field is required. Great improvements of the force-field have been made so far. The all-atom force-field is usually parameterized from DFT calculations in which dispersion interaction is included. Excellent agreements with experiments and/or quantum mechanical calculations have been achieved, even for flexible molecules [61, 67, 62]. Indeed if the force-field can be designed correctly, they are much appropriated for crystal structure prediction [4, 5, 6, 7, 8]. In the study of cholesterol, we will use a classical force-field in the local optimization, whose successful determination of the crystal structure of cholesterol polymorphs has been reported [16]. Good candidate structures are then relaxed better using accurate DFT calculations with non-local van der Waals functional. One thing must be noticed as mentioned in Ref. [61]. Even a good force-field may still have a noisy energy hypersurface, therefore with a good force-field, the best strategy is to optimize the structure three times: first step is done with the force-field in the crystal structure prediction search; the obtained structures are then relaxed with an accurate DFT method; and final one is relaxed with the force-field again. If the latter optimization with force-field produces a lower energy than the previous one,

the structure has escaped a local minima. We will also do this check.

2.4.3 Selection

It can happen that in one generation, one structure is found many times and the same parent structures are likely to create very similar children. Therefore if a repeated structure is included when choosing parents structures for the next generation, this may result in a waste of resource in the repeated optimization of the same structure. This duplication also reduces the diversity of the generation and makes it difficult for the algorithm to escape from local minima.

It is non-trivial to identify equivalent structures since there are many ways to represent a structure and numerical noise makes the problem more complicated. In principle one can not distinguish structures from their energy and volume alone [68]. In the USPEX code, the structures are classified by comparing their fingerprints. This function is related to the radial distribution function and diffraction spectrum. Details of the fingerprint function in USPEX are given in Ref. [69, 70]. There is a technical issue when calculating the fingerprint for molecular crystals. The intramolecular contributions, which are the same for all molecules with different orientation, are neglected [9]. The molecule is then treated as a single object with the coordinates of the molecular center.

Based on the desired property, different selection rules can be applied: if one looks for the most stable structure then all the structures are ranked by their energy. The worst structures with highest energy are discarded and the lowest ones are likely to be chosen as parents to create the new generation.

2.4.4 Variation operators

When the parent structures are selected, the children ones are created by applying some variation operators. Some operators are shown in Fig2.1(**b**). Typically it is a good idea to keep some of the best structures from the previous generation to make sure that the generations don't go worse. These best structures are required to have significantly different fingerprints. For coordinate mutation operator, the

molecules are moved in random directions and the rotational mutation operator chooses to rotate a certain number of molecules by a random angle. More details about mutation operators can be found in the original paper [9]. The diversity of the population is guaranteed by adding new random structures at each generation. Heredity operator is an essential part of the evolutionary algorithm. Without it, evolutionary algorithm becomes very similar to other random search methods. For other operators, the child structure is created from one single parent while for heredity, two structures act as parents. The heredity operator cuts planar slices from the two parents structures and combine them together to generate a child structure. One thing must be noticed in this case: if two good parents are very different the child structure can be very bad [71]. The reason is that the child structure falls into the high barrier region between two good local minima. In order to avoid this, in USPEX, the parents for heredity operators must have fingerprints that are not too different.

2.4.5 Convergence conditions

At each generation, the conditions for the convergence predefined by the user are checked. One may want to stop the calculation when it reaches a specified maximum number of generations, or when after a given number of consecutive generations, no better structures are found. The later condition is inspired by the fact that at a certain generation, if no better structure is found, the parents would be the same as in the previous one and therefore the generation has more chance to be repeated.

2.4.6 Further notes

In the evolutionary algorithm, several parameters can have an impact on the effectiveness of the search. One has to deal with incompatible requests: the diversity of the population and the convergence to the optimal solution. With higher diversity, one can explore the energy-landscape better however the chances to search in the most promising regions are lower. On the other hand, reducing diversity helps to investigate a given region but increases the risk of missing the global optimum. In

general, one seeks for a balance between diversity and convergence. Detailed study about this aspect can be found in Ref. [72].

There are advantages and disadvantages in evolutionary algorithms over other crystal structure prediction methods. The power of evolutionary algorithms is that they don't require previous system knowledge (except its chemical composition), which is good for the prediction of new materials. In each generation, the best structures are selected to "procreate" and, through new generations, increasingly better structures are found. However, in evolutionary algorithms, no information on mechanisms of phase formation is given. This information can be found from other methods, for example metadynamic [48, 49], basin hopping [50], simulated annealing [51, 52] and minima hopping [53]. While phase transition mechanisms can be simulated with such methods, real mechanisms present in experiments can be much more complicated and difficult to treat. Comparison of evolutionary algorithm with metadynamic and minima hopping can be found in Ref. [56] and Ref. [73, 74] respectively.

Chapter 3

Molecular crystal structure prediction of glycine

In this chapter, we tackle the challenge of crystal structure prediction in the difficult case of the repeatedly studied, abundantly used aminoacid glycine that hosts still little-known phase transitions [14, 75, 76] and illustrate the current state of the field through this example. First, we introduce glycine polymorphs and describe the method used in this study. By using clustering technique, the result of evolutionary algorithm is analyzed. We identify all experimentally known structures of glycine as well as the hitherto unresolved ζ phase. The challenge in exploring a polymorphic landscape is exposed in the case of α-glycine. We suggest a simple way to explore the energy-landscape more efficiently. Finally, we propose several new low-energy structures of glycine.

3.1 Introduction

Glycine, $H_3N^+CH_2COO^-$ as shown in Fig. 3.1**(a)**, the smallest aminoacid, is an excellent test case for organic crystal structure prediction studies as its already rich polymorphism under ambient conditions is amplified and becomes less understood at higher pressure (see Fig. 3.1 **(b)**). At ambient conditions, three polymorphs of glycine are known: α, β, and γ. The form readily obtained by evaporation of aqueous solutions is α-glycine, which for long was believed to be the most stable phase instead of the later discovered ground-state phase γ. However, the energy difference between the two phases is pretty small, about 0.27 kJ/mol per molecule [77]. The γ-form appears if the crystallization is performed in acidified solutions [78]

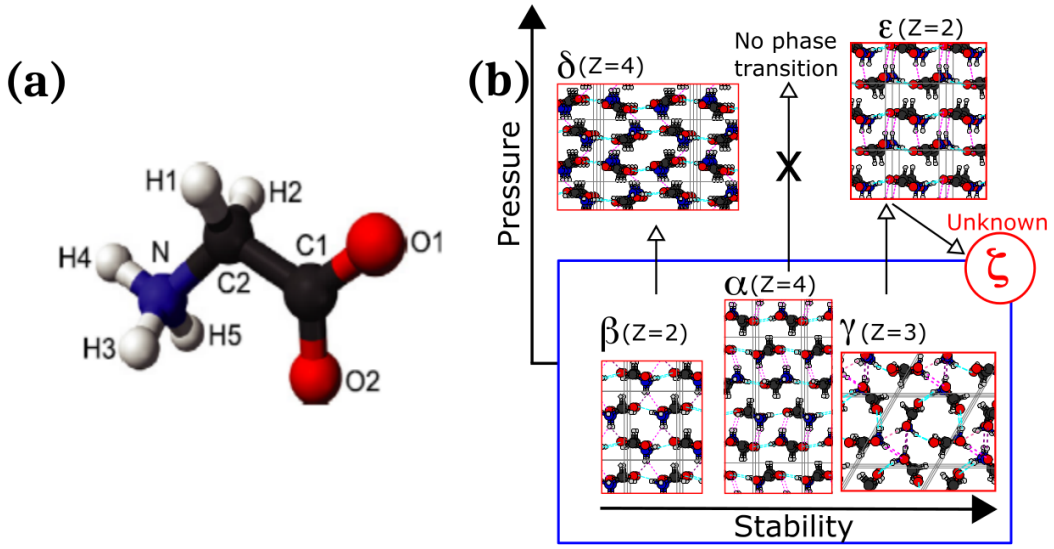

Figure 3.1: ***(a)*** *Glycine molecule with numbering of the atoms;* ***(b)*** *Glycine polymorphism under pressure. Stability order of polymorphs at ambient pressure , α, β, γ with indicated Z molecules in unit cell, is given. When the pressure increases, while phases β and γ transform into phases δ and ε respectively, α-glycine is still stable up to at least 23 GPa. The ζ-glycine is formed on decompression of phase ε at 0.62 GPa. However its crystal structure has not been resolved yet.*

or by using compounds that prevent the growth of α-glycine [79]. β-glycine can be obtained in the crystallization from aqueous solution with slow diffusion of ethanol [80]. Nevertheless, it is unstable and usually transforms spontaneously into either α or γ polymorphs [81].

Pressure evolution of ambient pressure phases shows that while γ and β phases quickly lose single crystal nature or undergo a phase transition within a few GPa, α phase stays stable up to 23 GPa, the highest pressure reached in experiments. A reversible, hysteresis-free single-crystal to single-crystal transition occurs from β to δ phase at 0.76 GPa. Single crystals of the γ phase undergo instead an extended polymorphic transformation in the wide range from 1.9 GPa to 7.6 GPa, to a high-pressure polymorph, the ε phase, accompanied with the fragmentation of single crystals into powder. Upon decompression, the ε phase is stable down to 0.62 GPa,

where a new, irreversible phase transition occurs to the ζ phase, a new polymorph which is reported to be stable at ambient conditions for at least three days [14]. Interestingly, despite its stability, and at least three crystal structure prediction studies devoted to glycine so far [9, 60, 82], a decade after its observation, the ζ phase has not been structurally resolved yet. In the study of the low-temperature heat capacity of β-glycine, a second order phase transition was observed at 252 K [75]. However, the crystal structure of this new β'' phase has not been reported so far and its existence has even been dismissed in subsequent work of the same group. In Ref. [76], glycine phases were studied when aqueous solutions were frozen, then the subsequent heating under different conditions resulted in an unknown X-glycine at 209-216 K. As temperature increases, this phase quickly transforms into the β-polymorph. The crystal structure of X-glycine is also not fully determined, the experimental and calculated diffraction patterns do not entirely agree. Motivated by the unknown structures, we perform the crystal structure prediction for glycine to search for possible low-energy structures and illustrate the current state of the field through this example.

3.2 Method

3.2.1 Evolutionary search

The complex polymorphism of glycine highlights the importance of performing an extensive search in phase space, while practical concerns limit any crystal structure prediction study to explore primarily the lowest energy structures. In this study, we use evolutionary algorithms as implemented in the USPEX package to address this interplay efficiently [9]. We perform three test suits with Z=2, 3 or 4 glycine molecules in the crystal unit cell. At the first generation, 30 structures are created randomly. After energy ordering, the 20% of the population that is energetically least favorable is discarded. Among the remaining structures, a fingerprint analysis is performed and potential parents whose fingerprint is within a threshold distance of 0.01 from any lower energy structure are also discarded. The so-determined unique structures are eligible as parents and are allowed to procreate. The structures of the

new generation are created from parents through the following operations: heredity (cross-over of two structures) (40%), softmutation (translation and rotation based on an estimate of soft vibrational modes) (20%) and rotation of the molecule (20%). The diversity of the population is guaranteed by addition of new random structures (20%) at each generation, while the three best parents are directly cloned to the next generation. The highest computational cost in this workflow is due to the *ab initio* geometry optimization of each structure considered. To keep this cost well within the capacity of modern high-performance computing technologies and within the budget of academic as well as industrial research, we limit the evolution to 20 generations at most.

3.2.2 *ab initio* calculations

For every structure generated by USPEX, the geometry and cell relaxation is performed using vdW-DF functional [10, 32] which was implemented in the QUANTUM ESPRESSO package [83]. A kinetic energy cutoff of 80 Ryd and a charge density cutoff of 560 Ryd are used. The Brillouin zone sampling resolution was gradually increased in three steps during relaxation: resolution of $2\pi \times 0.12$ Å^{-1}, $2\pi \times 0.10$ Å^{-1} and $2\pi \times 0.08$ Å^{-1} respectively. Energies and geometries of the last step with the densest k-point are used throughout the study. PAW pseudopotentials are taken from the PSLibrary project [84]. By using this setup all structures are fully relaxed within a convergence of less than 0.1 mRy for absolute total energy, 0.5 mRy/a.u. for the forces on atoms and less than 0.005 GPa for the stress tensor.

3.2.3 Cluster analysis

A cluster analysis of the structures generated during the crystal structure prediction runs is performed by using single linkage clustering, where two structures with fingerprint distance less than a distance threshold, d, are considered to belong to the same cluster. Since USPEX definition of fingerprint does not include any information on the enthalpy of the structure, a constraint is added such that two structures with enthalpy difference more than 0.5 kJ/mol are not allowed to form a cluster.

This constraint is found necessary only when the clustering analysis is performed for all the encountered structures, while when limiting the analysis to the low enthalpy region, such constraint was not necessary as each cluster was successfully identified with the distance only.

3.3 Results

3.3.1 Performances of several van der Waals functionals

First, we show the results of several van der Waals functionals to the structural and energetic properties of the three known polymorphs of glycine that are stable at ambient pressure. Four functionals are chosen in this study: the vdW-DF functional with revPBE for exchange part [32]; the rVV10 functional [37] which is the revised version of the VV10 functional [11] with the optimized parameter $b = 6.3$; the rVV10 functional with a modified parameter $b = 9.3$ as suggested in [38]; and the semi-empirical PBE+D one [28]. We use the notation rVV10-b6.3 and rVV10-b9.3 to indicate the rVV10 functional with the parameter b set to 6.3 and 9.3, respectively.

Table 3.1 shows the optimized lattice parameters at zero pressure for α, β and γ phases with the four van der Waals functionals. The experimental lattice parameters of these structures, are also given. The PBE+D and rVV10-b6.3 give the best descriptions of the cell parameters; changing parameter b to 9.3 makes the rVV10-b9.3 give slightly worse cell parameters than rVV10-b6.3; while vdW-DF too overestimates the cell parameters.

Next, we study the performance of these functionals in reproducing the structure of the single molecules in the crystal. In Table 3.2, we compare the bond lengths and torsion angle for α-glycine at ambient pressure. Changing parameter b from 6.3 to 9.3 of rVV10 functional doesn't make any effect in terms of bond lengths. The rVV10-b6.3, rVV10-b9.3 and PBE+D give very good values for bond lengths while vdW-DF again overestimates the bond lengths, especially the C1-C2 and C2-N bonds. In terms of torsion angles, PBE+D and rVV10-b6.3 give the best comparisons with experiment while vdW-DF and rVV10-b9.3 overestimate these angles.

In term of energetic properties, we compare the relative energies of glycine poly-

Table 3.1 Optimized cell parameters and the unit cell volume per molecule V_0 (Å^3) at zero pressure for glycine phases that are stable at ambient pressure. The results are shown for different van der Waals functionals and the experimental data are also given.

	a (Å)	b (Å)	c (Å)	β (deg)	V_0
α-glycine					
Exp.[85](77K)	5.069	11.801	5.448	111.7	75.70
vdW-DF	5.240	12.283	5.565	111.1	83.73
rVV10-b6.3	5.113	11.695	5.500	111.3	76.68
rVV10-b9.3	5.144	11.856	5.510	110.7	78.54
PBE+D	5.053	11.778	5.465	112.6	75.10
β-glycine					
Exp.[85](77K)	5.077	6.145	5.374	113.2	77.05
vdW-DF	5.212	6.409	5.500	112.4	84.74
rVV10-b6.3	5.091	6.119	5.437	113.3	76.68
rVV10-b9.3	5.117	6.210	5.439	112.6	79.77
PBE+D	5.039	6.107	5.406	113.8	76.11
γ-glycine					
Exp.[85](77K)	6.985	6.985	5.483	90.0	76.87
vdW-DF	7.233	7.232	5.587	90.0	84.34
rVV10-b6.3	6.985	6.986	5.523	90.0	77.82
rVV10-b9.3	7.057	7.057	5.540	90.0	79.65
PBE+D	6.919	6.919	5.491	90.0	75.87

Table 3.2 Optimized bond lengths (Å) and torsion angle (degree) for α-glycine at zero pressure. The results are shown for different van der Waals functionals and the experimental data are also given.

Bonds	Exp.[85]	vdW-DF	rVV10-*b*6.3	rVV10-*b*9.3	PBE+D
C1-O1	1.25	1.27	1.27	1.27	1.26
C1-O2	1.25	1.27	1.27	1.27	1.27
C2-N	1.48	1.51	1.49	1.49	1.48
C1-C2	1.52	1.55	1.53	1.53	1.53
C2-H1	1.05	1.09	1.10	1.10	1.10
C2-H2	1.04	1.09	1.09	1.09	1.10
N-H3	1.03	1.04	1.05	1.05	1.05
N-H4	1.09	1.06	1.06	1.06	1.06
N-H5	1.09	1.03	1.04	1.04	1.04
Torsion					
N-C2-C1-O1	18.6	23.1	22.5	23.2	22.5

Table 3.3 Relative energies (in kcal/mol per molecule) for the glycine polymorphs at zero pressure. The results are shown for different van der Waals functionals and the experimental data are also given.

Functionals	α	β	γ
Exp. [77]	0.064	0.142	0
vdW-DF	0.016	0.264	0
rVV10-*b*6.3	-0.107	0.295	0
rVV10-*b*9.3	0.013	0.231	0
PBE+D	0.223	0.758	0

morphs at ambient pressure and the pressure values for the β-δ and γ-ε phase transitions. As seen in Table 3.3, rVV10-*b*6.3 predicts a wrong energy ordering between α and γ phases; PBE+D gives right stability ordering, but the energy differences are too much overestimated; vdW-DF and rVV10-*b*9.3 are the ones which agree very well with experiment. Table 3.4 shows the pressure values at which the two known

phase transitions from β to δ and from γ to ε occur experimentally and those predicted by van der Waals functionals. While rVV10-b6.3 (PBE+D) underestimates (overestimates) these pressure values; vdW-DF and rVV10-b9.3 agree very well with experiment.

Among the functionals used in this study, we show that the ones that give a better description of the structural properties, give worse results for the energetic properties and vice versa. These results show that significant room for improvement of the van der Waals functionals still remains. For crystal structure prediction, the energy ordering is the most important aspect, therefore we choose to use the vdW-DF functional.

Table 3.4 The pressure phase transitions (in GPa) from β-δ and γ-ε of glycine polymorphs. The results are shown for different van der Waals functionals, and the experimental data are also given.

Phase transition	Exp.	vdW-DF	rVV10-b6.3	rVV10-b9.3	PBE+D
β-δ	0.76 [86]	1.5	0.15	1.1	1.25
γ-ε	1.90 [87]	1.8	1.00	1.8	2.54

3.3.2 Results of evolutionary algorithm

The results of crystal structure prediction can be visualized through the distribution of energy as a function of volume for the structures encountered during the search. Despite the exploration of a wide region in phase space (see left panel of Fig.3.2), about 40 % of all the structures lies within 4 kJ/mol of the experimentally known ground state structure, γ. Focusing on this region of the energy-landscape as shown in the right panels of Fig.3.2, we see structures forming islands with varying sizes and shapes. This feature illustrates the added complication in the case of molecular crystal structure prediction with respect to standard inorganic solids where well-defined, isolated minima would be observed for each phase. The shape and finite size of the islands can be understood considering that Glycine is very soft, therefore structures that are far off from the equilibrium lattice parameters are thermodynamically penalized only slightly as demonstrated in the inset of Fig.3.2.

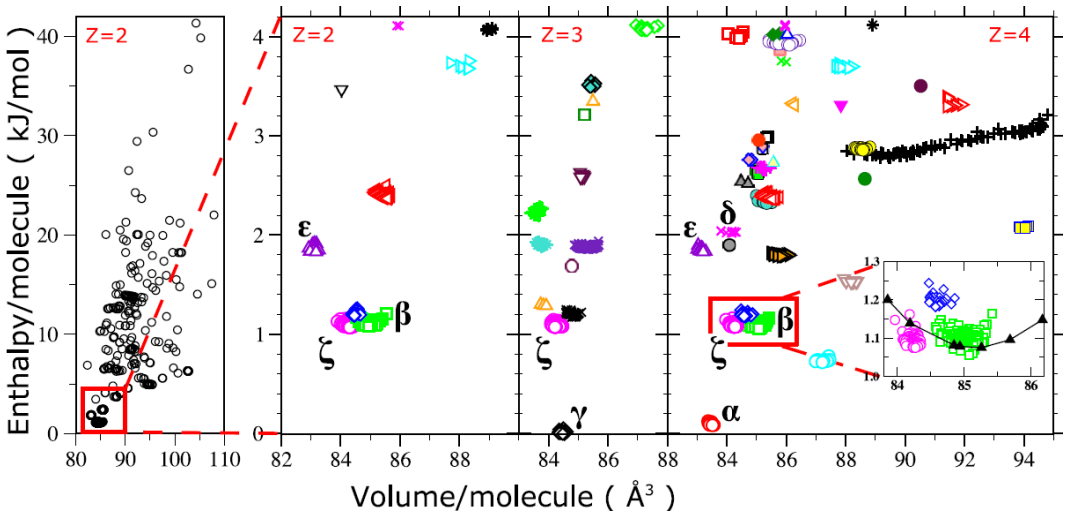

Figure 3.2: ***Results of ab initio crystal structure search for Glycine with cluster analysis.*** *Left panel: Enthalpy vs volume distribution of all encountered structures for 2 molecules per cell shows that CSP with evolutionary algorithm allows a wide energy range to be explored while "survival of the fittest" algorithm keeps the focus on the thermodynamically low lying structures. Right panels: Expanded view of all explored structures compatible with 2, 3 and 4 molecules per cell in the lowest 4 kJ/mol range. All known phases of glycine are identified with the right energy ordering along with a number of low-lying alternative polymorphs, including our prediction for the hitherto unresolved ζ phase. As shown in the inset of the Z=4 panel, crowding around each polymorph, when compared with its equation of state, is compatible with numerical noise due to incomplete relaxation. The fingerprint-and-energy based clustering techniques adopted here are however well suited to separate and identify the different low lying polymorphs even in presence of noise.*

This effect, combined with the numerical noise in geometry optimization, as well as an increased number of degrees of freedom in molecular crystals, is enough to give rise to crowding around each polymorphic minimum. Nevertheless islands are well separated and a clear assignment of polymorphs can be made for most of them. This is in stark contradiction with a very recent crystal structure prediction study

for glycine with empirical corrections for intermolecular interactions, which reported that the obtained energy-volume points were not well separated enough to clearly identify each polymorph, thus underlining the challenge of polymorphism for organic crystal structure prediction [60]. In this study instead the separation between several islands are well represented down to very small energy differences (inset of Fig.3.2). We believe this stems from the leap in accuracy and precision reached by the use of fully *ab initio* energetics together with last generation evolutionary algorithm tools.

3.3.3 Clustering algorithm

Reliable energetics from *ab initio* calculations is necessary but not sufficient to guarantee a reliable structure classification in crystal structure prediction. More than one polymorph can be present within a given extended island; or what appears to be two adjacent islands due to insufficient sampling and/or relaxation, may actually correspond to the same packing order. Indeed the most human-time consuming part of a crystal structure prediction procedure is known to be the stage where the output structures are comparatively examined in order to successfully separate the essential data from the crowd of repetitions [1]. Although not utilized to their full extent within crystal structure prediction, concepts from data mining, mainly clustering techniques, can be of great help in this stage of the analysis, as we demonstrate in the following.

We used the clustering technique to identify the unique polymorphs among all the structures obtained with crystal structure prediction. In this clustering analysis, a bottom-up distance-based hierarchical clustering approach with single linkage is used. In distance-based approaches, a similarity metric is defined so that a distance can be measured between data points, and clusters are constructed based on proximity.

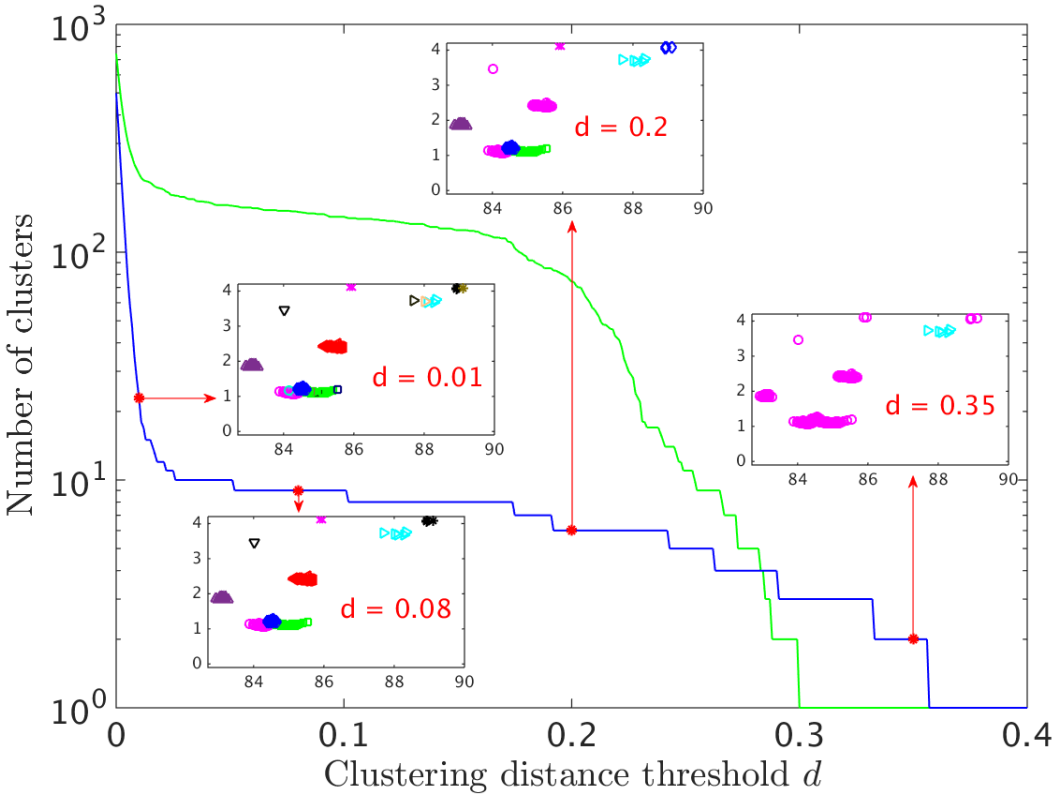

Figure 3.3: The number of clusters as a function of the distance threshold, d, *for all structures (green curve) and low-energy structures within approximately 4 kJ/mol above the ground state (blue curve) for the case of* $Z = 2$. *Insets show the enthalpy (kJ/mol) as a function of volume* (Å^3) *per molecule for different values of* $\text{d} = 0.01,\ 0.08,\ 0.2,\ 0.35$. *Different colors and point types in each inset correspond to different clusters. The value of* $\text{d} = 0.052 - 0.10$ *can distinguish different clusters successfully.*

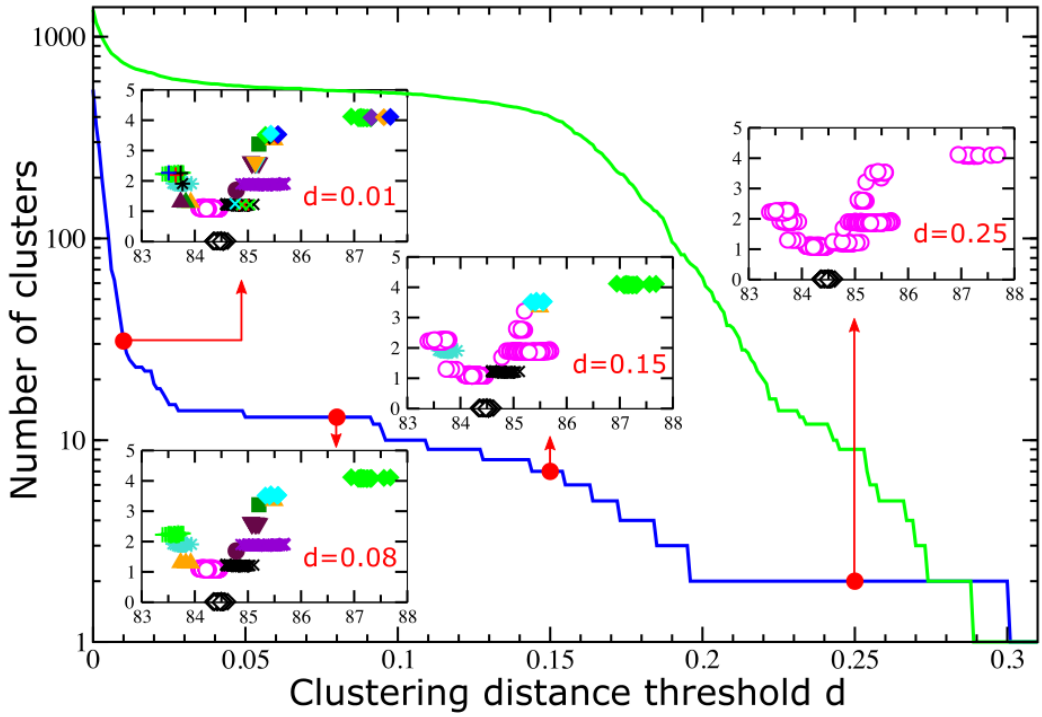

Figure 3.4: The number of clusters as a function of the distance threshold, d*, for all structures (green curve) and low-energy structures within approximately 4 kJ/mol above the ground state (blue curve) for the case of* $Z = 3$*. Insets show the enthalpy (kJ/mol) as a function of volume* (Å^3) *per molecule for different values of* $\mathrm{d} = 0.01,\ 0.08,\ 0.15,\ 0.25$*. Different colors and point types in each inset correspond to different clusters. The value of* $\mathrm{d} = 0.05 - 0.09$ *can distinguish different clusters successfully.*

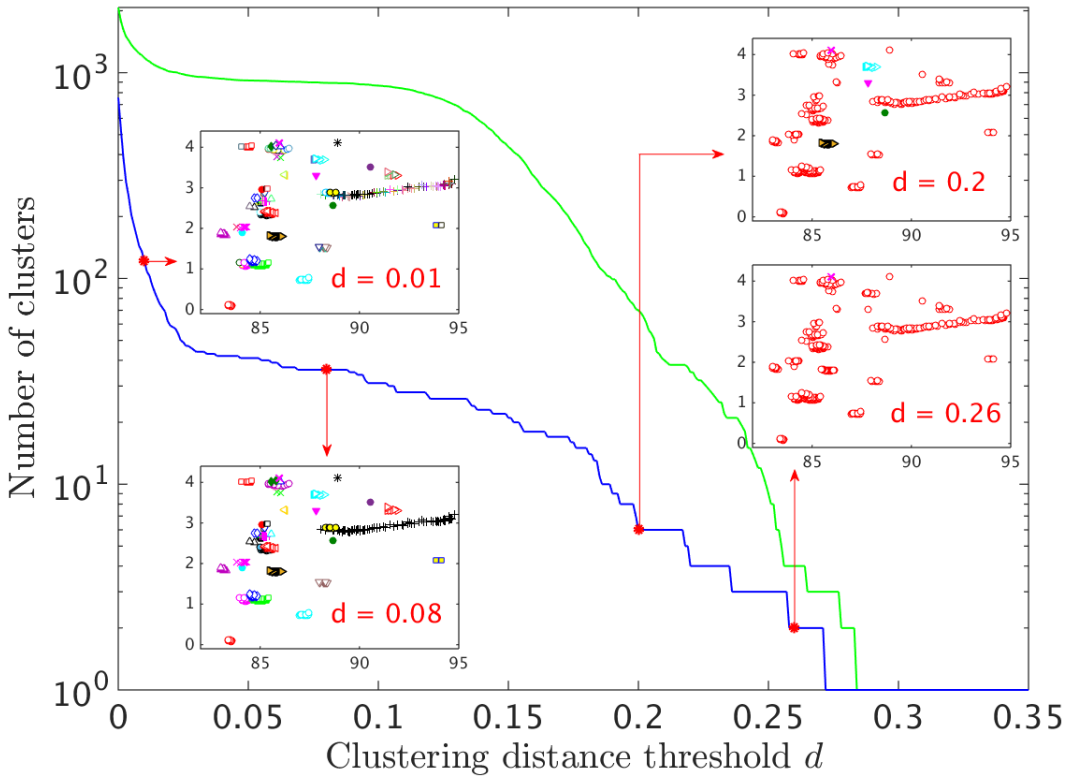

Figure 3.5: The number of clusters as a function of the distance threshold, d, *for all structures (green curve) and low-energy structures within approximately 4 kJ/mol above the ground state (blue curve) for the case of* $Z = 4$. *Insets show the enthalpy (kJ/mol) as a function of volume* (Å^3) *per molecule for different values of* $\mathrm{d} = 0.01,\ 0.08,\ 0.2,\ 0.26$. *Different colors and point types in each inset correspond to different clusters. The value of* $\mathrm{d} = 0.07 - 0.09$ *can distinguish different clusters successfully.*

In this study we use as the metric, the fingerprint-based cosine distance [9, 69, 70] defined in the evolutionary algorithm code USPEX [13]:

$$D_{\text{cosine}}(\mu,\nu) = \frac{1}{2}\left(1 - \frac{F_\mu * F_\nu}{|F_\mu|\,|F_\nu|}\right), \tag{3.1}$$

where individual structure fingerprints are defined as

$$F_{AB}(R) = \sum_{A_i,\text{cell}} \sum_{B_j} \frac{\delta(R - R_{ij})}{4\pi R_{ij}^2 \frac{N_A N_B}{V}\Delta} - 1, \tag{3.2}$$

where the double sum runs over all ith molecules of type A within the unit cell and all jth molecules of type B within a distance $R_{\max}$; $\delta(R-R_{ij})$ is a Gaussian-smeared delta function; R_{ij} is the distance measured from the centers of molecules i and j; V is the unit cell volume; the function $F_{AB}(R)$ is discretized over bins of width Δ; N_A and N_B are the numbers of molecules of type A and B, respectively.

In Figs. 3.3, 3.4 and 3.5, we display a step by step clustering analysis in the cases Z = 2, 3, 4 respectively. In these figures, the green (blue) curve is for the case in which the whole data set (the low-energy structures within 4.2 kJ/mol of the ground state) are considered. The distance threshold used to define whether two data points belong to the same cluster is then monotonically increased. As a result the cluster population evolves from the situation where every data point forms a distinct cluster to the situation in which all data points belong to the same global cluster, revealing the bottom-up and hierarchical nature of the approach. The most impressive feature of these figures is that the behavior of a rapid drop in the number of clusters is followed by a more or less constant plateau before the number of clusters eventually dies off. This behavior is quite general and applies to all cases of Z = 2, 3, 4 and when the whole data set or low-energy structures are considered. Translated to the crystal structure prediction problem, this data mining approach transforms the challenge of identification of unique polymorphs from the visual comparison of all structures into an easier decision on the value of the distance-threshold to be adopted. The optimal distance threshold is such that each data cluster matches a unique physical polymorph. In the case of glycine a distance threshold around 0.08 is found to be appropriate to identify the low energy polymorphs successfully. The so-determined optimal threshold can serve in advanced supervised learning

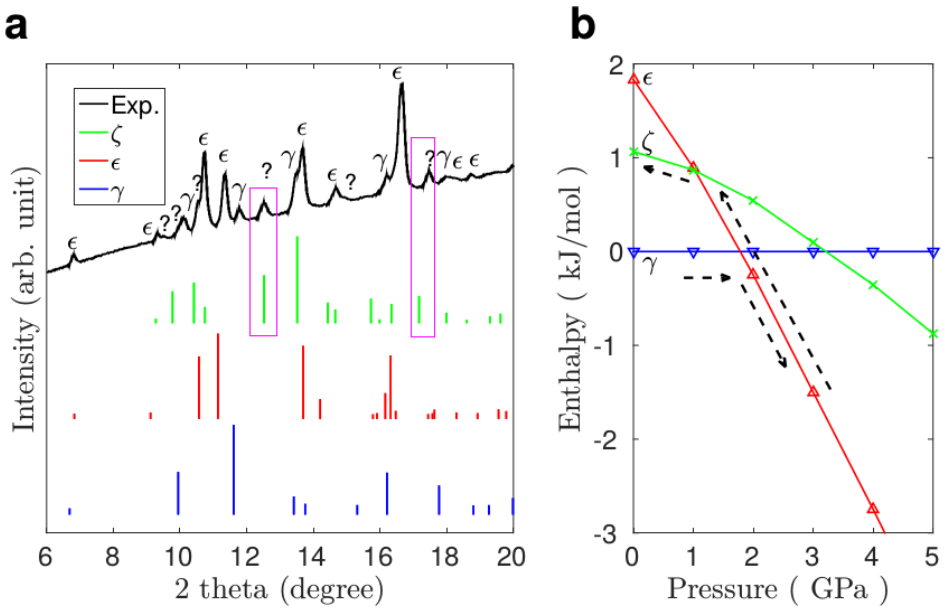

Figure 3.6: **a**. *Comparison of simulated x-ray diffraction patterns for ε-, γ- and ζ - glycine at 2 GPa with experimental data taken from [14] at 0.2 GPa. The XRD of proposed ζ-glycine can explain most of the unassigned peaks that were marked in the experimental spectrum, especially the ones highlighted in the purple boxes. The theoretical data is calculated at higher pressure to offset the overestimation of the ground state volume in ab initio calculations.* **b**. *Enthalpy per molecule as a function of pressure for ε-glycine and ζ-glycine with respect to the γ phase up to 5 GPa. The black arrows indicate the phase transitions observed in the experiment [14]: Under pressure, the γ phase undergoes a phase transition to ε-glycine. The decompression of ε-glycine instead results in the ζ phase.*

techniques and be fed back in the crystal structure prediction procedure to increase considerably the efficiency by reducing the generation of replicas of already explored structures.

3.3.4 Identification of ζ-glycine

The cluster analysis outlined above identifies all experimentally observed phases of glycine compatible with 2, 3 or 4 molecules per cell, as well as suggests oth-

ers, hereon named according to their enthalpy-per-molecule ordering. Phases 1 to 11 lie within approximately 2 kJ/mol of the experimentally most stable phase, γ. Among them one of the lowest energy polymorphs (phase 2) can be identified with ζ-glycine based on the excellent agreement with XRD results (Fig.3.6(a)) as well as its pressure evolution (Fig.3.6(b)). The structural identification of the ζ phase, previously experimentally observed but not resolved up to now, marks an important achievement for crystal structure prediction and is a key result of our study.

3.3.5 Challenges in exploring α-glycine

The search for α-glycine proved very demanding despite it being the experimentally most readily formed polymorph at ambient conditions. In this study the α phase could not be found even after 20 generations with the standard settings in USPEX. The detailed results of these searches are given in the Appendix in Figs. A.1, A.2, A.5 for Z = 2,3,4 respectively. We also performed two other simulations. One simulation allows to dynamically depopulate the enthalpy valley of the best parents found in the simulation while still keeping them in the list of eligible parents for the next generation. No experimental information is used in this simulation. The other simulation is done using the experimental parameters of α-glycine as defined by the non-standard space group $P2_1/n$ (equivalent to $P2_1/c$) [1]. Their results are shown in Fig. A.8 and A.9 respectively. These simulations fail to find the α phase. This difficulty revealed one of the remaining challenges of crystal structure prediction: the effective exploration of the topology of an erratic and vast configuration space.

Indeed if more system specific information is available it can be used to further constrain and guide the phase-space search: limiting the search to the experimentally known, $P2_1/c$, space group of α-glycine, or fixing the cell shape to its experimental value, resulted in its identification at the 15^{th} and 8^{th} generations, respectively. Combining the two constraints resulted in an even quicker discovery at the third

[1] In experiment, it has been noticed that a large lattice parameter β would give increased correlation between parameters with respect to the a and c axes, and therefore makes refinement becomes less stable [88]. Thus, in some cases, the non-standard space group $P2_1/n$ which has smaller value of β than the one of the $P2_1/c$, is used.

generation.

In the case of α-glycine, it is noteworthy that the crystal building block can be seen as a glycine dimer, with head to tail orientation. This feature is not seen in other ambient pressure polymorphs of glycine [89], and it can be speculated to be one of the reasons for the α phase not being readily connected with other phases in the energy-landscape. This correlates with the difficulty of generating the structure during the evolutionary algorithm procedure, as well as with its exceptional stability under pressure. Instead, if the dimer unit is taken as building block in a crystal structure prediction search, the α phase is found at the third iteration and new low-energy phases such as phase 8, phase 14, phase 24 and phase 38 are also discovered. The detailed result of the crystal structure search of glycine in the case of 4 molecules/cell using glycine dimer as a building block is given in Fig. A.7.

In the results present above, we showed that α-glycine can be found only if experimental information are used. To improve on this aspect we weighted the random selection of the space group of the candidate structures according to the frequency distribution appearing in known organic crystal structure database [$P2_1/c$ (36.59 %), $P\bar{1}$ (16.92 %), $P2_12_12_1$ (11.00 %), $C2/c$ (6.95 %), $P2_1$ (6.35 %) and $Pbca$ (4.24 %)]. Fig. A.6 shows the result of this simulation. This procedure successfully produced the α phase at the 14^{th} generation, demonstrating that incorporation of even mild and system unspecific experimental knowledge in the search strategy may have a significant impact to overcome the effectiveness challenge in the most demanding cases.

The difficulty of exploring the α phase as well as the finding of new phases only after a dimer unit is employed, underline the remaining challenges of crystal structure prediction and call for even more efficient methods for exploring new structures and innovative data analysis applications to guide the search for a full optimization of resources.

3.3.6 Notes on the γ-glycine

We also checked the procedure that uses information of space group symmetry from crystal structure database, in the case of 3 molecules/cell. Fig. A.3 shows the

results of this search. The most impressive feature in this case is that: structures with a much wider variety are found with respect to the simulation where all space groups were sampled with equal probability (see Fig. A.2). Again, it shows the effectiveness of the search using information learned from nature.

However in this case, the γ phase is not found within 20 generations. We performed another simulation with the same setting and γ-glycine was created randomly at the first generation. We notice that for the search with Z=3 using standard setting in USPEX (result shows in Fig. A.2), γ-glycine is also found by random generation. This can be interpreted as suggesting that in the energy-landscape, phase γ is also well separated from other phases (like phase α). In the case Z=3, the separation of γ-glycine from the other structures as shown in the inset of Fig. 3.4 can be seen as an evidence that supports this statement. Fortunately in this case the random generation operator can produce the γ-glycine.

3.3.7 New low-energy structures of glycine

Table 3.5 shows the list of low-enthalpy structures of glycine found in this study. Among them, phase-1 has a lower enthalpy than the β-glycine which exists at ambient pressure; the phase-2, which is identified with the ζ-glycine, also has an enthalpy very similar to β-glycine. The phases 3 to 8 have higher enthalpy than β-glycine, but lower than that of the high-pressure phase ε-glycine. Phases 9 to 11 have enthalpy in the region between the ε phase and δ phase. Other phases have higher enthalpy than δ-glycine. Several trends can be predicted based on the volume of the new structures.

Structures that are unstable at high pressure

The structures which have higher volume than the others, likely become unstable at high pressure. Fig. 3.7 **(a)** shows the crystal structure of a few large volume structures: phase-1, phase-6, phase-8, phase-12, and the enthalpies as functions of pressure for these phases are correspondingly shown as red curves in Fig. 3.7 **(b)** along with the enthalpies of $\alpha - \varepsilon$ phases. For phase-6, the unit-cell contains 4 molecules and can be seen as the combination of two subcells, each of which is the unit-cell of the β phase. However, they are combined to make an angle in space.

Table 3.5 Volume/molecule ($Å^3$) and enthalpy/molecule (kJ/mol) for all low-enthalpy structures found by USPEX at zero pressure. Enthalpy per molecule is calculated with respect to the γ phase. The number of molecules per unit-cell, Z, as identified by USPEX code is also given.

phases	Volume	Enthalpy	Z	phases	Volume	Enthalpy	Z
γ	84.477	0	3	phase-21	85.549	2.727	4
α	83.515	0.084	4	phase-22	84.722	2.759	4
phase-1	87.217	0.718	4	phase-23	89.497	2.778	4
β	85.083	1.055	2	phase-24	88.580	2.858	4
phase-2(ζ)	84.305	1.070	1	phase-25	85.198	2.870	4
phase-3	84.780	1.172	3	phase-26	85.068	2.956	4
phase-4	84.464	1.189	2	phase-27	85.206	3.214	3
phase-5	83.923	1.288	3	phase-28	91.409	3.296	4
phase-6	88.186	1.525	4	phase-29	86.274	3.309	4
phase-7	84.780	1.691	3	phase-30	87.840	3.311	4
phase-8	85.828	1.785	4	phase-31	85.496	3.348	3
ε	83.172	1.832	2	phase-32	84.023	3.465	2
phase-9	85.295	1.858	3	phase-33	85.408	3.489	3
phase-10	83.768	1.882	3	phase-34	90.527	3.504	4
phase-11	84.075	1.899	4	phase-35	88.294	3.677	2
δ	84.171	2.019	4	phase-36	86.005	3.745	4
phase-12	94.038	2.071	4	phase-37	85.786	3.870	4
phase-13	83.662	2.197	3	phase-38	86.106	3.915	4
phase-14	85.343	2.319	4	phase-39	84.416	3.985	4
phase-15	85.580	2.368	2	phase-40	85.552	4.020	4
phase-16	84.712	2.520	4	phase-41	86.031	4.022	4
phase-17	88.645	2.567	4	phase-42	88.915	4.062	2
phase-18	85.078	2.585	3	phase-43	87.309	4.064	3
phase-19	85.051	2.615	4	phase-44	85.975	4.109	2
phase-20	85.272	2.653	4				

There are two dimers in the crystal structure of phase-8. This phase has a two-layer structure, each layer is a chain that is made of dimers. Phase-1 and phase-12, instead, have a 3D hydrogen bond network in their crystal structure. As the pressure increases, these phases quickly become unstable. Phase-12 is even less stable than all experimental structures in the whole range of pressure from 0 to 20 GPa; while at high pressure, phase-1, phase-6 and phase-8 all have higher enthalpy than β-glycine.

Candidates for high pressures structures

The structures which have a small volume, become good candidates for high pressures phases. Fig. 3.8**(a)** shows the crystal structure of several such structures while their pressure evolution is shown in Fig. 3.8**(b)**. Phase-2, which is identified with the ζ-glycine, has only one molecule in the unit-cell. Phase-9 and phase-10 have single-layer structures connected by hydrogen bonds. Phase-3 has three layers in its structure, but two of them are nearly the same. At high pressure, the enthalpy of these phases are lower or comparable with the one of β-glycine. However until 20 GPa, none of them become more stable than any known high pressure phases δ or ε.

Fig. 3.9**(b)** shows the pressure evolutions of two structures, that are the most stable at high pressure among the new structures. Their crystal structures are shown in Fig. 3.9**(a)**. There are three layers in these structures. For phase-5, there is also a dimer in its structure. Phase-13 and δ-glycine have very similar enthalpy until 2 GPa and as the pressure increases, phase-13 becomes more stable. In this study, phase-5 has lower enthalpy than δ-glycine in the whole range of pressure from 0 to 20 GPa.

3.4 Conclusion

In conclusion, we presented a fully blind, fully *ab initio* crystal structure prediction (CSP) of glycine, a system that has been examined several times in the past yet never fully grasped. A remarkable precision and a broad sampling is obtained in an affordable computational time thanks to last generation van der Waals density functionals and evolutionary algorithms at the leading edge. The comparison of our

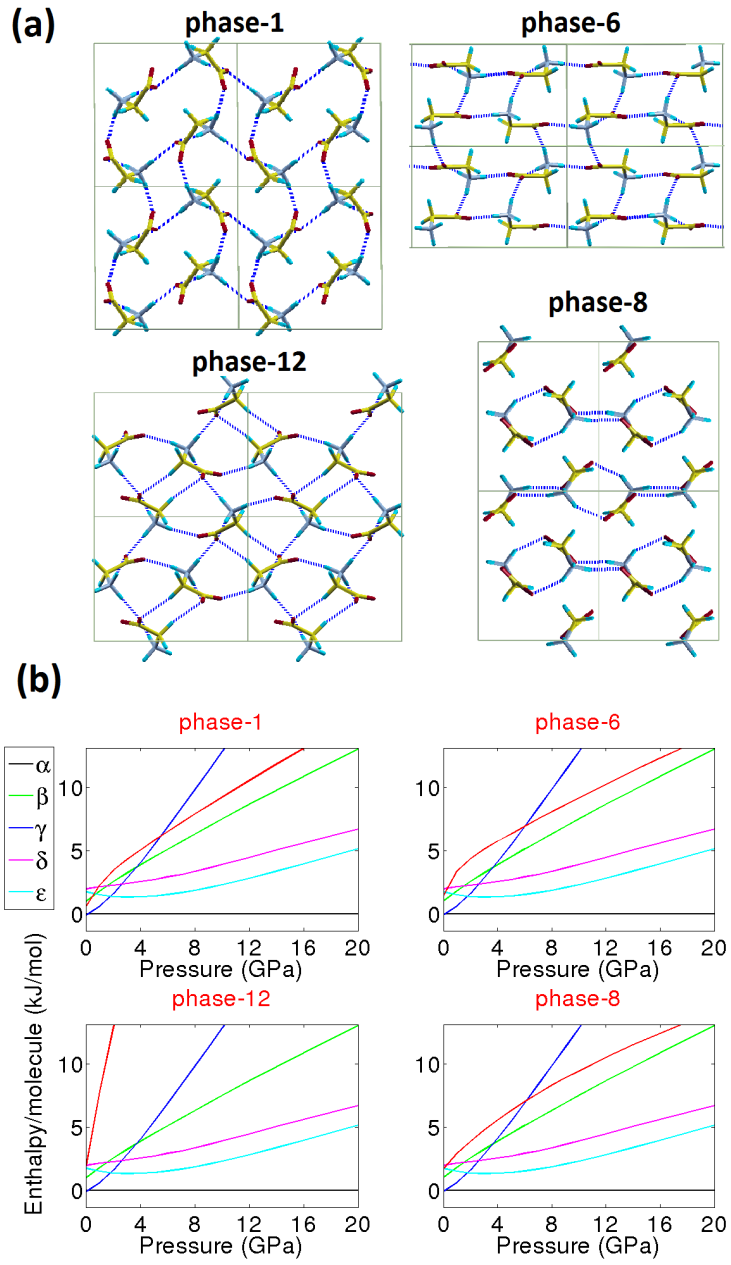

Figure 3.7: (**a**). *Crystal structure of phase-1 (above, left), phase-6 (above, right), phase-8 (below, right) and phase-12 (below, left).* (**b**). *Enthalpy/molecule with respect to the α phase as functions of pressure for phases shown in* (**a**) *are plotted as red curves correspondingly. The curves for known experimental structures are also given for comparison.*

results with existing experimental studies enabled us to resolve the so-far unidentified ζ phase a decade after its first experimental observation. Further analysis of the results of the blind test allowed us to propose several new thermodynamically plausible structures with varying volume, compressibility and polarization. To address the experimentally well established but CSP-wise challenging α phase, we introduced an intuitive sampling strategy based on crystal structure relative frequency found in nature. This strategy successfully found this challenging phase and allowed us further insight in the energy-landscape. Overall, the results of our blind test shows that a reliable crystal structure prediction procedure is possible with incorporation of several complementary recipes to reach success, emphasizing that one-size-fits-all solutions are yet to be discovered. Fortunately, the leap in precision and sampling capability we have demonstrated with these new generation tools open new paths for crystal structure prediction with data processing procedures such as clustering algorithms. Hence we strongly believe *ab initio* crystal structure prediction as presented here has come a long way and that a new standard for structure prediction for molecular crystals is set, and an interdisciplinary horizon for computational science within this field is now open.

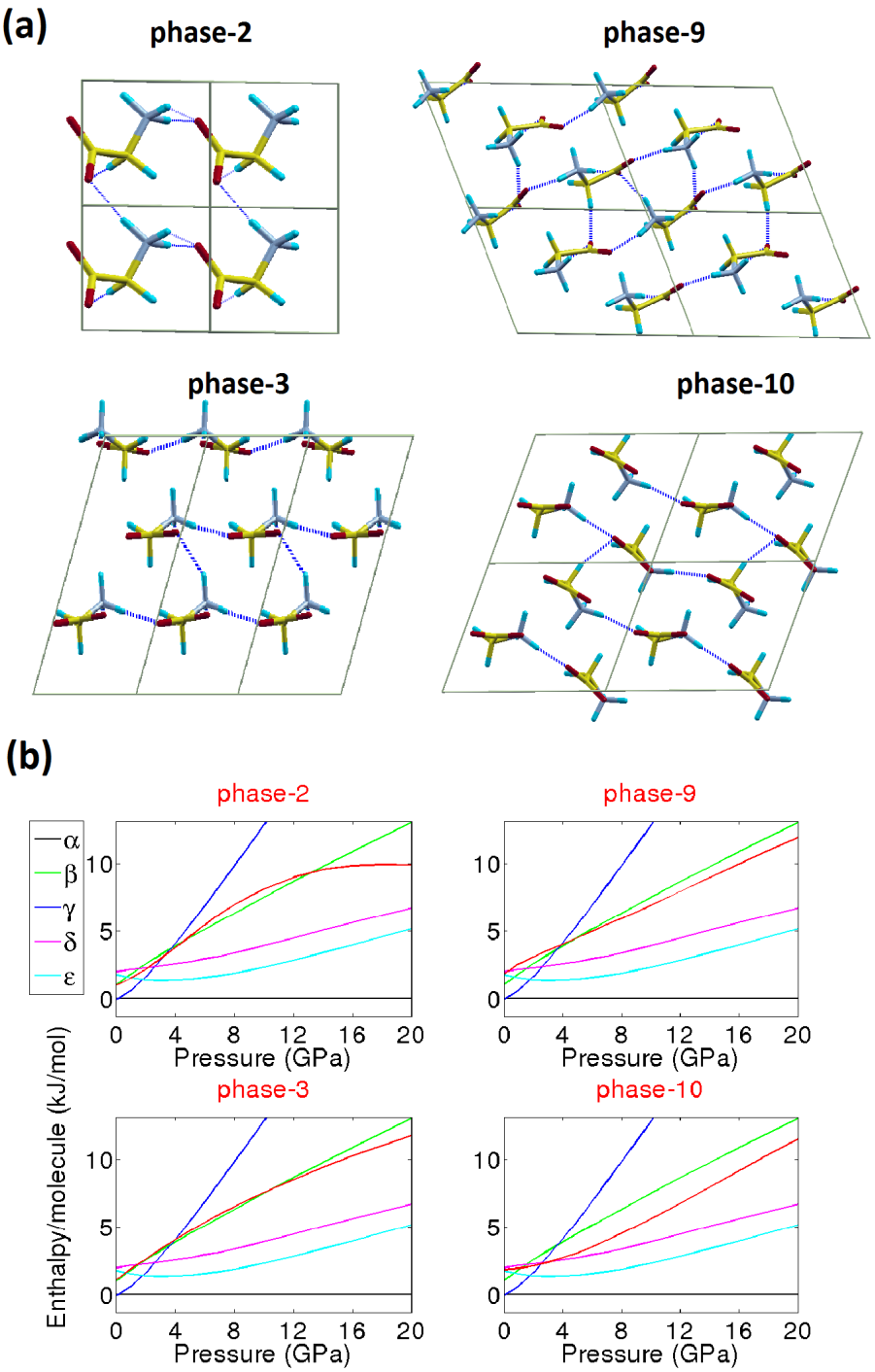

*Figure 3.8: (**a**). Crystal structure of phase-2 (above, left), phase-3 (below, left), phase-9 (above, right) and phase-10 (below, right). (**b**). Enthalpy/molecule with respect to the α phase as functions of pressure for phases shown in (**a**) are plotted as red curves correspondingly. The curves for known experimental structures are also given for comparison.*

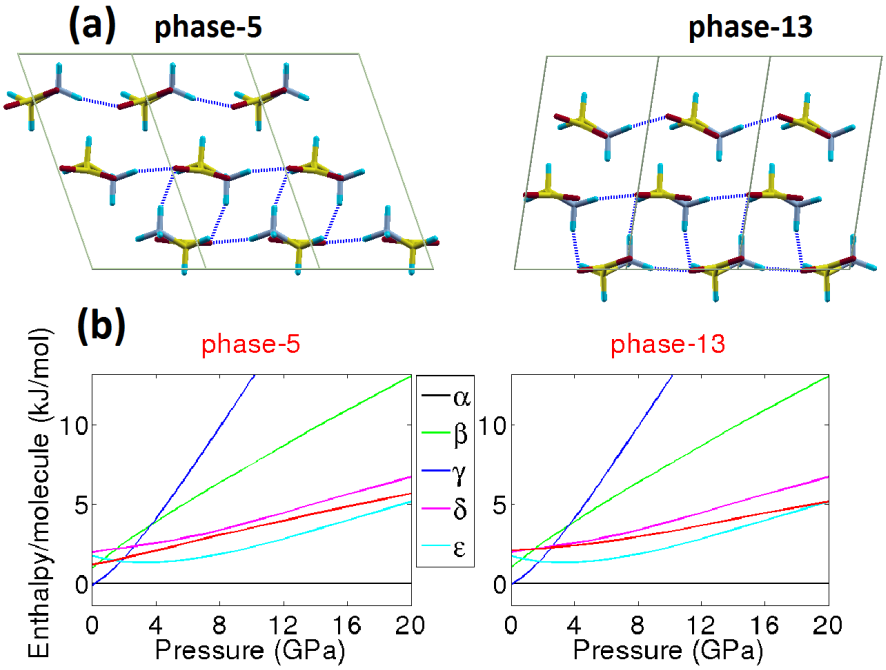

*Figure 3.9: (**a**). Crystal structure of phase-5 (left panel) and phase-13 (right panel). (**b**). Enthalpy/molecule with respect to the α phase as functions of pressure for phases shown in (**a**) are plotted as red curves correspondingly. The curves for known experimental structures are also given for comparison.*

Chapter 4

Molecular crystal structure prediction of cholesterol

In this chapter, the results of structure prediction in the challenging case of cholesterol molecular crystal are shown. First, we briefly introduce cholesterol polymorphs and our motivation for this study. Then we show the complexity of the energy-landscape of cholesterol. Next we describe the method that we used. In the results part, we provide the results of evolutionary search and the validation of the classical force-field used in this study. We also identify the experimental structure of ChAl, one known polymorph of cholesterol. Finally some new low-energy structures of cholesterol found in our simulation and their NMR characterizations are presented.

4.1 Introduction

Cholesterol, an organic molecule with chemical formula $C_{27}H_{46}O$, is a sterol biosynthesized by all animal cells and serves as a precursor for the biosynthesis of steroid hormones, bile acids, and vitamin D. The structure of cholesterol molecule is shown in Fig. 4.1(a). It is formed by four linked hydrocarbon rings forming the bulky steroid structure. The two ends of the steroid are terminated by a hydrocarbon tail and a hydroxyl group which is able to form hydrogen bonds in cholesterol solutions and/or in cholesterol crystal structures.

In nature, three polymorphs of cholesterol are known: Cholesterol Monohydrate (ChM) [90], Anhydrous Cholesterol - Low-Temperature Phase (ChAl) [91, 92], and Anhydrous Cholesterol - High-Temperature Phase (ChAh) [93, 94]. There are two types of cholesterol molecules in these phases as shown in Fig. 4.1(b). The first

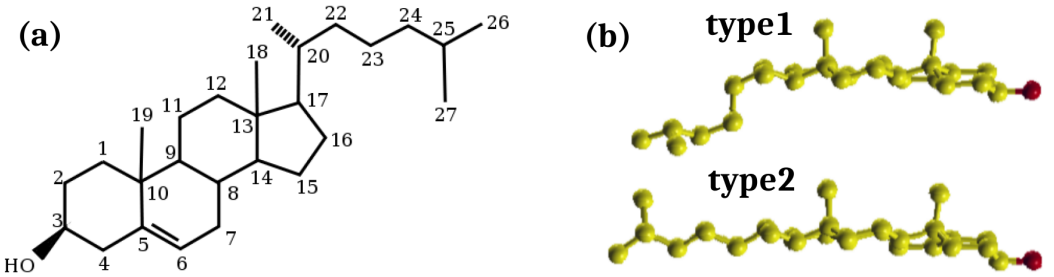

Figure 4.1: ***(a)*** *Cholesterol molecule with the numbering of carbon atoms;* ***(b)*** *two types of cholesterol molecule found in cholesterol polymorphs. Hydrogen atoms are not shown.*

molecule has a twisted tail while the other has a straight one. All three known polymorphs of cholesterol have triclinic cells with $P1$ space group symmetry. Fig. 4.2 sketches the crystal structures of the known polymorphs of cholesterol: (a) ChM, (b) ChAl and (c) ChAh.

The first known polymorph of cholesterol is ChM. It contains 8 cholesterol molecules and 8 water molecules in the unit-cell. The experimentally determined cell parameters are $a = 12.39$Å, $b = 12.41$Å, $c = 34.36$Å, $\alpha = 91.9°$, $\beta = 98.1°$, and $\gamma = 100.8°$. In the crystal structure of ChM, there are two nearly identical but differently oriented subcells with translation $(a, b/2)$ and $(b/2, a)$. The proton positions of ChM were not determined until the work of Frincu *et al.* [95].

The ChAl polymorph has 8 cholesterol molecules in the unit-cell. The cell parameters are $a = 14.172$Å, $b = 34.209$Å, $c = 10.481$Å, $\alpha = 94.64°$, $\beta = 90.67°$, and $\gamma = 96.32°$. In its crystal structure, there are two layers with two independent hydrogen bond chains that are parallel to the c-axis. Unlike ChM, there is no parallel packing in ChAl, the long molecule shows considerable variation in direction. For this phase, only positions of non-hydrogen atoms were reported experimentally. The procedure to add hydrogen atoms into this structure was suggested in Ref. [16].

At 304.8 K, the ChAl undergoes a reversible phase transition to the new phase ChAh. This phase has 16 molecules in the unit-cell with cell parameters $a = 27.565$Å, $b = 38.624$Å, $c = 10.748$Å, $\alpha = 93.49°$, $\beta = 90.90°$, and $\gamma = 117.15°$.

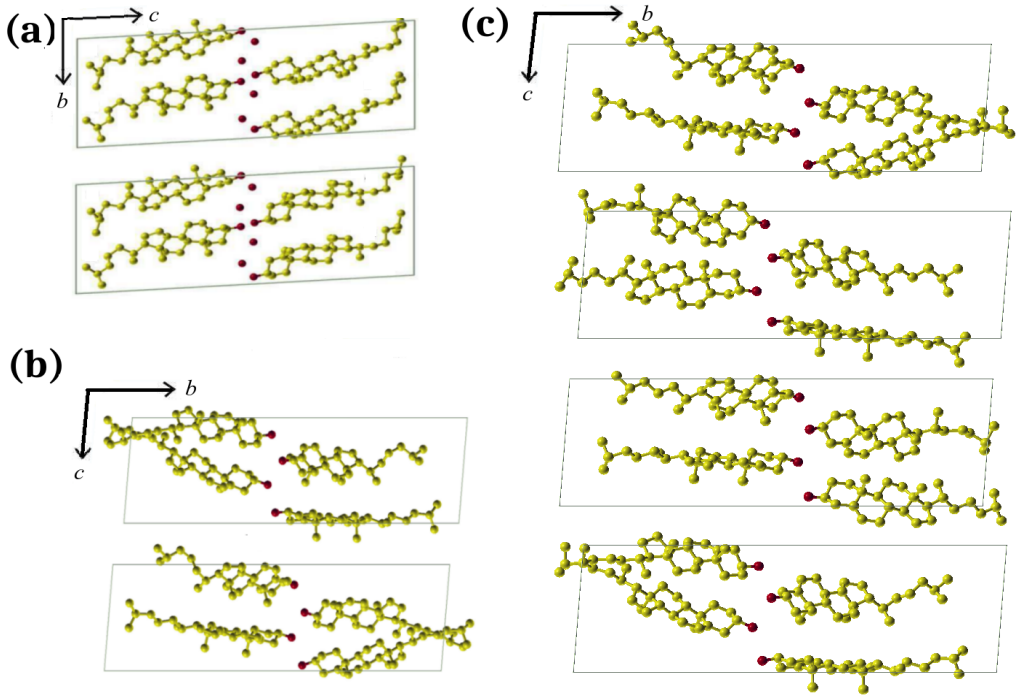

Figure 4.2: The three known polymorphs of cholesterol are shown: **(a)** *ChM,* **(b)** *ChAl and* **(c)** *ChAh. Hydrogen atoms are omitted in the case of ChAh while for ChM and ChAl, they are not experimentally determined. The cell is shown in the separate layers for clarity.*

The ChAh phase can be described as a doubling of the a-axis of the ChAl one, and therefore it also has a layered structure. For this polymorph, the crystal structure including proton positions has been fully determined [94].

Besides these three identified structures, cholesterol is suspected to have another crystalline phase that has been experimentally observed but not structurally resolved: In Ref. [15], a 13C solid state NMR experiment was conducted on human gallbladder stones made up of cholesterol, and a distinct NMR spectrum, unlike the known phases of cholesterol, was observed. Moreover this new phase could be associated to a specific gallbladder pathology, i.e. gallbladder cancer, as statistically more patients with this pathology had gallbladder stones with the new NMR signature, while patients with benign diseases such as chronic cholecystitis or xanthogranulomatous cholecystitis often had stones that yielded an NMR signature similar to the

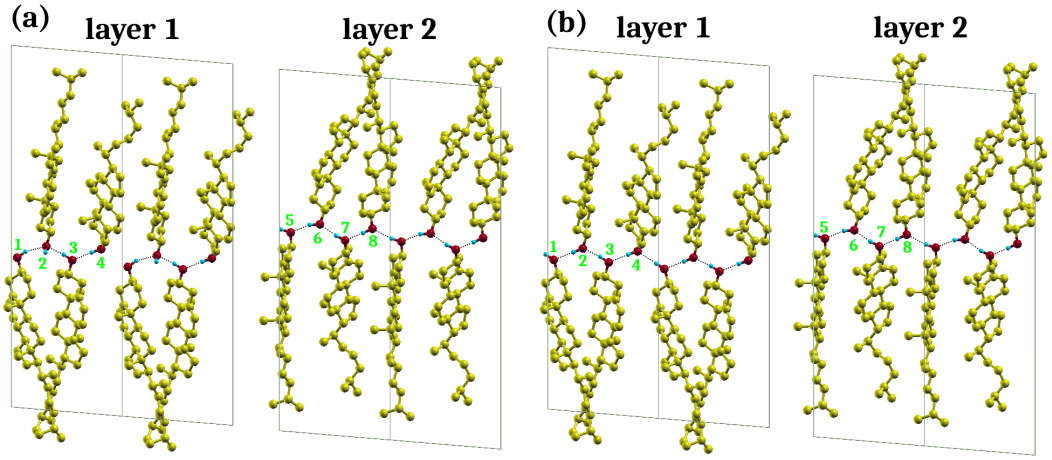

Figure 4.3: The two layers in the crystal structure of **(a)** *ChAl-1* **(b)** *ChAl-2. Only hydrogen atoms connected to oxygen ones are shown. The molecules are symbolized by their number. In each layer, the unit-cell is duplicated to display the hydrogen bond chains more clearly.*

one of ChM phase. This association of histopathology of gallbladder to the polymorphic structure of cholesterol stones reveals the importance of understanding the cholesterol polymorphism, which is the main motivation behind our work. Such better understanding of the cholesterol polymorphism, combined with the studies on in vitro crystal growth [95] holds the key to understanding the gallbladder conditions that favor the formation of such crystals and ultimately shed light on first steps of prevention strategies.

4.2 The complexity of the energy-landscape of ChAl

Before discussing the crystal structure prediction of cholesterol, we first show the complexity of the energy-landscape of this system examining ChAl, an experimental known polymorph of cholesterol with 8 molecules in the unit-cell. The cell parameters and atomic positions of carbon and oxygen atoms are experimentally well determined. However the positions of hydrogen atoms are not measurable in X-ray diffraction experiments [91, 92]. Theoretically, hydrogen positions are often assigned

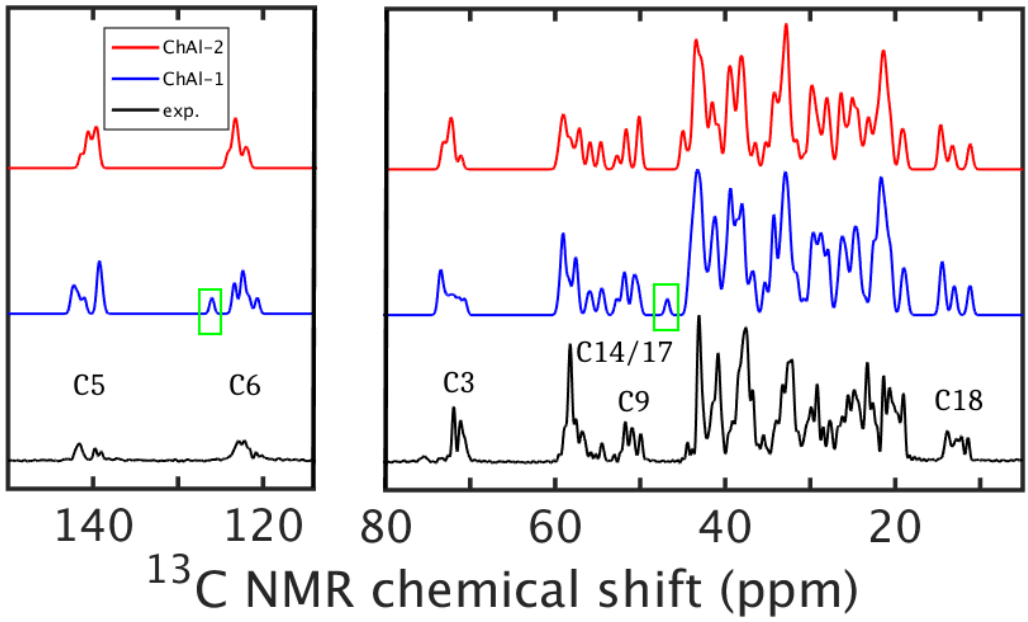

Figure 4.4: The calculated ^{13}C GIPAW spectra for ChAl-1 (blue curve) and ChAl-2 (red curve). The experimental NMR spectrum is shown as black curve for comparison.

by randomly distributing them in the neighborhood of the heavy atoms according to the expected coordination and then by relaxing the structure in order to obtain a fully optimized structure. This procedure works very well in some cases, for example cellulose polymorphs [96], flurbiprofen polymorphs [97], thymol polymorphs [98].

By using this method, the structure ChAl-1, which is shown in Fig. 4.3(a), was found in Ref. [16]. For a convenient discussion, we identify the molecules in the cell by their number as indicated in Fig. 4.3: the molecule with the number i is named as `mol`i where $i = 1, 2, ..., 8$. The NMR spectrum of this structure is shown as the blue curve in Fig. 4.4. In this section, the NMR spectra are calculated by GIPAW method described in section 4.3.3 from the structure optimized at a very tight threshold level (`relax2` level specified in section 4.3.2). We can notice that there is some disagreements between the experimental NMR spectrum and the calculated one. Some peaks of the calculated NMR spectrum, shown in the green boxes, are not present in the experimental one. Also, in comparing with

experimental spectrum, the isolated regions of C3 in the calculated spectrum is too spread. The reason for this disagreement is due to the inaccurate position of two hydrogen atoms in `mol1` and `mol2` in `layer1`. Notice that the hydrogen bond chain between cholesterol molecules is broken between `mol1` and `mol4`.

We modified these hydrogen positions restoring the hydrogen bond connectivity and relaxed again, obtaining the new structure named ChAl-2 (shown in Fig. 4.3(b)) Its NMR spectrum is presented as the red curve in Fig. 4.4 and is in better agreement with the experimental one not only at the peaks which are shown in the green box but also at the C3 peak. This improvement also implies the sensitivity of the NMR spectra to the structural details. The ChAl-2 structure has lower energy than the ChAl-1 one (5.84 kJ/mol per molecule). From this simple example one can see that in the energy-landscape of ChAl, there are many local minima and one can easily get stuck at these minima when exploring the energy-landscape.

Table 4.1 Enthalpy of all structures found by USPEX at ambient GPa.

Polymorphs	a (Å)	b (Å)	c (Å)	α (deg.)	β (deg.)	γ (deg.)	Volume(^{3})
Exp. ChAl	14.172	34.209	10.481	94.64	90.67	96.32	5032.772
ChAl-1	14.141	34.178	10.539	94.54	90.41	97.32	5035.469
ChAl-2	14.231	34.110	10.458	94.53	90.20	97.50	5017.026

It is quite unexpected that the inaccurate position of hydrogen atoms connected to oxygen can affect the chemical environment near carbon C6 which is in the second carbon ring far from the oxygen atom. In order to understand the relation between structural properties and NMR chemical shifts, the two structures ChAl-1 and ChAl-2 are compared in more details in terms of cell-parameters, bond-lengths and angles. We show, in Table 4.1, the optimized cell parameters of those structures. In both cases they agree very well with the experimental ones with a difference in volume of +0.054 % and -0.313 %, respectively. ChAl-2 has a lower volume than the ChAl-1 probably because the hydrogen bond makes the structure become more "compressed". Interestingly, it was reported that the vdW-DF functional usually overestimates the cell volume [12], however in the case of ChAl-2, we found that

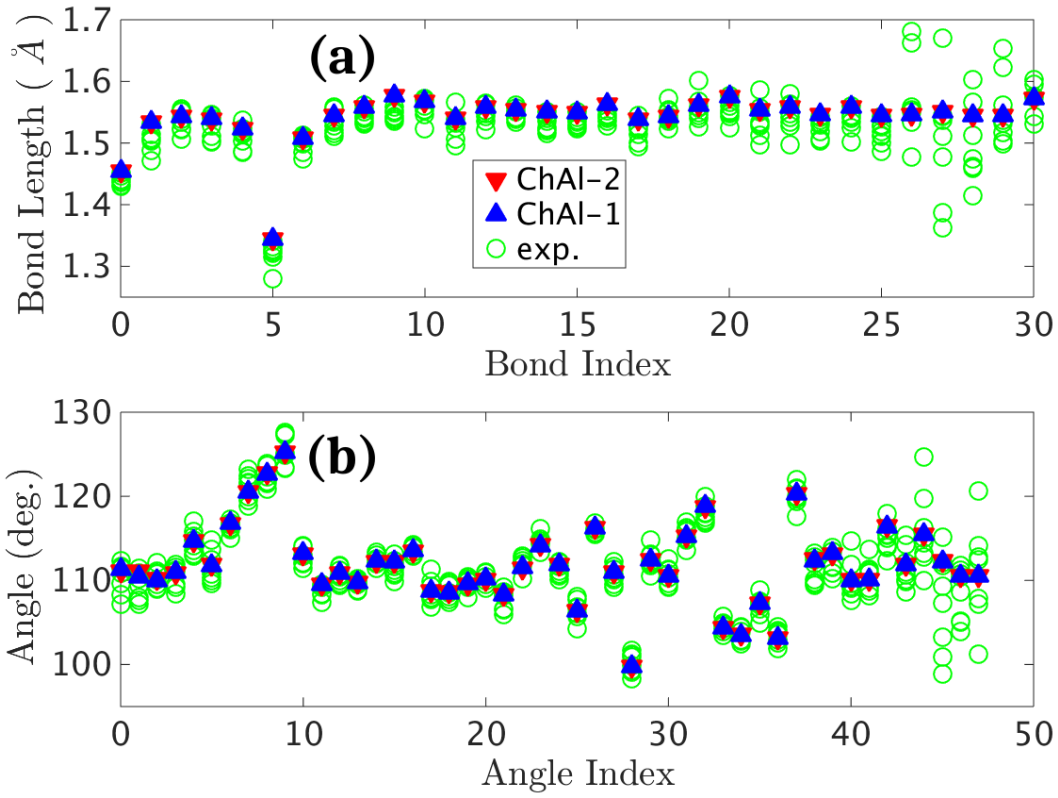

Figure 4.5: Average bond lengths (in Å) and angles (in deg.) for ChAl-1 and ChAl-2 and the bond lengths (in Å) and angles (in deg.) for experimental structure are shown in ***(a)*** *and* ***(b)****, respectively. For experimental structure, the bond lengths (angles) belonging to all 8 molecules are shown for a given bond (angle) index. The bond index is given in Fig.4.6(a).*

the optimized volume obtained with this functional is slightly smaller than the experimental one.

The results of the averages of bond lengths and angles of ChAl-1 and ChAl-2 are shown in Fig. 4.5. The experimental data for all 8 molecules are also given for comparison. The bond index is shown in Fig. 4.6(a) while the angle index is not shown. The most impressive feature of Fig. 4.5 is that on average, the bond lengths and angles of ChAl-1 and ChAl-2 are very similar and they are all very well in the range of experimental measurements. It looks like the structural differences between ChAl-1 and ChAl-2 are not large and the calculation of averages does not

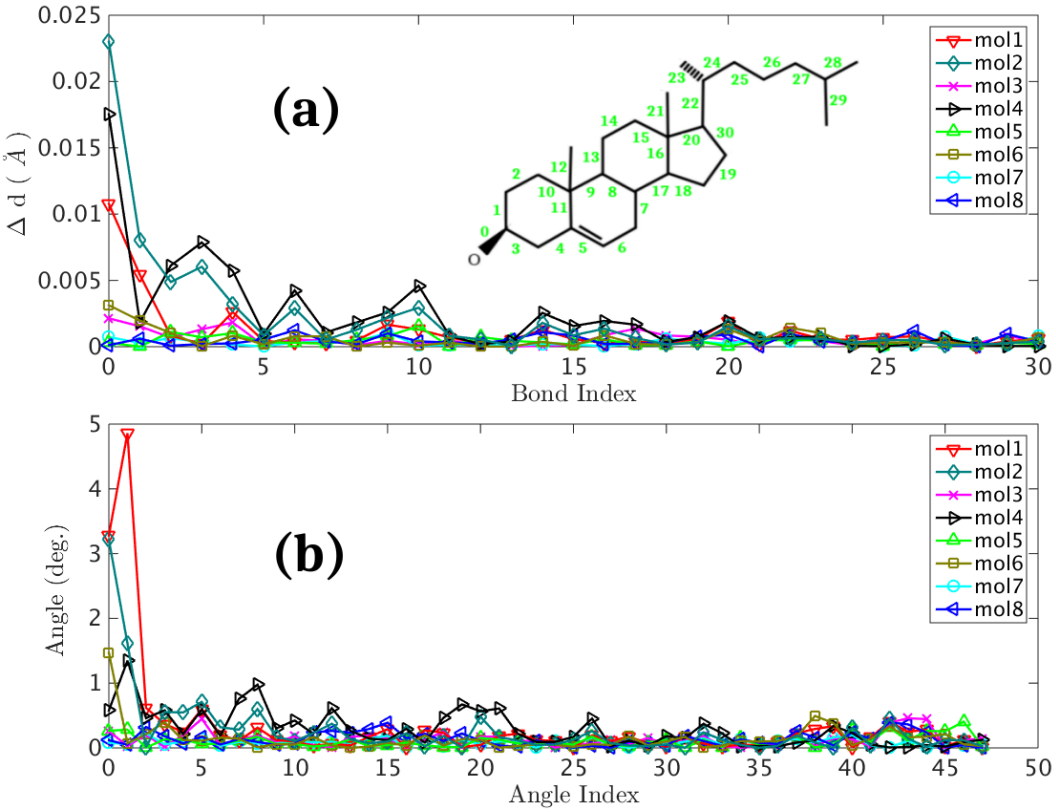

Figure 4.6: The differences in bond lengths (in Å) ***(a)*** *and angles (in deg.)* ***(b)*** *between ChAl-1 and ChAl-2 for all 8 molecules are shown. The bond index is given in* ***(a)****.*

tell where NMR differences come from.

The detailed differences in bond lengths and angles for the 8 molecules of two structures are shown in Fig. 4.6. They come mostly from the molecules in `layer1` affected by the change in the hydrogen bond network. In this layer, `mol3` is less affected by the "wrong" hydrogen bond network since the hydrogen (oxygen) atom in this molecule still makes the correct hydrogen bond with the surrounding oxygen (hydrogen) atom. As a result, the differences between the two structures are pretty small in `mol3`. In `layer2`, the largest differences between the two structure are found in `mol6` probably because it is the one closest to the `mol2` of `layer1`.

Another feature that we can notice in Fig. 4.6 is that the largest differences belong to bond lengths and/or angles of the oxygen atoms and/or the carbon atoms

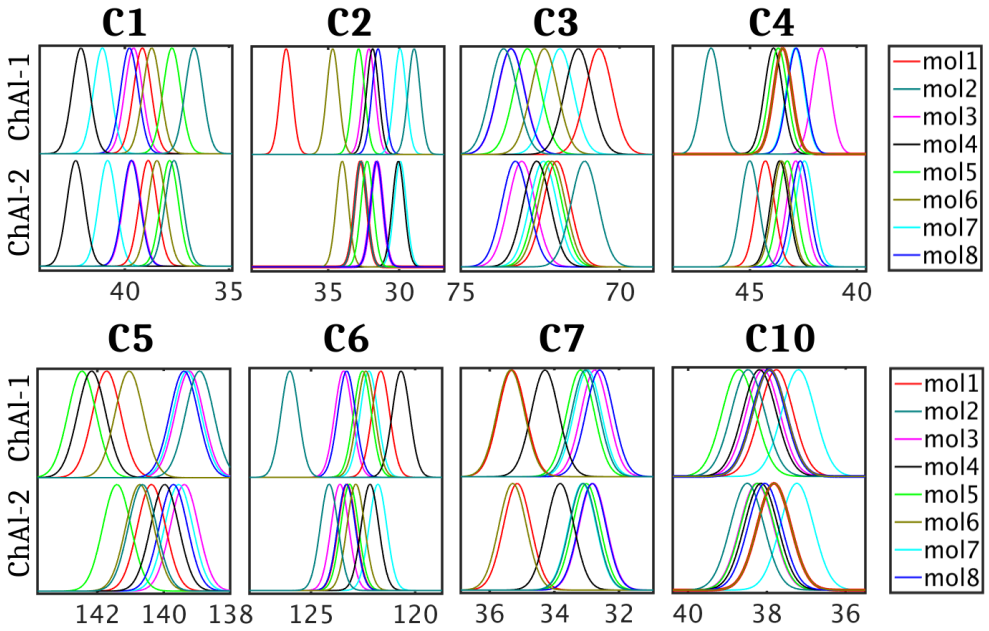

Figure 4.7: The details calculated NMR spectra of ChAl-1 and ChAl-2 of 8 molecules at the peaks of C1, C2, C3, C4, C5, C6, C7 and C10.

around oxygen. The largest differences in bond lengths are those with the index 0 (O-C1 bond), 1 (C2-C3 bond), 2 (C1-C2 bond), 3 (C3-C4 bond), 4 (C4-C5 bond), 5 (C5-C6 bond), 6 (C6-C7 bond) and 10 (C1-C10 bond). The largest differences in angles belong to those with index 0 (O-C3-C2 angle), 1 (O-C3-C4 angle) and 8 (C10-C5-C6 angle). It is interesting to notice that the effect of the wrong assignment of the hydrogen bond network can be long ranged involving chains of carbon atoms C1-C2-C10 and C3-C4-C5-C6-C7 connected to the oxygen. Also, the bond length with index 5, which is the double bond, is not affected much by the "wrong" hydrogen bond network. For others distances and angles, the differences are smaller than 0.004 Å and 0.5 deg. respectively.

The differences in crystal structure induce significant differences in NMR spectra of the two structures as shown in Fig.4.4. We also show, in Fig.4.7, the details of the calculated NMR spectra of the 8 molecules for the carbon atoms which have the largest differences between the two spectra. Among them, the differences of C6 are seen also in the full spectrum in Fig.4.4 while although C2 and C4 also display

huge differences between the two spectra, they are all in the "crowded" region of the spectrum, and therefore can not be seen easily in Fig.4.4.

The results in this part show that the assignment of proton positions, which are not measurable in the XRD experiments, is non-trivial. Inaccurate determination of proton positions can result in different structure from the ground state one. The complexity of the energy-landscape of ChAl and the sensitivity of NMR spectra to the structural details are also revealed.

4.3 Method

4.3.1 Evolutionary search

In this study we used evolutionary algorithm in the USPEX code to search for possible low-energy structures of cholesterol with 8 molecules/cell. We choose this number for easy comparison with the experimental structure of the well-known phase ChAl. Since the number of atoms in the unit-cell in this case is huge (592 atoms/cell), the use of a primary energy screening with a classical force-field is needed. We use the all-atom force-field as parameterized in Ref. [16]. All the classical molecular dynamic calculations are done using the LAMMPS code [99].

The structural relaxation process is done in three steps. In the first step of the relaxation, the cell is allowed to change while in the second one, only the atom coordinates are optimized. The last step is a molecular dynamics calculations with a Langevin thermostat where the temperature is kept at 0 K for 25 ns with a time step of 0.5 fs. All calculations are performed at zero pressure, and the search is continued up to 40 generations.

Motivated by the fact that in the experimental structure of ChAl, there are two types of cholesterol molecules, we consider two different molecular conformations as shown in Fig. 4.1(b). The crystal structure prediction is performed for structures containing i molecule(s) of `type1` and $8-i$ molecule(s) of `type2` where $i = 0, 1, 2, ..., 8$. The first generation consisting of 30 structures was created randomly. New generations were created 20% randomly and 80% through heredity (40%), soft mutation (20%) and transmutation (20%) from the remaining struc-

tures after discarding 40% of the energetically less stable structures.

4.3.2 *ab initio* calculations

When the USPEX calculation is finished, some best structures are relaxed using DFT with the vdW-DF exchange-correlation functional as implemented in the QUANTUM ESPRESSO package [83]. A kinetic energy cutoff of 45 Ryd, a charge density cutoff of 220 Ryd and PAW pseudopotentials from Ref. [100, 101] are used. In this work, two levels of relaxation are discussed. The `relax1` has a convergence of less than 0.1 mRy for total energy and less than 0.0005 Ry/a.u. for the forces on atoms and less than 0.005 GPa for the stress tensor. The `relax2` is the same as `relax1`, except that the forces on atoms are less than 0.00025 Ry/a.u.

4.3.3 GIPAW calculations

The theoretical chemical shielding tensor is calculated using the GIPAW code in Quantum Espresso package. The isotropic chemical shifts is related to the isotropic chemical shielding by using the secondary reference scheme to improve comparison with experimental data [100, 101] $\delta_{\text{iso}} = \sigma_{\text{ref}} - c - \sigma_{\text{iso}}$. In this work we used $\sigma_{\text{ref}} = 167.5$ ppm and the correction values c are different for different local chemical environments of carbon atoms, as used in Refs. [100, 101]. The theoretical spectra are drawn using normalized Lorentzian distribution centered at the chemical shifts with an arbitrary broadening corresponding to a FWHM of 0.5 ppm.

4.4 Results

4.4.1 Results of evolutionary algorithm

First we emphasize again that for crystal structure prediction, the use of accurate DFT calculations in the relaxation step is a very good option. In this study however since the number of atom in the unit-cell is huge, we have to use classical force-fields as a primary energy screening. The crystal structure prediction as described in section 4.3.1, is done for different combinations of two cholesterol molecules.

Notation (i, j) is used for structures with i molecule(s) of `type1` and j molecule(s) of `type2`. In this study, the number of molecules in the unit-cell is Z=8, i.e. $i+j = 8$ and nine simulations (corresponding to $i = 0$ to 8) are performed.

The results of crystal structure prediction can be analyzed through the distribution of energy as a function of volume for the structure encountered during the search as shown in Fig. 4.8. In contrast to the study of glycine in Fig. 3.2 in which the structures forming isolated islands, the energy-volume points are not well separated enough to identify different structures. One reason for this problem is that the energy-landscape of the classical force-field is quite noisy. Another reason is the variation of the hydrogen bond network. As shown in section 4.2, different hydrogen bond networks can result an energy difference of 1.4 kcal/mol per molecule.

4.4.2 Validation of the classical force-field

The ability of the force-field to describe accurately the properties of cholesterol crystals must be verified in order to do crystal structure prediction successfully. A force-field that can give good energy ordering for cholesterol crystals is desired.

All lowest energy structures in each case in Fig. 4.8 are optimized at level of `relax1` using DFT with the vdW-DF functionals. [10, 32]. In order to compare the energy-landscapes of the force-field and DFT, some low-lying energy structures for each (i, j) pair are also relaxed. We show in Fig. 4.9(a) the DFT energies (Ryd) as a function of the force-field energies E_1 (kcal/mol) for different structures of different (i, j). The line in the figure shows the correlation between DFT energies and force-field ones (the conversion between Ryd and kcal/mol). If the force-field and DFT had the same energy hypersurface, all structures would be on this line. In this case, we notice that the distribution of the structures is kind of "disordered" and it seems that force-field and DFT energies have no correlation at all, i.e. the potential-energy hypersurfaces of the force-field and of DFT are very different.

However as mentioned in Sec. 2.3.2, the no correlation between force-field and DFT may be due to a very noisy potential-energy hypersurface as illustrated in Fig. 4.9(b). Even when these hypersurfaces of the force-field and of DFT are very similar, there usually is some noise that produces many local minima in the energy-

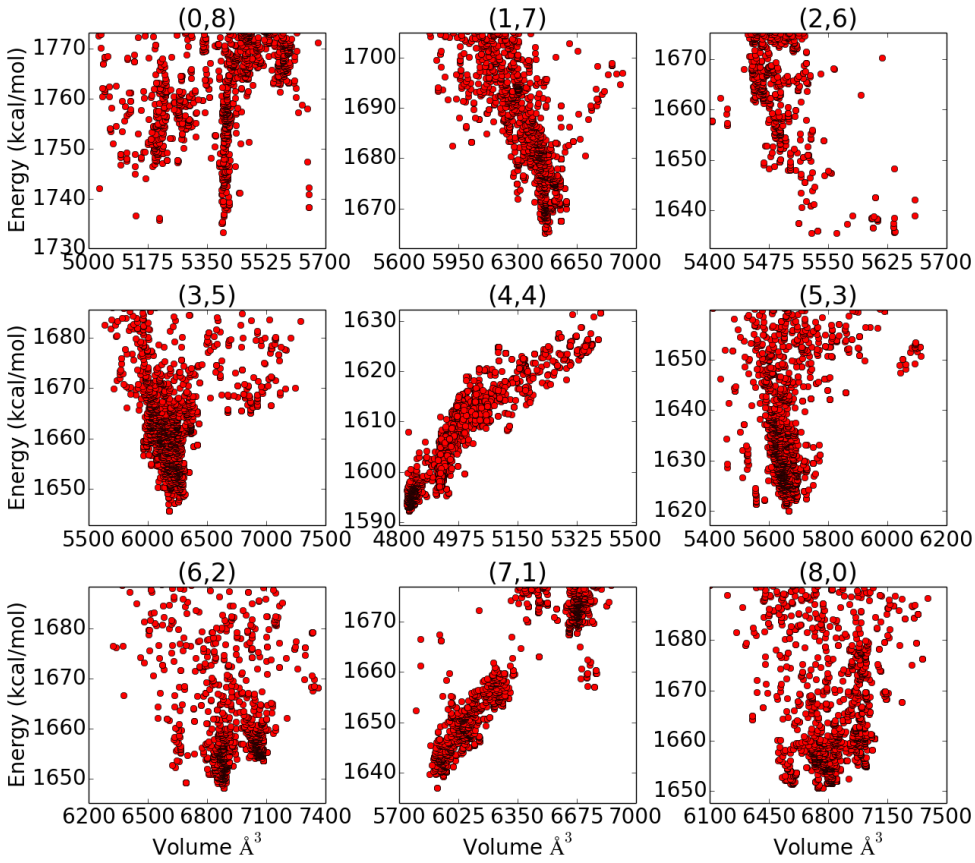

Figure 4.8: Energy vs. volume distribution of all evolutionary searches in the lowest 40 kcal/mol range. The number of molecules in the unit-cell is Z=8 and the searches are done for different combinations of two cholesterol molecules (see text).

landscape of the force-field. The structures optimized with this force-field may get stuck at these local minima; and switching to the DFT minimization can result in a significant energy change. To examine this aspect, the optimized structures at DFT level were relaxed again with the force-field. The force-field energies E_1 then changed to E_2. Fig. 4.9(c) shows the force-field energies E_2 as a function of force-field energies E_1. We notice that most structures after this second optimizing with the force-field have $E_2 < E_1$, sometimes significantly so and the overall picture is very similar to Fig. 4.9(a). More interestingly, we show in Fig. 4.9(d) the DFT energies as a function of the new force-field energies E_2. The structures show rather good correlation meaning that the energy hypersurfaces of the force-field and of

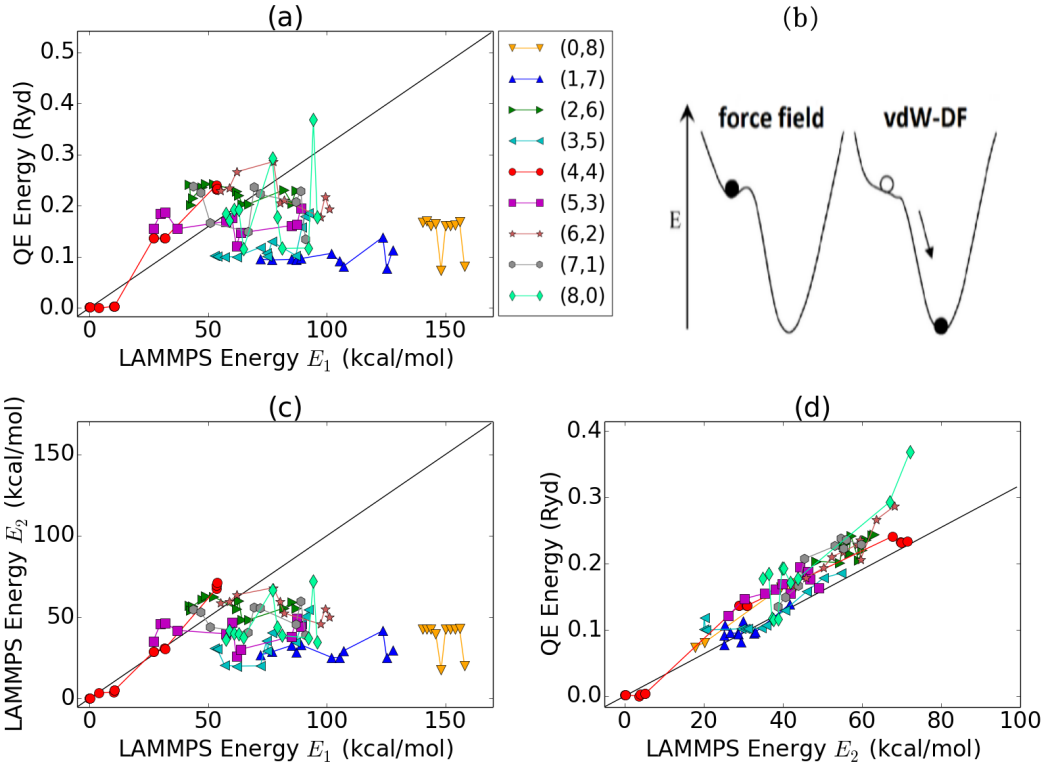

Figure 4.9: For different structures (i, j), DFT energies (with vdW-DF functional) as a function of force-field energies E_1 and E_2 shown in **(a)** *and* **(d)**, *respectively, The dependence of force-field energies E_2 on E_1 is presented in* **(c)**. *The similarity of the hypersurfaces of the force-field and DFT is schematized in* **(b)**.

DFT (with vdW-DF functional) are very similar and we can use the force-field as a primary energy screening.

It is worth mentioning that in Fig. 4.9 the (4,4) structures show good correlations in all cases. The reason is the classical force-field used in this study is designed for the experimental structure ChAl which has 4 molecules of each type (`type1` and `type2`) in the unit-cell.

4.4.3 Relations between levels of relaxation, structural properties and NMR spectra

In order to determine whether two structures (namely μ and ν) are the same or not we compare the mean deviation between bond lengths (MDBL) and mean deviation between angles (MDA) defined as

$$\mathrm{MDBL}(\mu,\nu) = \frac{\sum_{i=1}^{N_{\mathrm{bond}}} |d_i^{(\mu)} - d_i^{(\nu)}|}{N_{\mathrm{bond}}}, \qquad \mathrm{MDA}(\mu,\nu) = \frac{\sum_{i=1}^{N_{\mathrm{angle}}} |\alpha_i^{(\mu)} - \alpha_i^{(\nu)}|}{N_{\mathrm{angle}}}, \tag{4.1}$$

where N_{bond} and N_{angle} are the number of bond-lengths and angles in the structure, $d_i^{(\mu)}$ and $\alpha_i^{(\mu)}$ are the bond-length and angle of the index i in the structure μ, respectively. The bond-lengths and angles are calculated for non-hydrogen atoms only.

Before discussing the relations between levels of relaxation, structural properties and NMR spectra of cholesterol structures, we need to quantify the differences between two calculated NMR spectra. A similar technique as in the comparison of bond lengths and angles, can be used. The mean deviation between NMR spectra of structure μ and structure ν is defined as

$$\mathrm{MDNMR}(\mu,\nu) = \frac{\sum_{i=1}^{N_{\mathrm{peak}}} |\delta_i^{(\mu)} - \delta_i^{(\nu)}|}{N_{\mathrm{peak}}}, \tag{4.2}$$

where N_{peak} is the number of carbon atoms in the unit-cell; the $\{\delta_i^{(\mu)}\}$, which are sorted in ascending order, are the calculated ^{13}C chemical shielding values of the structure μ.

First the structures are optimized at the level of `relax1` and the NMR spectra are calculated correspondingly. We only show the results for three structures (namely struc-1, struc-2 and struc-3) in the lowest energy region in Fig. 4.9(d) obtained in the $(4,4)$ simulation. Table 4.2 shows the MDBL, MDA and MDNMR for these structures.

In term of structural properties at optimization of `relax1`, struc-1 and struc-2 are very similar while the differences between struc-1 (struc-2) and struc-3 are significantly larger. As a result, the NMR spectra of struc-1 and struc-2 shown in Fig. 4.10(a), are nearly identical with a MDNMR of 0.0807 ppm per carbon atom

Table 4.2 MDBL, MDA and MDNMR for struc-1, struc-2, struc-3 at the levels of `relax1` and `relax2`

MDBL

`relax1`				`relax2`			
	struc-1	struc-2	struc-3		struc-1	struc-2	struc-3
struc-1	0.0000	0.0004	0.0028	struc-1	0.0000	0.0002	0.0020
struc-2		0.0000	0.0027	struc-2		0.0000	0.0020
struc-3			0.0000	struc-3			0.0000

MDA

`relax1`				`relax2`			
	struc-1	struc-2	struc-3		struc-1	struc-2	struc-3
struc-1	0.0000	0.0974	0.8541	struc-1	0.0000	0.0515	0.6516
struc-2		0.0000	0.8660	struc-2		0.0000	0.6563
struc-3			0.0000	struc-3			0.0000

MDNMR

`relax1`				`relax2`			
	struc-1	struc-2	struc-3		struc-1	struc-2	struc-3
struc-1	0.0000	0.0807	0.2995	struc-1	0.0000	0.0591	0.1428
struc-2		0.0000	0.2864	struc-2		0.0000	0.1434
struc-3			0.0000	struc-3			0.0000

whereas MDNMR of struc-1 and struc-3 is 0.2995 ppm per carbon atom. The other quantity that can be considered is the maximum difference between corresponding chemical shieldings $\delta_{\max} = \max_{\{i\}}|\delta_i^{(\mu)} - \delta_i^{(\nu)}|$. For the comparison of struc-1 and struc-2, $\delta_{\max}$ is 0.32 ppm which is smaller than typical numerical error in the calculation of the chemical shielding. In the case of struc-1 and struc-3, $\delta_{\max}$ is 1.22 ppm which is significant. Therefore in Fig. 4.10(a) struc-3 shows a different NMR spectrum compared with the others two. All NMR spectra of the chosen structures do not agree very well with the experimental one.

We did the calculations again using the `relax2` level in the relaxation. The better

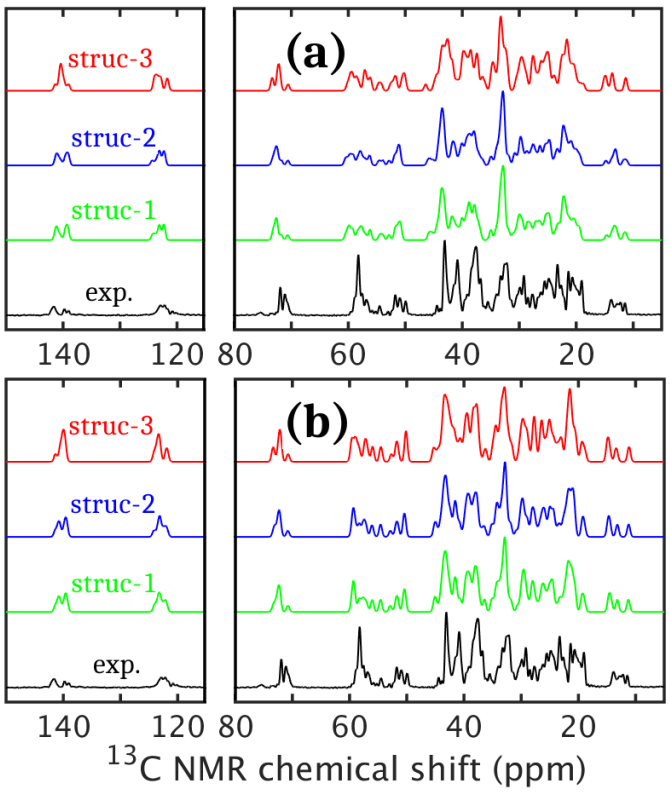

Figure 4.10: The calculated NMR spectra of struc-1 (green curve), struc-2 (blue curve) and struc-3 (red curve) (see text) are shown at the level of `relax1` *in* ***(a)*** *and* `relax2` *in* ***(b)****. The experimental spectrum is also given for comparison.*

optimization improves not only the structural properties but also the NMR spectra. As shown in Table 4.2, at `relax2` level, the values of MDBL, MDA, MDNMR for all structures are smaller when compared with those at `relax1` level. We notice that in Fig. 4.10(b) the differences in NMR spectra between struc-3 and struc-1 (struc-2) are smaller than those in Fig. 4.10(a). For struc-1 and struc-3 comparison, $\delta_{\max}$ is now 0.5 ppm which is the typical numerical error in the calculation of the chemical shielding. Therefore we conclude that struc-1, struc-2 and struc-3 are indeed the same structure although a tighter relaxation would be necessary to bring struc-3 to approach the other two structures to a similar degree. Another important result is that since cholesterol is a complex molecule with a noisy energy-landscape and the NMR spectrum is very sensitive to structural changes, a rather tight convergence

condition is required to identify the crystal structure and the corresponding NMR spectrum.

4.4.4 Identifying the experimental structure of ChAl

The lowest energy structure found at the level of `relax2` is struc-1 which is later named [4, 4], i.e. the lowest energy structure found in the simulation (4, 4). Its optimized cell parameters are $a = 14.220$Å, $b = 34.158$Å, $c = 10.457$Å, $\alpha = 94.46°$, $\beta = 90.36°$, $\gamma = 96.89°$ and the cell volume is 5026.494 Å^3. These parameters are very close to those of the structure ChAl-2 in Table 4.1. They both agree very well with the cell parameters of the experimental structure.

We show in Table 4.3 the comparison of MDBL and MDA for three structures: the experimental structure (Exp. ChAl) [91, 92], the ChAl-2 one obtained in this study after restoring the hydrogen bond network as explained earlier and the struc-1 found by USPEX. The MDBL and MDA of ChAl-2 and struc-1 are pretty small. They are much smaller than those values for struc-1 (or ChAl-2) and Exp. ChAl. The reason is that ChAl-2 and struc-1 are optimized using the same functional with a tight convergency condition. In term of structural properties, the agreement of struc-1 and the Exp. ChAl is fairly good. Their crystal structures are shown in Fig. 4.11.

The NMR spectrum of struc-1 agrees very well with the experimental one, as shown in Fig. 4.10(b) (the green curve). We also compare the calculated NMR spectra of struc-1 and ChAl-2. The MDNMR of struc-1 and ChAl-2 is 0.1069 while $\delta_{\max} = 0.47$ ppm which is smaller than the typical numerical error in the calculation of the chemical shielding.

From the results obtained in this section we can conclude that the crystal structure prediction simulation by USPEX with the adopted force-field was indeed able to identify the experimental ChAl structure successfully.

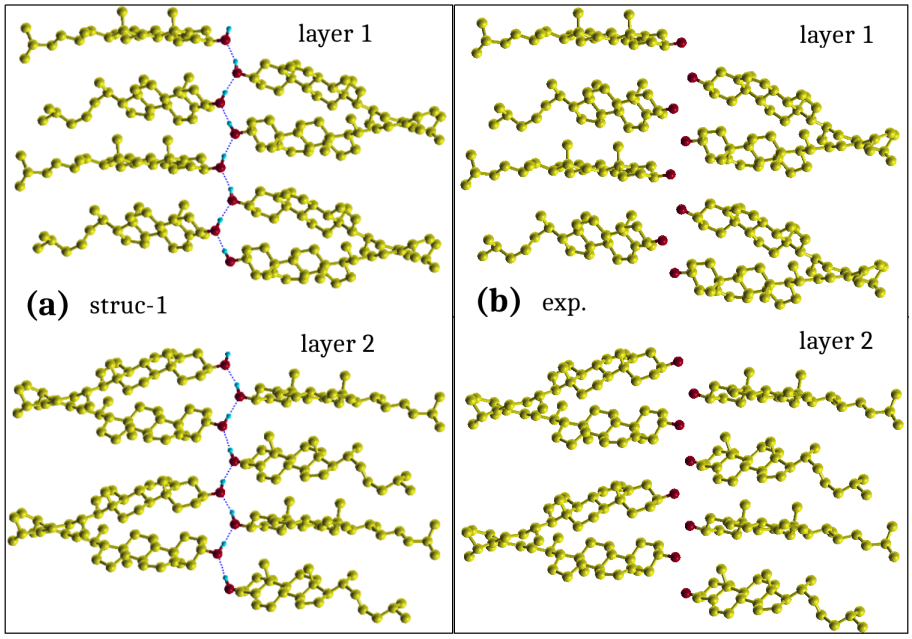

Figure 4.11: The two layers in the crystal structure of **(a)** *struc-1 and* **(b)** *experimental structure. Only hydrogen atoms connected to oxygen ones are shown in the struc-1. In each layer, the unit-cell is duplicated to display the hydrogen bond chains more clearly.*

Table 4.3 MDBL and MDA for struc-1, Exp. ChAl , ChAl-2 at the levels of `relax2`

Structure μ	Structure ν	MDBL (Å)	MDA (deg.)
Exp. ChAl	ChAl-2	0.038956	2.695331
Exp. ChAl	struc-1	0.038938	2.693725
ChAl-2	struc-1	0.000325	0.093429

4.4.5 Prediction of new structures of cholesterol and their NMR spectra

The successful identification of the experimental ChAl structure is an important result that validates our approach. Our simulations allow us to identify other low-lying structures that might be relevant in the interpretation of experiments [15]. We suggest some low-energy structures as below.

We use notation $[i, j]$ for the lowest energy structure found in the simulation

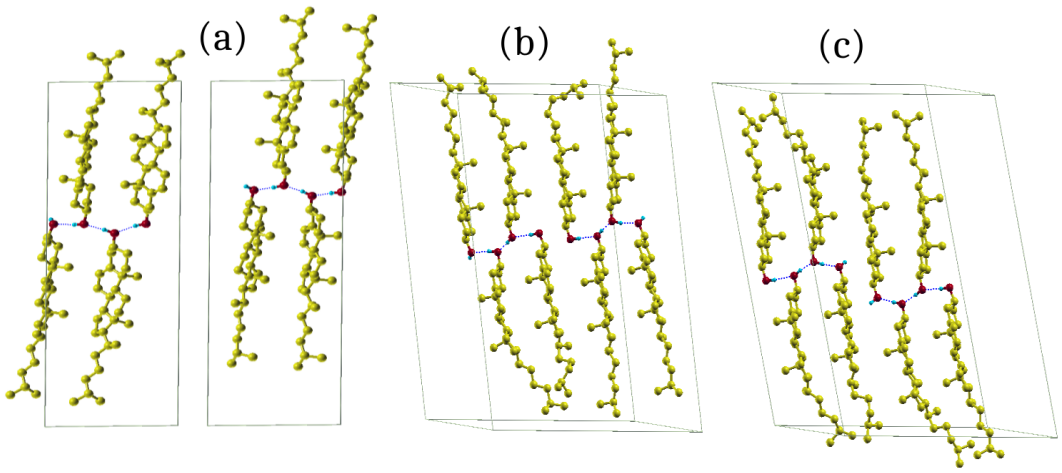

Figure 4.12: Three low-energy structures found by USPEX: [0,8], [1,7], and [3,5] are shown in ***(a)****,* ***(b)****, and* ***(c)*** *respectively. Only hydrogen atoms connected to oxygen ones are shown.*

(i, j), i.e. for structures having i molecule(s) of `type1` and j molecule(s) of `type2` (see section 4.4.1). Particularly low-energy structures are found in the case (0,8), (1,7), and (3,5). We show the crystal structures of structures [0,8], [1,7], and [3,5] in Fig. 4.12 (only hydrogen atoms which contribute to hydrogen bonds are shown). One thing can be noticed. The structures are generated from two types of melecules with different tails as shown in Fig. 4.1(b). However, in the relaxation process the molecular shapes are allowed to change, therefore, the resulted structures at the end may have different molecular conformations as seen in structure [1,7]. The [1,7] and [3,5] display cholesterol molecules organized in a single layer connected by periodically broken hydrogen bond chains, in contrast with the experimentally known structure ChAl which has a bilayer structure. The [0,8] also has a bilayer structure and is shown in Fig. 4.12(a) with cell parameters $a = 12.468$Å, $b = 11.812$Å, $c = 34.972$Å, $\alpha = 89.53°$, $\beta = 84.87°$, $\gamma = 94.52°$ and cell volume is 5113.331 Å^3. The structure $[1, 7]$ in Fig. 4.12(b) has cell parameters $a = 12.588$Å, $b = 25.627$Å, $c = 33.927$Å, $\alpha = 94.95°$, $\beta = 96.24°$, $\gamma = 27.75°$ and cell volume is 5064.438 Å^3. Fig. 4.12(c) shows the structure $[3, 5]$ with cell parameters $a = 12.455$Å, $b = 25.203$Å, $c = 34.561$Å, $\alpha = 97.80°$, $\beta = 95.92°$, $\gamma = 28.41°$ and cell

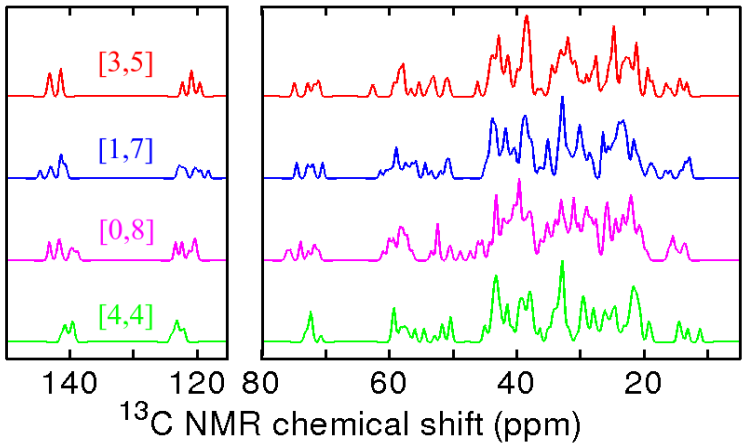

Figure 4.13: Calculated NMR spectra of three lowest energy structures found by USPEX: [4,4], [0,8], [1,7], and [3,5]. These spectra show distinct features.

volume is 5110.447 $Å^3$. In term of energy, structures $[0, 8]$, $[1, 7]$, and $[3, 5]$ are 0.173 (kJ/mol) per atom, 0.183 (kJ/mol) per atom, and 0.228 (kJ/mol) per atom higher than the experimental ChAl-2, respectively.

Fig. 4.13 shows the calculated NMR spectra of the 4 lowest energy structures found by USPEX: [4,4], [0,8], [1,7], and [3,5]. It is clear that each structure shows distinct NMR features that could be used in the comparison with experimental spectra to detect whether structures [0,8], [1,7], and [3,5] can exist experimentally.

4.5 Conclusion

In this chapter, we presented a crystal structure search for cholesterol polymorphs. First we show the complexity of the energy-landscape of cholesterol: the strong effect of an inaccurate determination of the proton positions in the hydrogen bond network leading to unsuccessfully finding the true ground state structure of ChAl; and the level of sensitivity of the NMR spectra with respect to the structural details. In the search for low-energy structures of cholesterol, we first validate the classical force-field used in this study and suggest an efficient scheme for crystal structure prediction of large molecule using classical force-field. We then identify the lowest energy structure found in the USPEX simulation as the experimental structure ChAl

by comparing their structures and the corresponding NMR spectra. We also propose a few new low-energy structures of cholesterol and characterize them by their NMR spectra.

Chapter 5

Summary and Outlook

The present work has focused on the *ab initio* crystal structure prediction of molecular crystal using evolutionary algorithm. Promising results have been obtained in two different cases: a simple molecule with rich polymorphism in the case of glycine and the cholesterol molecule, which is long, flexible, with a noisy energy-landscape. In both cases, the use of last generation van der Waals density functionals allows us to obtain stability ordering correctly.

In the study of glycine, clustering technique is used to analyze the results of evolutionary algorithms more effectively. In order to explore the energy-landscape efficiently, we propose an intuitive sampling strategy based on crystal structure symmetry relative frequency found in nature. This strategy successfully address the experimentally well-established α-glycine, which a simulation using standard setting could not. The comparison of our calculations with available experimental data identifies ζ-glycine, a metastable phase that is experimentally observed but whose crystal structure was not resolved yet. We also propose several new low-energy structures of glycine at ambient and high pressure. Through this example, we believe that the *ab initio* structure prediction of molecular crystal has come a long way toward a new standard. The situation is now such that polymorphs stability-order can be successfully addressed thanks to last generation van der Waals density functionals; and the energy-landscape can be efficiently explored thanks to recent developments in evolutionary crystal structure search.

The crystal structure prediction problem of cholesterol is more challenging because of two reasons. First the number of atoms in the unit-cell is huge and therefore the computational time would be too long for the optimization of candidate structures to be done using accurate DFT methods. The second one comes from the

nature of the molecule. Since cholesterol is a flexible molecule, cholesterol polymorphs usually have more than one inequivalent molecular conformation in the unit-cell and therefore the energy-landscape is more complicated. In this study, to overcome the second issue, we use two molecular types and the simulations are done for different conformations that are made from these two types. The use of classical force-field can be a good solution to the first problem if the force-field can describe "reasonably well" the energy ordering. We therefore validated the force-field used in this study as a primary selection method. Then accurate DFT calculations with a functional that includes van der Waals interaction, were used to obtain higher accuracy in the descriptions of structural and energetic properties. The predicted lowest-energy structure is identified with the experimental ChAl one based on their good agreement for crystal structure parameters as well as NMR spectra. In this study, a few more low-energy structures of cholesterol are proposed and their NMR features are characterized.

Appendix A

Detailed results of evolutionary algorithm searches for low-energy structure of glycine

---

In this appendix, we present the detailed results of all search attempts for the low-enthalpy structures of glycine at zero pressure for $Z = 2$, $Z = 3$ and $Z = 4$. These figures are mentioned in sections 3.3.5 and 3.3.6.

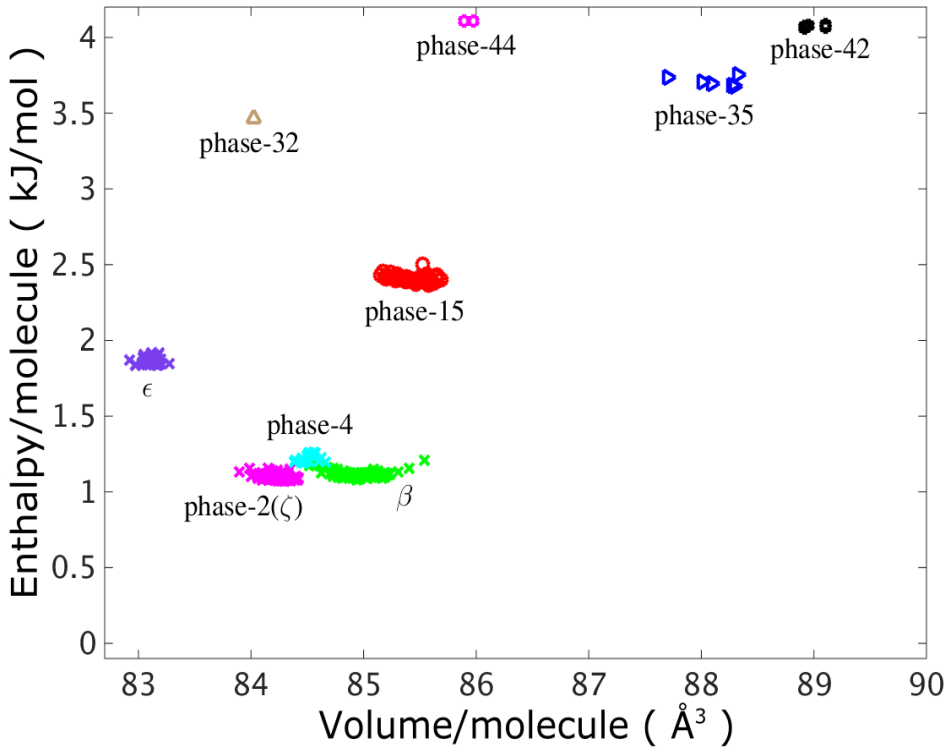

Figure A.1: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 2 molecules/cell without using any experimental information. The simulation is run for 18 generations with settings described in Section 3.2. The clusters are labeled according to Table 3.5.

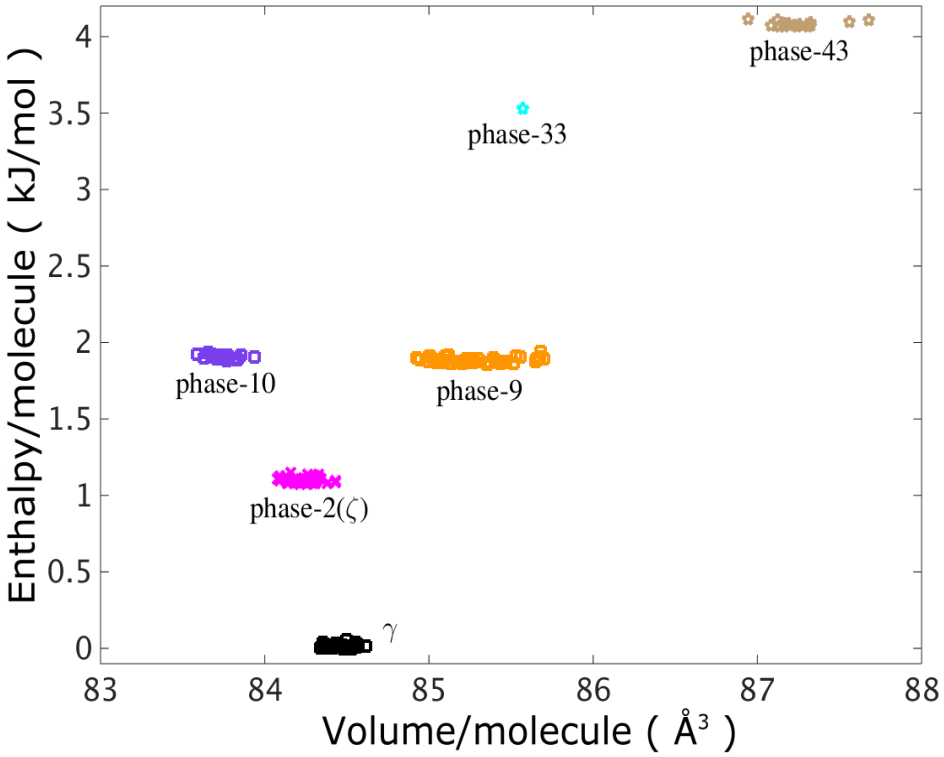

Figure A.2: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 3 molecules/cell without using any experimental information. The simulation is run for 15 generations with settings described in Section 3.2. The clusters are labeled according to Table 3.5.

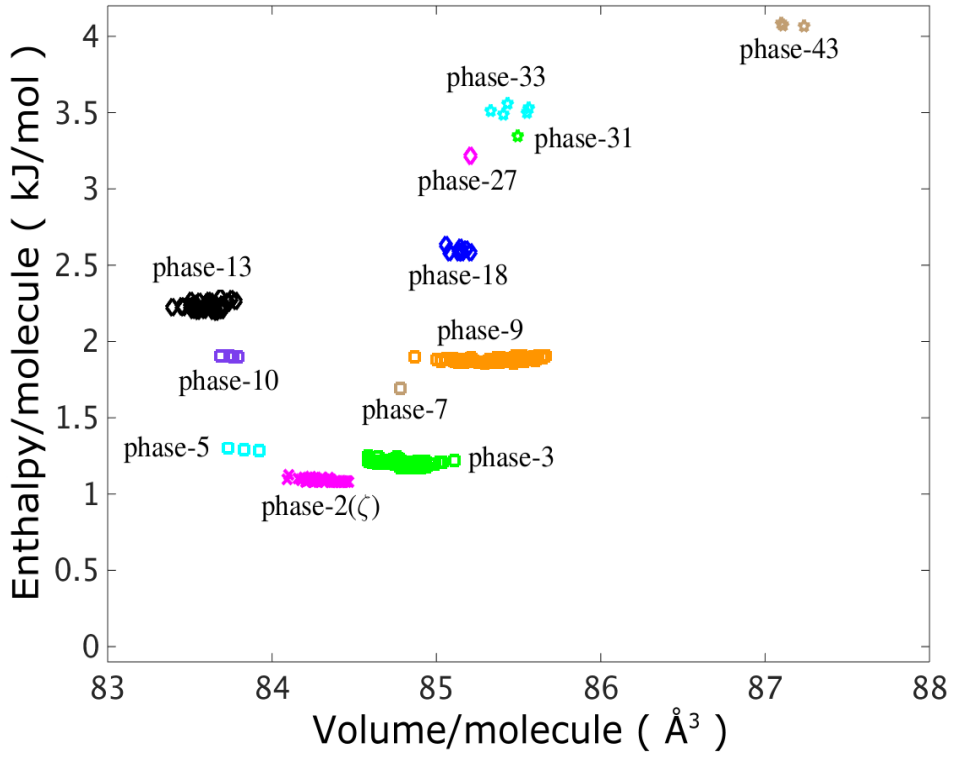

Figure A.3: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 3 molecules/cell. For the structures generated randomly at each generation, the space group is selected according to the frequency distribution appearing in known organic crystal structure database [$P2_1/c$ (36.59 %), $P\bar{1}$ (16.92 %), $P2_12_12_1$ (11.00 %), $C2/c$ (6.95 %), $P2_1$ (6.35 %), $Pbca$ (4.24 %), and uniform otherwise]. The simulation is run for 20 generations with settings described in Section 3.2. The clusters are labeled according to Table 3.5. This simulation fails to find the most stable structure compatible with Z=3, γ-glycine, while structures with a much wider variety are found with respect to the simulation where all space groups were sampled with equal probability (see Fig. A.2).

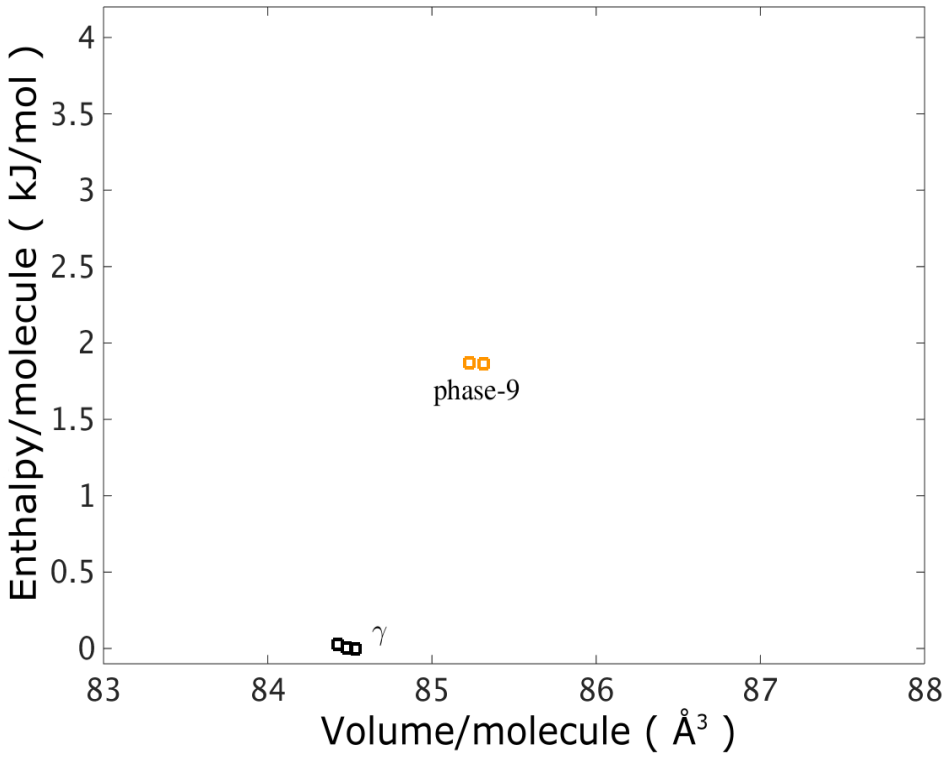

Figure A.4: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 3 molecules/cell. For the structures generated randomly at each generation, the space group are selected according to the frequency distribution appearing in known organic crystal structure database [$P2_1/c$ (36.59 %), $P\bar{1}$ (16.92 %), $P2_12_12_1$ (11.00 %), $C2/c$ (6.95 %), $P2_1$ (6.35 %), $Pbca$ (4.24 %), and uniform otherwise]. The simulation finds γ-glycine at the first generation by random generation and is stopped subsequently at the second generation. The clusters are labeled according to Table 3.5.

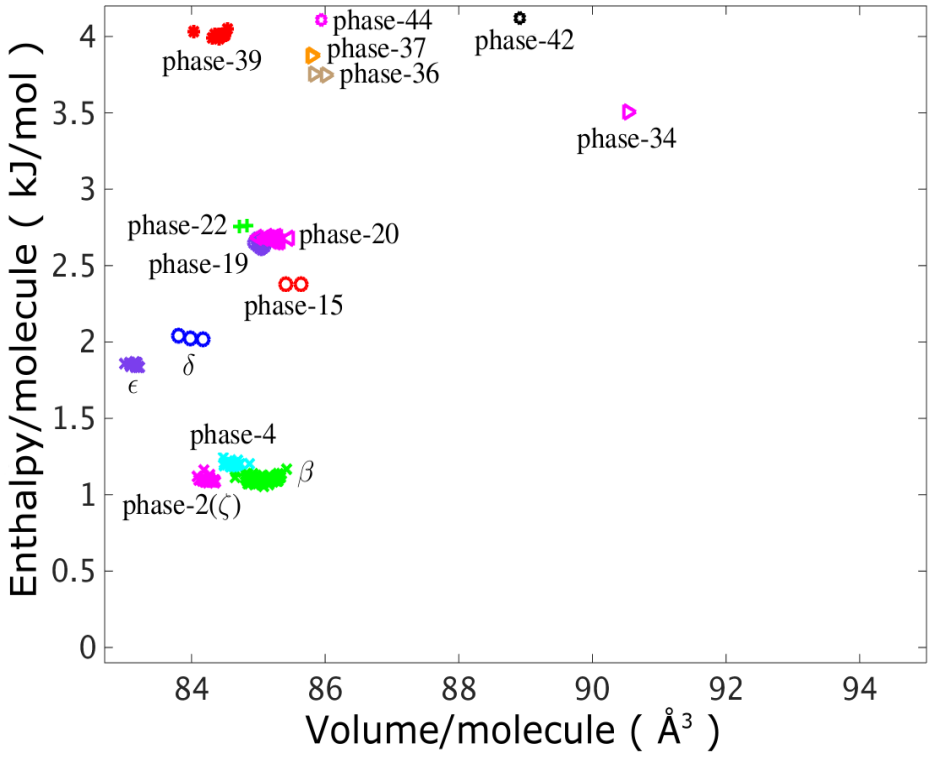

Figure A.5: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 4 molecules/cell without using any experimental information. The simulation is run for 20 generations with settings described in Section 3.2. The clusters are labeled according to Table 3.5. This simulation fails to find the most stable structure compatible with Z=4, α glycine.

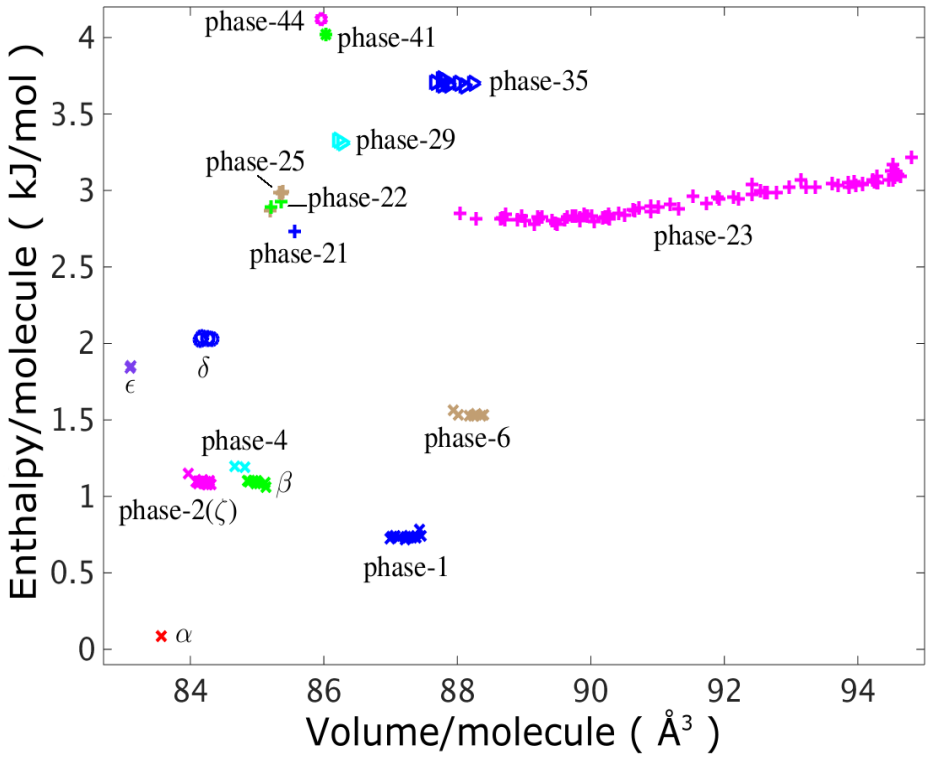

Figure A.6: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 4 molecules/cell. For the structures generated randomly at each generation, the space group are chosen according to the frequency distribution appearing in known organic crystal structure database [$P2_1/c$ (36.59 %), $P\bar{1}$ (16.92 %), $P2_12_12_1$ (11.00 %), $C2/c$ (6.95 %), $P2_1$ (6.35 %), $Pbca$ (4.24 %), and uniform otherwise]. The simulation is run for 16 generations with settings described in Section 3.2. The clusters are labeled according to Table 3.5.

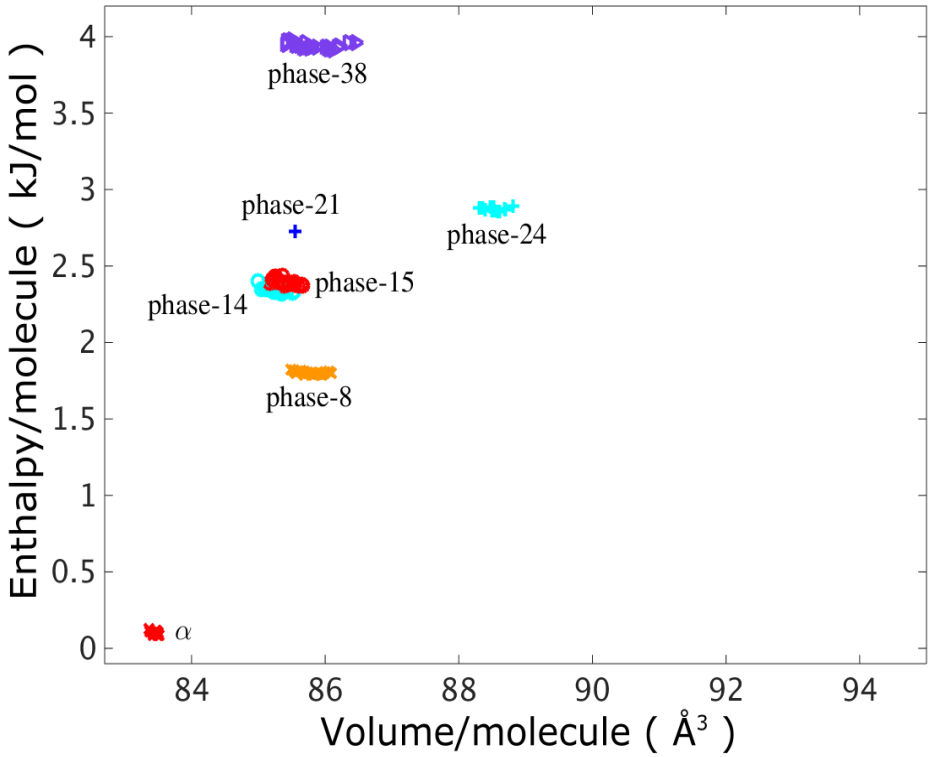

Figure A.7: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 4 molecules/cell using glycine dimer as a building block, instead of the single molecule. The simulation is run for 7 generations with settings described in Section 3.2. The clusters are labeled according to Table 3.5. Note that some of these structures (phase 15, 24) were also encountered in other simulations for Z=4. Hence both the vastness of the phase space and poor connectivity between valleys around the α phase can contribute to the challenge of finding this phase in evolutionary search.

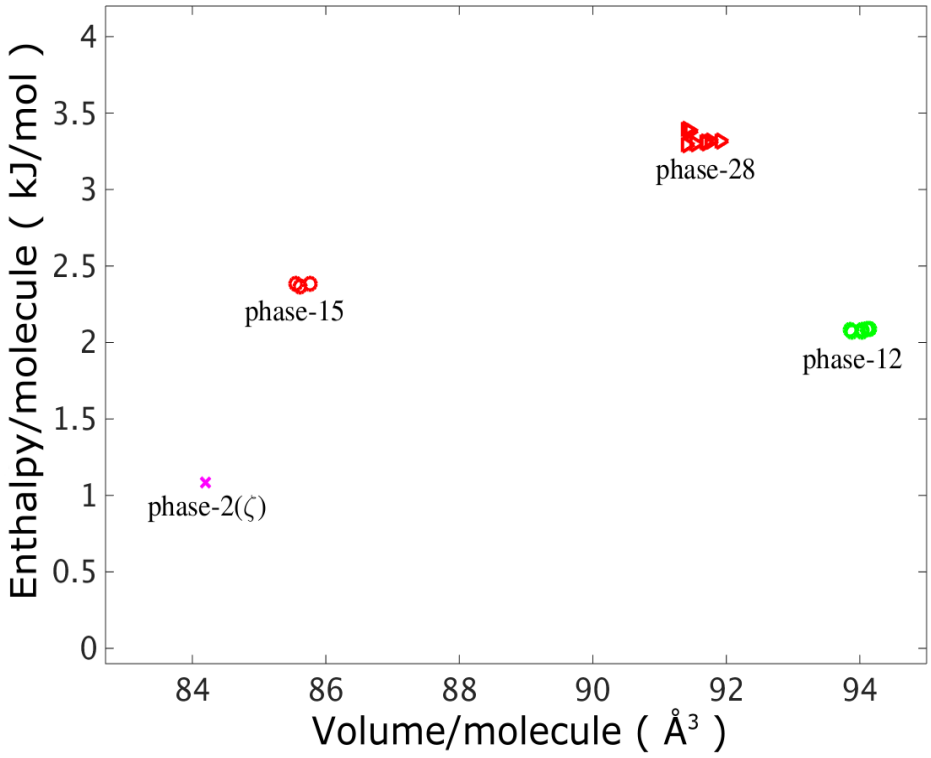

Figure A.8: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 4 molecules/cell without using any experimental information but with modified criterion for the selection of suitable parents: As described in main text, once the energetically unfavorable structures are discarded from the list of potential parents, a fingerprint analysis is performed for the remaining candidates. The potential parents whose fingerprint is within a threshold distance of 0.01 from any lower energy structure are discarded as well. In this simulation, we double this threshold value at each generation when one of the low energy structures is found to be a clone from the previous generations. This simple modification allows to dynamically depopulate the enthalpy valley of the best parents found in the simulation while still keeping them in the list of eligible parents for the next generation. The simulation is run for 5 generations with settings described in Section 3.2, for the rest of the parameters. The clusters are labeled according to Table 3.5.

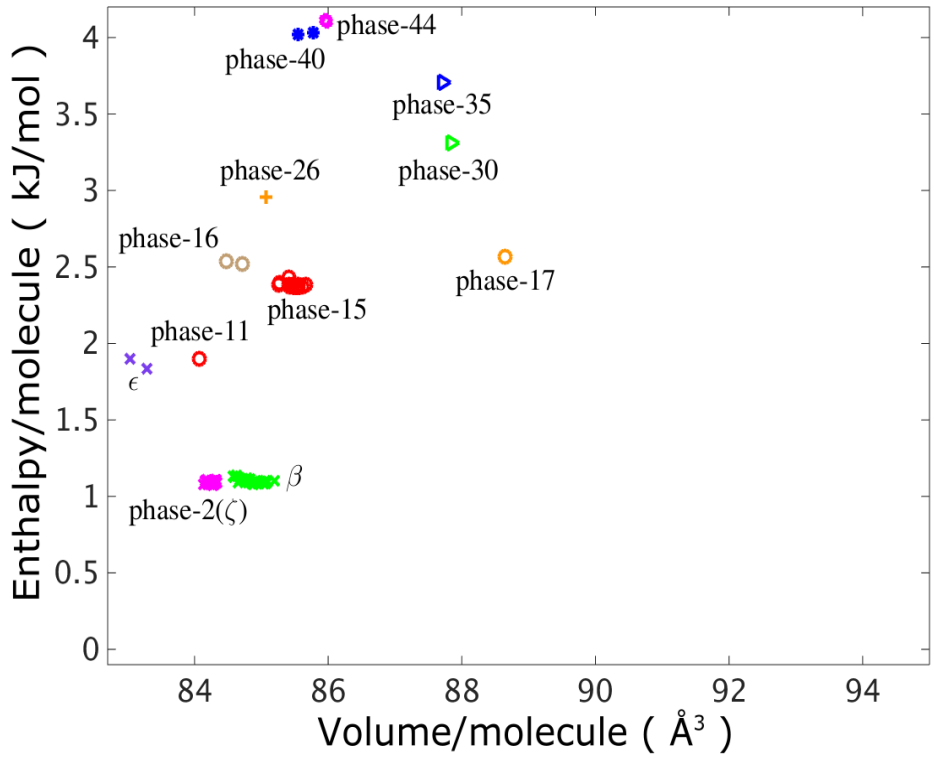

Figure A.9: Enthalpy as a function of volume for ab initio crystal structure search of Glycine in the case of 4 molecules/cell. The cell parameters of the randomly generated structures are taken from the experimental structure as defined by the non-standard space group $P2_1/n$. The simulation is run for 14 generations with settings described in Section 3.2, for the rest of the parameters. The clusters are labeled according to Table 3.5.

Bibliography

[1] Sarah L Price. Predicting crystal structures of organic compounds. *Chemical Society Reviews*, 43(7):2098–2111, 2014. (Cited on pages 7 and 42.)

[2] Jun Yang, Weifeng Hu, Denis Usvyat, Devin Matthews, Martin Schütz, and Garnet Kin-Lic Chan. Ab initio determination of the crystalline benzene lattice energy to sub-kilojoule/mole accuracy. *Science*, 345(6197):640–643, 2014. (Cited on page 7.)

[3] Sarah L Price. The computational prediction of pharmaceutical crystal structures and polymorphism. *Advanced drug delivery reviews*, 56(3):301–319, 2004. (Cited on pages 7 and 23.)

[4] Jos PM Lommerse, WD Sam Motherwell, Herman L Ammon, Jack D Dunitz, Angelo Gavezzotti, Detlef WM Hofmann, Frank JJ Leusen, Wijnand TM Mooij, Sarah L Price, Bernd Schweizer, et al. A test of crystal structure prediction of small organic molecules. *Acta Crystallographica Section B: Structural Science*, 56(4):697–714, 2000. (Cited on pages 7 and 28.)

[5] WD Sam Motherwell, Herman L Ammon, Jack D Dunitz, Alexander Dzyabchenko, Peter Erk, Angelo Gavezzotti, Detlef WM Hofmann, Frank JJ Leusen, Jos PM Lommerse, Wijnand TM Mooij, et al. Crystal structure prediction of small organic molecules: a second blind test. *Acta Crystallographica Section B: Structural Science*, 58(4):647–661, 2002. (Cited on pages 7 and 28.)

[6] GM Day, WDS Motherwell, HL Ammon, SXM Boerrigter, RG Della Valle, E Venuti, A Dzyabchenko, JD Dunitz, B Schweizer, BP Van Eijck, et al. A third blind test of crystal structure prediction. *Acta Crystallographica Section B: Structural Science*, 61(5):511–527, 2005. (Cited on pages 7 and 28.)

[7] Graeme M Day, Timothy G Cooper, Aurora J Cruz-Cabeza, Katarzyna E Hejczyk, Herman L Ammon, Stephan XM Boerrigter, Jeffrey S Tan, Raffaele G Della Valle, Elisabetta Venuti, Jovan Jose, et al. Significant progress

in predicting the crystal structures of small organic molecules–a report on the fourth blind test. *Acta Crystallographica Section B: Structural Science*, 65(2):107–125, 2009. (Cited on pages 7 and 28.)

[8] David A Bardwell, Claire S Adjiman, Yelena A Arnautova, Ekaterina Bartashevich, Stephan XM Boerrigter, Doris E Braun, Aurora J Cruz-Cabeza, Graeme M Day, Raffaele G Della Valle, Gautam R Desiraju, et al. Towards crystal structure prediction of complex organic compounds–a report on the fifth blind test. *Acta Crystallographica Section B: Structural Science*, 67(6):535–551, 2011. (Cited on pages 7 and 28.)

[9] Qiang Zhu, Artem R Oganov, Colin W Glass, and Harold T Stokes. Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. *Acta Crystallographica Section B: Structural Science*, 68(3):215–226, 2012. (Cited on pages 7, 8, 24, 25, 26, 27, 29, 30, 35 and 46.)

[10] Max Dion, Henrik Rydberg, Elsebeth Schröder, David C Langreth, and Bengt I Lundqvist. Van der waals density functional for general geometries. *Physical review letters*, 92(24):246401, 2004. (Cited on pages 8, 15, 28, 36 and 68.)

[11] Oleg A Vydrov and Troy Van Voorhis. Nonlocal van der waals density functional: The simpler the better. *The Journal of chemical physics*, 133(24):244103, 2010. (Cited on pages 8, 16 and 37.)

[12] Riccardo Sabatini, Emine Küçükbenli, Brian Kolb, Timo Thonhauser, and Stefano de Gironcoli. Structural evolution of amino acid crystals under stress from a non-empirical density functional. *Journal of Physics: Condensed Matter*, 24(42):424209, 2012. (Cited on pages 8, 28 and 62.)

[13] Artem R Oganov and Colin W Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics*, 124(24):244704, 2006. (Cited on pages 8, 22 and 46.)

[14] Elena V Boldyreva, Svetlana N Ivashevskaya, Heidrun Sowa, Hans Ahsbahs, and Hans-Peter Weber. Effect of hydrostatic pressure on the γ-polymorph of glycine. 1. a polymorphic transition into a new δ-form. *Zeitschrift für Kristallographie/International journal for structural, physical, and chemical aspects of crystalline materials*, 220(1/2005):50–57, 2005. (Cited on pages 8, 33, 35 and 47.)

[15] K Jayalakshmi, Kanchan Sonkar, Anu Behari, VK Kapoor, and Neeraj Sinha. Solid state 13 c nmr analysis of human gallstones from cancer and benign gall bladder diseases. *Solid state nuclear magnetic resonance*, 36(1):60–65, 2009. (Cited on pages 8, 59 and 75.)

[16] Zoe Cournia, Jeremy C Smith, and G Matthias Ullmann. A molecular mechanics force field for biologically important sterols. *Journal of computational chemistry*, 26(13):1383–1399, 2005. (Cited on pages 8, 28, 58, 61 and 66.)

[17] Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927. (Cited on page 11.)

[18] Richard M Martin. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2004. (Cited on pages 11, 12 and 13.)

[19] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical review*, 136(3B):B864, 1964. (Cited on page 12.)

[20] Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133, 1965. (Cited on page 13.)

[21] David M Ceperley and BJ Alder. Ground state of the electron gas by a stochastic method. *Physical Review Letters*, 45(7):566, 1980. (Cited on page 14.)

[22] John P Perdew and Yue Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Physical Review B*, 45(23):13244, 1992. (Cited on page 14.)

[23] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996. (Cited on page 14.)

[24] John F Dobson, Keith McLennan, Angel Rubio, Jun Wang, Tim Gould, Hung M Le, and Bradley P Dinte. Prediction of dispersion forces: is there a problem? *Australian Journal of Chemistry*, 54(8):513–527, 2002. (Cited on page 15.)

[25] Huy-Viet Nguyen and Stefano de Gironcoli. Efficient calculation of exact exchange and rpa correlation energies in the adiabatic-connection fluctuation-dissipation theory. *Physical Review B*, 79(20):205114, 2009. (Cited on page 15.)

[26] Ngoc Linh Nguyen, Nicola Colonna, and Stefano de Gironcoli. Ab initio self-consistent total-energy calculations within the exx/rpa formalism. *Physical Review B*, 90(4):045138, 2014. (Cited on page 15.)

[27] Nicola Colonna, Maria Hellgren, and Stefano de Gironcoli. Correlation energy within exact-exchange adiabatic connection fluctuation-dissipation theory: Systematic development and simple approximations. *Physical Review B*, 90(12):125150, 2014. (Cited on page 15.)

[28] Stefan Grimme. Semiempirical gga-type density functional constructed with a long-range dispersion correction. *Journal of computational chemistry*, 27(15):1787–1799, 2006. (Cited on pages 15 and 37.)

[29] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15):154104, 2010. (Cited on page 15.)

[30] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Physical review letters*, 102(7):073005, 2009. (Cited on page 15.)

[31] Guillermo Román-Pérez and José M Soler. Efficient implementation of a van der waals density functional: application to double-wall carbon nanotubes. *Physical review letters*, 103(9):096102, 2009. (Cited on page 15.)

[32] Timo Thonhauser, Valentino R Cooper, Shen Li, Aaron Puzder, Per Hyldgaard, and David C Langreth. Van der waals density functional: Self-consistent potential and the nature of the van der waals bond. *Physical Review B*, 76(12):125112, 2007. (Cited on pages 16, 28, 36, 37 and 68.)

[33] Jiří Klimeš, David R Bowler, and Angelos Michaelides. Chemical accuracy for the van der waals density functional. *Journal of Physics: Condensed Matter*, 22(2):022201, 2010. (Cited on page 16.)

[34] Valentino R Cooper. Van der waals density functional: An appropriate exchange functional. *Physical Review B*, 81(16):161104, 2010. (Cited on page 16.)

[35] Kristian Berland and Per Hyldgaard. Exchange functional that tests the robustness of the plasmon description of the van der waals density functional. *Physical Review B*, 89(3):035412, 2014. (Cited on page 16.)

[36] Kyuho Lee, Éamonn D Murray, Lingzhu Kong, Bengt I Lundqvist, and David C Langreth. Higher-accuracy van der waals density functional. *Physical Review B*, 82(8):081101, 2010. (Cited on page 16.)

[37] Riccardo Sabatini, Tommaso Gorni, and Stefano de Gironcoli. Nonlocal van der waals density functional made simple and efficient. *Physical Review B*, 87(4):041108, 2013. (Cited on pages 16 and 37.)

[38] Giacomo Miceli, Stefano de Gironcoli, and Alfredo Pasquarello. Isobaric first-principles molecular dynamics of liquid water with nonlocal van der waals interactions. *The Journal of chemical physics*, 142(3):034501, 2015. (Cited on pages 16 and 37.)

[39] Peter E Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953, 1994. (Cited on pages 16 and 17.)

[40] Chris J Pickard and Francesco Mauri. All-electron magnetic response with pseudopotentials: Nmr chemical shifts. *Physical Review B*, 63(24):245101, 2001. (Cited on pages 16 and 19.)

[41] Jonathan R Yates, Chris J Pickard, and Francesco Mauri. Calculation of nmr chemical shifts for extended systems using ultrasoft pseudopotentials. *Physical Review B*, 76(2):024401, 2007. (Cited on pages 16 and 19.)

[42] DR Hamann, M Schlüter, and C Chiang. Norm-conserving pseudopotentials. *Physical Review Letters*, 43(20):1494, 1979. (Cited on page 17.)

[43] JCp Slater. Wave functions in a periodic potential. *Physical Review*, 51(10):846, 1937. (Cited on page 17.)

[44] Paul M Marcus. Variational methods in the computation of energy bands. *International Journal of Quantum Chemistry*, 1(S1):567–588, 1967. (Cited on page 17.)

[45] John Maddox. Crystals from first principles. *Nature*, 335(6187), 1988. (Cited on page 22.)

[46] Angelo Gavezzotti. Are crystal structures predictable? *Accounts of chemical research*, 27(10):309–314, 1994. (Cited on page 22.)

[47] Stefano Curtarolo, Dane Morgan, Kristin Persson, John Rodgers, and Gerbrand Ceder. Predicting crystal structures with data mining of quantum calculations. *Physical review letters*, 91(13):135503, 2003. (Cited on page 22.)

[48] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002. (Cited on pages 22 and 31.)

[49] R Martoňák, Alessandro Laio, and Michele Parrinello. Predicting crystal structures: the parrinello-rahman method revisited. *Physical review letters*, 90(7):075503, 2003. (Cited on pages 22 and 31.)

[50] David J Wales and Jonathan PK Doye. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, 1997. (Cited on pages 22 and 31.)

[51] J Pannetier, J Bassas-Alsina, J Rodriguez-Carvajal, and V Caignaert. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature*, 346(6282):343–345, 1990. (Cited on pages 22 and 31.)

[52] J Christian Schön and Martin Jansen. First step towards planning of syntheses in solid-state chemistry: Determination of promising structure candidates by global optimization. *Angewandte Chemie International Edition in English*, 35(12):1286–1304, 1996. (Cited on pages 22 and 31.)

[53] Stefan Goedecker. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *The Journal of chemical physics*, 120(21):9911–9917, 2004. (Cited on pages 22 and 31.)

[54] Chris J Pickard and RJ Needs. Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011. (Cited on pages 23 and 28.)

[55] IA Baburin, Stefano Leoni, and G Seifert. Enumeration of not-yet-synthesized zeolitic zinc imidazolate mof networks: a topological and dft approach. *The Journal of Physical Chemistry B*, 112(31):9437–9443, 2008. (Cited on page 23.)

[56] R Martoňák, Artem R Oganov, and CW Glass. Crystal structure prediction and simulations of structural transformations: metadynamics and evolutionary algorithms. *Phase Transitions*, 80(4-5):277–298, 2007. (Cited on pages 24 and 31.)

[57] Roman Martoňák, Alessandro Laio, Marco Bernasconi, Chiara Ceriani, Paolo Raiteri, Federico Zipoli, and Michele Parrinello. Simulation of structural phase transitions by metadynamics. *Zeitschrift für Kristallographie*, 220(5/6/2005):489–498, 2005. (Cited on page 24.)

[58] Maximilian Amsler, José A Flores-Livas, Tran Doan Huan, Silvana Botti, Miguel AL Marques, and Stefan Goedecker. Novel structural motifs in low energy phases of lialh 4. *Physical review letters*, 108(20):205505, 2012. (Cited on pages 24 and 25.)

[59] Tran Doan Huan, Maximilian Amsler, Miguel AL Marques, Silvana Botti, Alexander Willand, and Stefan Goedecker. Low-energy polymeric phases of alanates. *Physical review letters*, 110(13):135502, 2013. (Cited on pages 24 and 25.)

[60] Albert M Lund, Gabriel I Pagola, Anita M Orendt, Marta B Ferraro, and Julio C Facelli. Crystal structure prediction from first principles: The crystal structures of glycine. *Chemical physics letters*, 626:20–24, 2015. (Cited on pages 24, 25, 35 and 42.)

[61] Marcus A Neumann. Tailor-made force fields for crystal-structure prediction. *The Journal of Physical Chemistry B*, 112(32):9810–9829, 2008. (Cited on pages 25, 26 and 28.)

[62] Seonah Kim, Anita M Orendt, Marta B Ferraro, and Julio C Facelli. Crystal structure prediction of flexible molecules using parallel genetic algorithms with a standard force field. *Journal of computational chemistry*, 30(13):1973–1985, 2009. (Cited on pages 25 and 28.)

[63] Hyoungki Park, Michael R Fellinger, Thomas J Lenosky, William W Tipton, Dallas R Trinkle, Sven P Rudin, Christopher Woodward, John W Wilkins, and Richard G Hennig. Ab initio based empirical potential used to study the mechanical properties of molybdenum. *Physical Review B*, 85(21):214121, 2012. (Cited on page 25.)

[64] S Brodersen, S Wilke, FJJ Leusen, and G Engel. A study of different approaches to the electrostatic interaction in force field methods for organic crystals. *Physical Chemistry Chemical Physics*, 5(21):4923–4931, 2003. (Cited on page 25.)

[65] Andriy O Lyakhov, Artem R Oganov, and Mario Valle. How to predict very large and complex crystal structures. *Computer Physics Communications*, 181(9):1623–1632, 2010. (Cited on page 27.)

[66] Aurora J Cruz Cabeza, Elna Pidcock, Graeme M Day, WD Sam Motherwell, and William Jones. Space group selection for crystal structure prediction of solvates. *CrystEngComm*, 9(7):556–560, 2007. (Cited on page 27.)

[67] Alston J Misquitta, Gareth WA Welch, Anthony J Stone, and Sarah L Price. A first principles prediction of the crystal structure of. *Chemical Physics Letters*, 456(1):105–109, 2008. (Cited on page 28.)

[68] Cong Huy Pham, Emine Küçükbenli, and Stefano de Gironcoli. Crystal structure prediction of molecular crystals from first principles: Are we there yet? *submitted*, 2015. (Cited on page 29.)

[69] Artem R Oganov and Mario Valle. How to quantify energy landscapes of solids. *The Journal of chemical physics*, 130(10):104504, 2009. (Cited on pages 29 and 46.)

[70] Mario Valle and Artem R Oganov. Crystal fingerprint space-a novel paradigm for studying crystal-structure sets. *Acta Crystallographica Section A: Foundations of Crystallography*, 66(5):507–517, 2010. (Cited on pages 29 and 46.)

[71] Andriy O Lyakhov, Artem R Oganov, and Mario Valle. Crystal structure prediction using evolutionary approach. *Modern Methods of Crystal Structure Prediction*, pages 147–180, 2010. (Cited on page 30.)

[72] Anubhav Jain, Ivano E Castelli, Geoffroy Hautier, David H Bailey, and Karsten W Jacobsen. Performance of genetic algorithms in search for water splitting perovskites. *Journal of Materials Science*, 48(19):6519–6534, 2013. (Cited on page 31.)

[73] Sandro E Schönborn, Stefan Goedecker, Shantanu Roy, and Artem R Oganov. The performance of minima hopping and evolutionary algorithms for cluster

structure prediction. *The Journal of chemical physics*, 130(14):144108, 2009. (Cited on page 31.)

[74] Min Ji, Cai-Zhuang Wang, and Kai-Ming Ho. Comparing efficiencies of genetic and minima hopping algorithms for crystal structure prediction. *Physical Chemistry Chemical Physics*, 12(37):11617–11623, 2010. (Cited on page 31.)

[75] VA Drebushchak, EV Boldyreva, Yu A Kovalevskaya, IE Paukov, and TN Drebushchak. Low-temperature heat capacity of β-glycine and a phase transition at 252 k. *Journal of thermal analysis and calorimetry*, 79(1):65–70, 2005. (Cited on pages 33 and 35.)

[76] NV Surovtsev, SV Adichtchev, VK Malinovsky, AG Ogienko, VA Drebushchak, A Yu Manakov, AI Ancharov, AS Yunoshev, and EV Boldyreva. Glycine phases formed from frozen aqueous solutions: Revisited. *The Journal of chemical physics*, 137(6):065103, 2012. (Cited on pages 33 and 35.)

[77] GL Perlovich, L Kr Hansen, and A Bauer-Brandl. The polymorphism of glycine. thermochemical and structural aspects. *Journal of thermal analysis and calorimetry*, 66(3):699–715, 2001. (Cited on pages 33 and 39.)

[78] Lian Yu and Kingman Ng. Glycine crystallization during spray drying: the ph effect on salt and polymorphic forms. *Journal of pharmaceutical sciences*, 91(11):2367–2375, 2002. (Cited on page 33.)

[79] Isabelle Weissbuch, Leslie Leisorowitz, and Meir Lahav. "tailor-made" and charge-transfer auxiliaries for the control of the crystal polymorphism of glycine. *Advanced Materials*, 6(12):952–956, 1994. (Cited on page 34.)

[80] Elena S Ferrari, Roger J Davey, Wendy I Cross, Amy L Gillon, and Christopher S Towler. Crystallization in polymorphic systems: the solution-mediated transformation of β to α glycine. *Crystal growth & design*, 3(1):53–60, 2003. (Cited on page 34.)

[81] E Boldyreva, V Drebushchak, T Drebushchak, I Paukov, Y Kovalevskaya, and E Shutova. Polymorphism of glycine, part ii. *Journal of thermal analysis and calorimetry*, 73(2):419–428, 2003. (Cited on page 34.)

[82] James A Chisholm, Sam Motherwell, Paul R Tulip, Simon Parsons, and Stewart J Clark. An ab initio study of observed and hypothetical polymorphs of glycine. *Crystal growth & design*, 5(4):1437–1442, 2005. (Cited on page 35.)

[83] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502, 2009. (Cited on pages 36 and 67.)

[84] Andrea Dal Corso. Pseudopotentials periodic table: From h to pu. *Computational Materials Science*, 95:337–350, 2014. (Cited on page 36.)

[85] VM Kozhin. Tensors of thermal expansion of α-, β-, and γ-polymorphs of glycine. *Kristallografiya*, 23:1211–1215, 1978. (Cited on pages 38 and 39.)

[86] SV Goryainov, EN Kolesnik, and EV Boldyreva. A reversible pressure-induced phase transition in β-glycine at 0.76 gpa. *Physica B: Condensed Matter*, 357(3):340–347, 2005. (Cited on page 40.)

[87] Alice Dawson, David R Allan, Scott A Belmonte, Stewart J Clark, William IF David, Pamela A McGregor, Simon Parsons, Colin R Pulham, and Lindsay Sawyer. Effect of high pressure on the crystal structures of polymorphs of glycine. *Crystal growth & design*, 5(4):1415–1427, 2005. (Cited on page 40.)

[88] George C Feast, James Haestier, Lee W Page, Jeremy Robertson, Amber L Thompson, and David J Watkin. An unusual methylene aziridine refined in p21/c and the nonstandard setting p21/n. *Acta Crystallographica Section C: Crystal Structure Communications*, 65(12):o635–o638, 2009. (Cited on page 48.)

[89] Isabelle Weissbuch, Meir Lahav, and Leslie Leiserowitz. Toward stereochemical control, monitoring, and understanding of crystal nucleation. *Crystal growth & design*, 3(2):125–150, 2003. (Cited on page 49.)

[90] Bryan M Craven. Crystal structure of cholesterol monohydrate. 1976. (Cited on page 57.)

[91] HS Shieh, LG Hoard, and CE Nordman. Crystal structure of anhydrous cholesterol. 1977. (Cited on pages 57, 60 and 74.)

[92] H-S Shieh, LG Hoard, and CE Nordman. The structure of cholesterol. *Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry*, 37(8):1538–1543, 1981. (Cited on pages 57, 60 and 74.)

[93] Leh-Yeh Hsu and CE Nordman. Phase transition and crystal structure of the 37 degrees c form of cholesterol. *Science*, 220(4597):604–606, 1983. (Cited on page 57.)

[94] L-Y Hsu, Jeff W Kampf, and Christer E Nordman. Structure and pseudosymmetry of cholesterol at 310 k. *Acta Crystallographica Section B: Structural Science*, 58(2):260–264, 2002. (Cited on pages 57 and 59.)

[95] M Crina Frincu, Sean D Fleming, Andrew L Rohl, and Jennifer A Swift. The epitaxial growth of cholesterol crystals from bile solutions on calcite substrates. *Journal of the American Chemical Society*, 126(25):7915–7924, 2004. (Cited on pages 58 and 60.)

[96] Ulrich Sternberg, Frank-Thomas Koch, Wolfram Prieß, and Raiker Witter. Crystal structure refinements of cellulose polymorphs using solid state 13c chemical shifts. *Cellulose*, 10(3):189–199, 2003. (Cited on page 61.)

[97] Jonathan R Yates, Sara E Dobbins, Chris J Pickard, Francesco Mauri, Phuong Y Ghi, and Robin K Harris. A combined first principles computational and solid-state nmr study of a molecular crystal: flurbiprofen. *Physical Chemistry Chemical Physics*, 7(7):1402–1407, 2005. (Cited on page 61.)

[98] Elodie Salager, Robin S Stein, Chris J Pickard, Bénédicte Elena, and Lyndon Emsley. Powder nmr crystallography of thymol. *Physical Chemistry Chemical Physics*, 11(15):2610–2621, 2009. (Cited on page 61.)

[99] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of computational physics*, 117(1):1–19, 1995. (Cited on page 66.)

[100] Emine Küçükbenli, Kanchan Sonkar, Neeraj Sinha, and Stefano de Gironcoli. Complete ^{13}c nmr chemical shifts assignment for cholesterol crystals by combined cp-mas spectral editing and ab initio gipaw calculations with dispersion forces. *The Journal of Physical Chemistry A*, 116(14):3765–3769, 2012. (Cited on page 67.)

[101] Emine Küçükbenli. Phd thesis: Study of complex molecular crystals from first principles: Case of cholesterol. 2011. (Cited on page 67.)