ENHANCEMENTS TO THE MODIFIED GENETIC ALGORITHM FOR CRYSTAL STRUCTURE PREDICTION

by

Albert Merrill Lund

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry

The University of Utah

May 2016

Copyright © Albert Merrill Lund 2015

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Albert Merrill Lund**

has been approved by the following supervisory committee members:

Thanh Truong	, Chair	**1/21/2016** Date Approved
Peter Francis Flynn	, Member	**1/21/2016** Date Approved
Haitao Ji	, Member	**1/21/2016** Date Approved
Jon D Ranier	, Member	**1/21/2016** Date Approved
Julio Cesar Facelli	, Member	**1/22/2016** Date Approved

and by **Cynthia Burrows**, Chair/Dean of the Department/College/School of **Chemistry**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Crystal structure prediction is an important field of study, both for the development of new compounds and materials, and for the advancement of understanding crystallization processes. The Modified Genetic Algorithm for Crystal Structure Prediction, MGAC, is a software package for structure prediction that has had varying success in predicting the structures of many molecules. However, several advancements in the field of structure prediction have prompted a revision to the software, both from a scientific and technical standpoint.

In this dissertation, the evaluation of a new method for energy calculation and structural optimization, dispersion corrected density functional theory, is presented, along with practical parameterizations for using density functional theory in crystal structure prediction. Next, a preliminary implementation of MGAC using density functional theory is outlined, including some key changes to the construction of unit cells, along with successful prediction results for the molecules glycine and histamine. Finally, a new implementation of MGAC is proposed to handle multiple space group prediction effectively, with accompanying preliminary prediction results for histamine using the new implementation of MGAC, called MGAC2.

TABLE OF CONTENTS

ABSTRACT iii

LIST OF TABLES vi

GLOSSARY OF TERMS vii

ACKNOWLEDGEMENTS ix

Chapters

1 INTRODUCTION 1

Solving the CSP Problem 2
Accuracy in Energetic Calculations 5
Sampling the Search Space 7
Automation Concerns 11
The Modified Genetic Algorithm for Crystals (MGAC) 11
The MGAC1 Schema 14
Volume Filtering 16
Practical Considerations for CSP with MGAC 17
Concluding Remarks 18

2 OPTIMIZATION OF CRYSTAL STRUCTURES OF ARCHETYPICAL PHARMACEUTICAL COMPOUNDS: A PLANE WAVE DFT-D STUDY USING QUANTUM ESPRESSO 20

Abstract 20
Introduction 21
Methods 24
Results and Discussion 26
Conclusion 40

3 AN IMPROVED METHOD FOR BUILDING CRYSTAL STRUCTURES FROM GENETIC SCHEMA IN CRYSTAL STRUCTURE PREDICTION 41

MGAC1 Fitcell 44
MGAC2 Fitcell 48

4 CRYSTAL STRUCTURE PREDICTION FROM FIRST PRINCIPLES: THE CRYSTAL STRUCTURES OF GLYCINE 53

Abstract 53
Introduction 54
Methods 58
Results and Discussion 60
Conclusions 67

5 DEPOSITION OF HISTAMINE PREDICTION RESULTS 69

Methods 69
Results 69
Conclusion 70

6 MGAC2: A NEW ALGORITHM FOR CRYSTAL STRUCTURE PREDICTION 74

Abstract 74
Introduction 74
Multiple Space Group Schema 77
Genetic Algorithm Refinements 83
Computational Changes 93
Steady-State Algorithm 97
Conclusion 98

7 DISSEMINATION OF HISTAMINE RESULTS USING THE MGAC2 ALGORITHM 102

Methods 102
Results 102

REFERENCES 109

LIST OF TABLES

2.1 Parameters explored to determine the optimal conditions for QE optimizations 27

2.2 Summary of QE results for the K & P dataset 30

2.3 Results of reranking MGAC-CHARMM lowest energy structures using QE optimization 35

4.1 Comparison of the energies and geometries of the α-glycine, β-glycine and γ-glycine structures found by MGAC-QE with the reference experimental structures 63

5.1 Parameters for the predicted structures compared against the experimentally determined histamine structure 72

7.1 Cell parameters for the MGAC2 histamine prediction 104

GLOSSARY OF TERMS

CCDC – Cambridge Crystallographic Data Centre

CHARMM – Chemistry at HARvard Molecular Mechanics.

Compute node – A computer server in a computing cluster or supercomputer that primarily performs calculations.

Core hour – A measurement of computer resources. One core hour means that a core in a CPU is occupied by calculations or other operations for the span of one hour.

Crossover – The mixing of genes in a genetic algorithm, analogous to genetic recombination in sexual reproduction.

CSP – Crystal structure prediction

DFT-D – Dispersion corrected density functional theory

Fitcell – An algorithm in MGAC that minimizes the volume of the unit cell given an arbitrary set of unit cell parameters, molecular positions, rotations, and internal flexibility.

GAFF – General Amber Force Field

Generation – All of the individuals in the population in one iteration of a genetic algorithm.

Genetic Algorithm (GA) – A population based optimization algorithm that uses natural selection and genetic recombination to find global minima.

Glide plane – A compound symmetry operation that combines a reflection with a translation along the plane of reflection.

ITC – International Tables of Crystallography

K-P dataset – Karamertzanis-Price dataset

MGAC – Modified Genetic Algorithm for Crystals

MGAC1, MGAC1-CHARMM – The first version of MGAC that relied on the molecular mechanics package CHARMM for energy calculations.

MGAC1-QE – An update to MGAC to use the DFT-D based software package Quantum Espresso.

MGAC2 – An updated version of MGAC that improved on several deficiencies of the original algorithm and implementation.

MPI – Message Passing Interface, used in parallel computing to establish communication between compute nodes.

Mutation – The random modification of the genome of an individual in a genetic algorithm.

Polymorphism – A phenomena exhibited by some molecules where different crystallization methods result in different three dimensional structures.

Precluster – A step in the MGAC2 algorithm where a representative set of structures is generated in the initialization step through the use of clustering techniques.

Pseudopotential – A potential energy function representing the core electrons in plane-wave based DFT-D.

QE – Quantum Espresso, a plane-wave based DFT-D solver.

Replacement – The set of individuals generated in a new generation in a genetic algorithm.

Roulette wheel – A selection method in genetic algorithms where the fitness of an individual is proportional to the probability that individual will be selected for breeding.

Schema – The representation of independent physical parameters as a genome in a genetic algorithm.

Screw axis – A compound symmetry operation that combines a rotation with a translation along the axis of rotation.

Supercell – A crystal structure comprising more than one unit cell.

Unit cell – The smallest, translatable repeating volume unit in a solid crystalline material.

Volume filter – A filter in MGAC that restricts the candidate structures for optimization and evaluation to certain range of volumes based on an estimate of the single molecule volume of the crystal structure.

Z – The number of molecules in a unit cell.

Z' – The number of molecules in an asymmetric unit cell.

ACKNOWLEDGEMENTS

Financial support and computer resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

Special thanks to Thanh Truong for acting as head of my committee and the Department of Chemistry for permitting my dissertation work under Julio Facelli. Extra special thanks to Julio for agreeing to be my advisor.

CHAPTER 1

INTRODUCTION

Crystal structure prediction (CSP) is the elucidation of the solid state crystalline structure of an arbitrary molecule using nothing but the knowledge of the chemical diagram of that molecule and first-principles or semiempirical calculations (Price, 2004; Day, 2011). CSP is generally understood to be the prediction of small organic molecules and solid state materials, as opposed to the field of protein structure prediction, which employs a distinct and separate set of techniques to achieve solutions. As a field, the problem of CSP is still quite unresolved (Price, 2013), in part because of the complexity of the calculations required, and the lack of suitable first-principles techniques. However, the advent of commodity super-computing in the 1990s and the development of several new techniques in computational chemistry has made CSP a real possibility.

The authoritative organization surrounding CSP is the Cambridge Crystallographic Data Centre, which periodically organizes blind tests to assess the state of CSP research (Motherwell et al., 2002; Day et al., 2005, 2009; Bardwell DA et al., 2011). So far there have been six blind tests, starting in 1999 and most recently being concluded in September 2015 (with results to be published later in the year). At this stage, there are subclasses of the CSP problem that are considered to be generally understood, but there remain several problems that are nontrivial and unsolved. Among these, the problems of polymorphism and flexible molecules remain among the most challenging problems.

The value of CSP lies in the importance of screening pharmaceutical drugs (Bauer et al., 2001; Morissette et al., 2004; Price, 2004; Jones et al., 2006; Issa et al.,

2009), explosives (Foltz et al., 1994; Larionov, 1997; Miller and Garroway, 2001; Deschamps et al., 2008), and other materials for undesirable properties before costly synthesis experiments are performed. A famous and costly example of this is the drug Ritonavir, an antiretroviral drug targeted at the human immunodeficiency virus (HIV) that exhibited polymorphism after passing clinical trials (Bauer et al., 2001). It was discovered that the original formulation of the drug was a metastable polymorph, which later converted to a more stable and less soluble form. Designed to be taken orally, the loss of solubility impacted the bioavailability of the drug, leading to an eventual recall of the drug. Pharmaceutical companies now routinely perform expensive and labor intensive polymorph screens as a consequence of this phenomenon. The same problem impacts explosives development, since the reactivity of explosive compounds strongly depends on space group lattice type. **A tool capable of performing consistent, accurate, and effective CSP *in silico* would be invaluable to pharmaceutical developers and materials scientists.**

Solving the CSP Problem

The nature of crystallization as a process is only partially understood. Numerous experimental techniques for finding optimal crystallization conditions are known, and some controllable properties have been elucidated, but ultimately the knowledge surrounding crystallization techniques relates to issues of reproducibility, scaling, and quality assurance (Nagy et al., 2008). Knowledge of the basic science surrounding crystallization processes on the molecular level remains lacking, especially in regards to polymorphism and the processes that initiate crystal formation. Some modelling and theoretical work has been done in this area, but generally the field is still in its infancy, bound by the limitations of current theoretical knowledge and the lack of methods that enable the inspection of crystal growth at the molecular level.

Despite these shortcomings, insight can be gained from Levinthal's paradox, which is especially potent when applied to CSP. Although the thought experiment was originally based on protein folding, the implications hold true for small organic molecules as well. The premise of Levinthal's paradox is that the number of valid structures a polypeptide chain can adopt is enormous, but a protein will adopt no more than a handful of conformations in nature (Zwanzig et al., 1992). The adoption of an idealized conformation is achieved on a very short time scale (μs - ms), but to sample a large representative number of possible conformations randomly would take more time than the life of the universe. Consequently, there must be underlying forces that drive the folding process in an efficient manner. Applied in the context of small organic molecules, a similar condition must hold true; even for small simple molecules, the number of potential crystal structures is very high, but the number of naturally or experimentally occurring structures is clearly constrained by the energetics of the system.

Levinthal's paradox implies that there is a means of calculating a solution that will be more efficient than a brute force search. Since crystallization is a dynamic process, molecular dynamics (MD) could be used to emulate crystal formation, leading directly to a correct solution on the implication that a kinetically desirable path will be taken to the solution. However, due to problems with generating accurate generic force fields (discussed later) and the lack of atomic level models of crystal formation, this is probably not tenable for small organic molecules at this time. Furthermore, given the timescales of crystal formation and the required femtosecond time-step granularity of MD, obtaining results would take a significant amount of time such that effectively getting solutions in a high throughput way becomes unfeasible, especially when considering polymorphism. Moreover, since kinetically or thermodynamic paths could be taken to a solution, multiple MD experiments would need to be performed.

Having ruled out MD simulations as a possibility, the only other viable solution is

to take point measurements of the lattice energy of stochastically or procedurally generated crystal structures using first principles or semiempirical techniques. Since nature tends to the lowest energy state, successfully finding the lowest energy configuration for an arbitrary molecule will likely match any experimentally determined structures. Based on this key fact, three primary issues with CSP can be elucidated, which characterize the problem domain:

1) The underlying first principles calculations used in CSP must provide an accurate energetic ranking of structures. Generally, crystal formation can take either a thermodynamically favored or kinetically favored route; however, a kinetically favored structure which is not at an energetic minimum may possibly convert to a more thermodynamically stable form if the barrier to conversion is low enough (as in the famous case of Ritonavir) (Bauer et al., 2001). Given this fact, as a single metric, lattice energy is the most important when considering candidate structures, and by extension, an energy calculation with poor accuracy will result in bias towards structures which do not reflect nature.
2) The search space for CSP is very large. A first approximation of the degrees of freedom for an arbitrary molecule in a repeating lattice is three times the number of atoms in the molecule multiplied by the number of molecules in the repeating unit. By simplifying the atomic model so that most atomic positions are fixed relative to the position of the molecule in the repeating crystal unit, the degrees of freedom can be reduced to approximately 10 degrees of freedom, plus additional degrees to account for internal molecule flexibility, but this still represents a nontrivial search space. In addition to this, the presence of local minima on the energy hypersurface means that Newtonian approximation cannot be used to find the global minima, and so global optimization

techniques must be used in order to obtain good results (Liberti, 2008).

3) In relation to the first two issues, many global optimization techniques rely on grid based or population based techniques to identify solutions (Bardwell DA et al., 2011), which multiplies the amount of calculation work to be performed. Other solutions require significant initial setup of force fields, which may also require nontrivial human intervention to obtain good results (Neumann, 2008). Because of these factors, there is strong imperative to automate as much of the CSP process as possible, and to eliminate human contribution beyond the initial setup of the molecular system. Otherwise, the CSP process becomes inefficient.

Summarizing, a good CSP method produces accurate energetic rankings, adequately samples the search space, and is maximally automated. A discussion of each of these issues and a summary of techniques currently employed follows.

Accuracy in Energetic Calculations

Since the fitness of structures is measured solely by energy, the calculations that that are performed to assess the quality of candidate structures must be accurate and bias free. Since many potential structures will be evaluated, and many of those structures will be nonideal, it is important that the contributions of intra and intermolecular forces be correctly estimated and balanced (Karamertzanis and Price, 2006). Furthermore, the error of such calculations must be small enough that polymorphs can be energetically distinguishable (Yu et al., 2005).

A primary concern is the handling of flexible molecules. For first row atoms, rigid molecules are considered a solved problem in CSP (Bazterra et al., 2002a; Karamertzanis C. C., 2005). The interaction model can be largely reduced to the

intermolecular potentials for Van der Waals and ionic charges, and dipole moments where relevant. However, the introduction of internal degrees of freedom that have similar energetics to the intermolecular forces severely complicates the fitness evaluation process. This is especially true when using force fields, where the semiempirical assignment of constants to torsions generally mischaracterizes the electronic interactions (Karamertzanis and Price, 2006). So, a level of theory that incorporates electron level calculations is most likely needed for flexible molecules, especially the class of molecule typically found in pharmaceutical formulations.

Several methods gaining popularity in the recent blind tests are calculations based around dispersion corrected density functional theory (DFT-D) (Neumann and Perrin, 2005; Bardwell DA et al., 2011). DFT-D shows promise in that the calculation quality is sufficient to accurately rank flexible molecules, but the calculation time is still low enough that full-scale predictions are tenable. Furthermore, several studies suggest that a final re-ranking step using DFT-D is very effective in correctly ranking structures, regardless of the means that those structures were generated (van de Streek and Neumann, 2010; Lund et al., 2013). An older but still practical method is ranking and optimization using molecular mechanics (MM) methods and force fields. The practicality of MM is a consequence of the simplicity of the calculation; compared to first principles methods MM is computationally lightweight. The subject of force fields is problematic, however: in many cases force fields suffer from bias that result in poor ranking of structures (Kim et al., 2009). In other cases, force fields may be hand built through appropriate first-principles reasoning, but this remains a labor intensive process that requires human intervention. There exist several different force fields, most of which are designed to handle specific classes of molecules, and so it is tempting to combine multiple force fields in a way that maximizes accuracy. However, the decision process on what force fields to use in such a method would strongly rely on human expertise, and it would not be practical

from a computational standpoint to try to encode this as an expert system. Consequently, a general method employing DFT-D or a similar technique will most likely come to dominate CSP calculations because of the transferability of such methods to different molecular systems.

Sampling the Search Space

The shape of the energetic landscape of an arbitrary molecule is generally punctuated by numerous local minima. This complicates the global optimization process; if Newton's method could be used to traverse the energy hypersurface effectively then CSP would be a solved problem. Because of the nonuniformity of the hypersurface, more advanced techniques that sample the energy landscape broadly and capture multiple local minima are necessary. In practice, this usually entails increasing degrees of refinement where a large sampling of structures is identified and filtered across a multistep process. Figure 1.1 shows a pictorial representation of this process. In a first pass CSP refinement, a very large population on the scale of 10^6-10^8 structures might be filtered using a low cost energetic calculation or other mechanism (such as the elimination of structures by density), reducing to a final population of 10^3-10^4 structures. On the second pass, more complex levels of theory and local optimization might be employed to further refine this population until a small set of structures (10-100) suitable for in-depth calculations can be identified. Typically, at this stage a pharmaceutical developer may take this as a representative sample of potential crystal structures, and use that to inform risk-based decisions on the potential outcome of a crystal screening and viability. However, an underlying risk remains; a global minimum with a narrow energy-potential well may not be easily found and is strongly dependent on the search method used (Figure 1.2).

The search methods in use for CSP are essentially all based on Monte Carlo simulation to a degree. Methods span from Monte Carlo-based simulated annealing, to

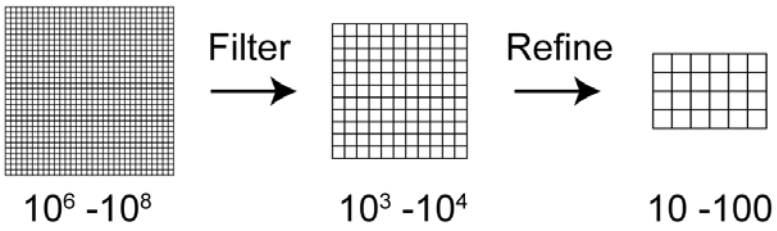


Figure 1.1: A representation of the refining process that might be used in CSP, with relative sizes of structure populations at each step. Initially a large number, possibly millions, of structures might be generated. After filtering this might be reduced to a few thousand structures, which could then be further refined through local optimization techniques to obtain a representative set of structures. This small set, in the tens of structures, could be even further refined if needed.

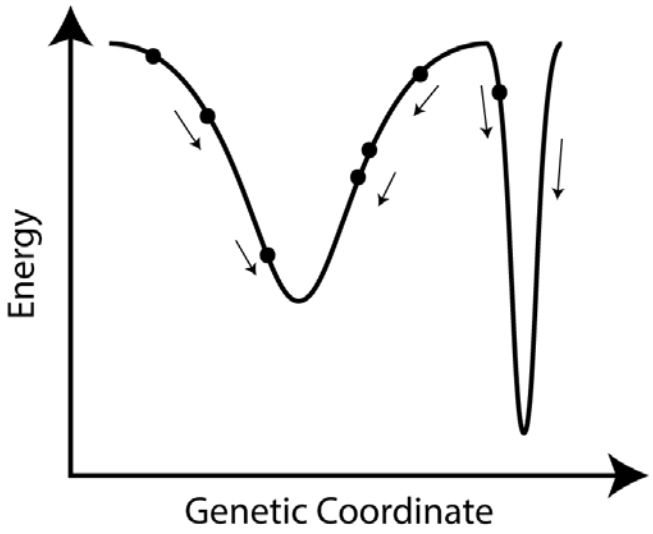


Figure 1.2: An example of a narrow energy well scenario. Dots represent the initial energies of candidate structures along the energy hypersurface. In a multidimensional system there are typically many energy potential wells, with different local minima. Depending on the search algorithm used, the left energy well might be oversampled relative to the right energy well, as shown above. In both cases, the use of local optimization will approximate the true local minima, taking a path represented by the small arrows.

search via Sobol sequence, to more fundamentally complex methods like genetic or evolutionary algorithms (Bardwell et al., 2011). Generally, fully Monte Carlo-based search methods suffer from nonuniformity issues, and rely heavily on luck to find a solution quickly (Niederreiter, 1988; Morokoff and Caflisch, 1995). Simulated annealing methods improve on this substantially, in that local minima can be escaped effectively after local optimization. However, given the dimensionality of the crystal search space, it can be difficult to effectively choose a directional vector for the annealing process, hence, the Monte Carlo aspect of nonuniformity presents itself again. Sobol sequences are excellent for overcoming the uniformity issues present in Monte Carlo based sampling, but suffer the pitfall of requiring an exponentially increasing number of samples as the dimensionality of the system increases (Sobol, 1998). At a first approximation a Sobol sequence is only more effective than a grid search method if uniform effective sampling is better than a grid, and only if the number of sampling points is less than the grid. Genetic algorithms and evolutionary algorithms are promising in the sense that the negative effects of Monte Carlo sampling are diminished by the preservation of high-fitness structures (Goldberg and Holland, 1988; Falkenauer, 1998). Genetic algorithms suffer from a different set of issues, however, in that a high number of samples is typically required for a good solution to be found, and are highly sensitive to the fitness function used (Holland, 1973; Fitzpatrick and Grefenstette, 1988). Furthermore, the identification of a solution is very sensitive to the starting conditions of the GA; a poorly formed initial population will exclude valid solutions and make it impossible to predict a structure by virtue of gene exclusion (Kim et al., 2009). As a consequence of these issues, the main problem with picking a search algorithm is selecting a method that increases the probability of finding good and valid solutions without compromising the limits on computational effort.

Automation Concerns

CSP should ideally be as fully automated as possible, especially for the purpose of molecule screening. Pharmaceutical development usually includes the testing of several hundred or even thousands of molecules for viability, which includes solubility testing and polymorph screening. To this end the full automation of CSP is greatly desired. However, because of the nature of the problem, significant guesswork is still involved in the refinement process. Furthermore, many groups still employ methods that require a fair amount of human intervention to produce accurate results, such as in the case of the software GRACE, which uses tailor-made, proprietary potentials (Neumann and Perrin, 2005; Neumann, 2008; Neumann et al., 2008). The elimination of human interaction as much as possible is the main barrier to fully automating CSP.

In addition to this, software design and incorporation of parallelism is especially crucial to the automation process. The development of CSP software to be robust and scalable closely mirrors the general needs in scientific computing; as the field moves to Exascale computing and increasingly multicore architectures, it is recognized that commodity scientific computing is becoming available, and with that increased capacity will come greater scientific progress. In addition, the incorporation of graphical processing units (GPUS) and other accelerator technologies like many-in-core (MIC) architectures, into CSP workflows has significant potential value, if key limitations in the algorithms used in CSP and the associated energetic calculations can be overcome. Consequently, CSP is well poised to take advantage of increased computing resources, but will only be able to do so if CSP can be made “hands-off” (but not necessarily a black box).

The Modified Genetic Algorithm for Crystals (MGAC)

One algorithm that is used to solve the problem of CSP is the genetic algorithm (GA). Genetic algorithms use the concepts of survival of the fittest, coupled with genetic

inheritance, to solve hard configurational problems (Goldberg and Holland, 1988; Falkenauer, 1998). CSP is well suited to this search method because crystal structures can be represented with a simple and consistent schema that can be used as a genome. This simple representation, coupled with an effectively unbiased means of generating a volume-minimized three-dimensional structure, and a high quality way to rank structures energetically, effectively enables the use of genetic algorithms. As an important aside, genetic algorithms are not necessarily well suited to all problems, primarily because they require high multiplicity of fitness evaluations, which may not be tenable depending on the complexity of the fitness calculation, as may be the case with first principles calculations in CSP. However, results so far have shown that GAs can be successful in identifying solutions in CSP, and so further exploration is warranted.

The Modified Genetic Algorithm for Clusters and Crystals (called MGAC1 in this dissertation) is the software used in the Facelli group to perform CSP (Bazterra et al., 2002a, 2002b, 2004, 2007; Kim et al., 2009). It was originally designed in the early 2000s and has been iteratively updated since then. The genetic algorithm used in MGAC1 can be broadly summarized in a few steps (and also shown in Figure 1.3):

1) A population of individuals is created; each individual represents a crystal structure and is wholly defined by a simple schema consisting of crystal structure parameters and configurational properties of the molecule and crystal system (described in the following section on the MGAC1 Schema).
2) A three dimensional representation of each structure is generated and then filtered by volume, using a method to estimate the likely volume of the true structure of the molecule. The filter is discussed in a later section.
3) All structures that pass the volume filter are structurally optimized and evaluated to determine their energy using a suitable computational method.
4) The structures are ranked, with the highest ranked structure having the lowest

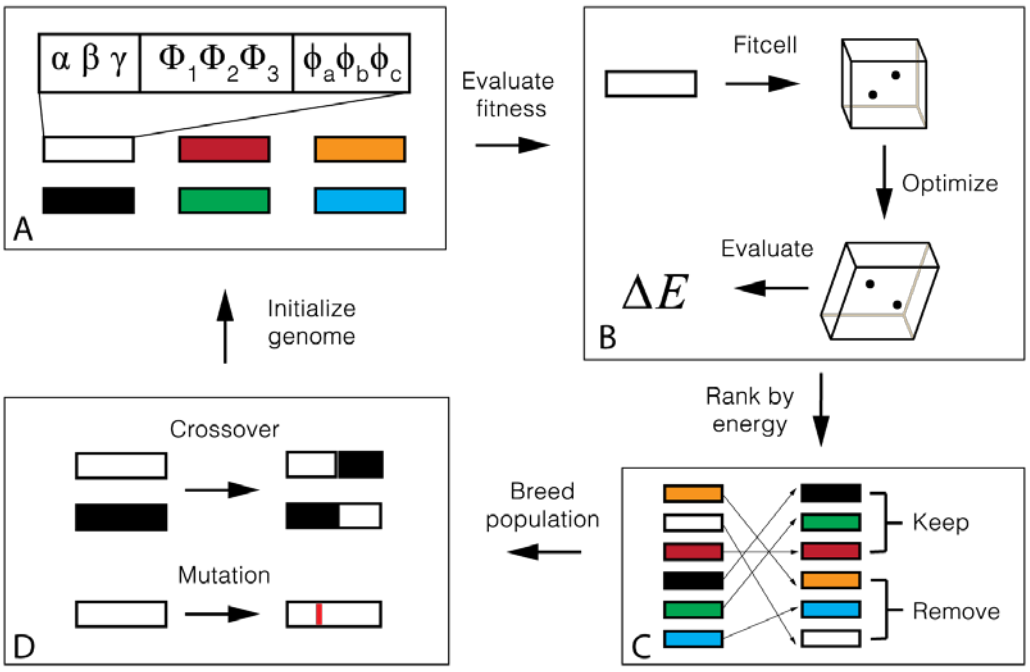


Figure 1.3: A graphical outline of the MGAC1 algorithm. In panel A, the MGAC1 schema is represented by each of the small boxes, which form a population of candidate structures. The fitness of each structure is evaluated in panel B through a three-part process, where the schema is transformed into a three-dimensional representation, which is volume minimized and filtered based on volume constraints. The structures that pass the volume filter are optimized and then evaluated to determine their hypothetical energy. The final set of structures is ranked by energy in panel C, with a subset being removed from the population. In panel D, the remaining structures are crossbred with each other, using a roulette wheel selection method. A subset of the structures is also randomly mutated. The newly generated structures form the new population to be evaluated, while the old structures only participate in the reranking and breeding process. This is repeated until some arbitrary convergence criteria is met.

energy. A fraction of poorly ranked structures may be removed at this step if an elitism GA model is used.

5) Crossover and mutation are applied to generate new structures. Crossover effectively represents sexual reproduction, whereas mutation represents environmentally caused changes to the genome.

6) The new population is evaluated, effectively repeating all previous steps until a convergence criterion is met.

In the original tests, MGAC1 used CHARMM, a molecular mechanics solver (Brooks et al., 1983; MacKerell et al., 1998), coupled with the General Amber Force Field (Wang et al., 2006), to perform optimizations and establish energetic fitness. This is a very efficient process; CHARMM is very fast, even in larger systems (>300 atoms). MGAC1 is also parallel enabled, using MPI to distribute work across many nodes, relying on a server-client model to distribute work. Typically, a population size in the range of 30-100 individuals is used, with 50% replacement each generation under an elitism model. In the elitism model, structures are ranked according to energy, and then removed if their rank is larger than the population size, effectively eliminating structures with undesirable traits, decreasing diversity and increasing average fitness in the population as the number of generations increases. Convergence is established when a particular structure or set of features comes to dominate the population, which depends on the population size to a degree, whereas the loss of diversity depends largely on the replacement.

The MGAC1 Schema

In a well-packed crystal lattice, the atomic positions ultimately define the energy of the structure. However, because the internal structure of a molecule is more or less fixed, then the degrees of freedom can be largely reduced to the dimensions of the crystal lattice, the effective symmetry operations present in the lattice, and the positions and

rotations of molecules relative to the crystal lattice origin. The only feature not taken into account by these properties is the rotation of torsional bonds. This generalizes to approximately 10+n degrees of freedom, where n is the number of torsional bonds. A schema of mostly independent parameters is then defined from these parameters (Bazterra et al., 2002a):

1) Symmetry operations; this is a discrete parameter bounded on the 230 mathematically defined crystallographically valid space groups (Hahn, 2002). Historically this has been handled by selecting a subset of space groups that are statistically overrepresented among known crystal structures in the CSD database (Allen, 2002).
2) The unit cell angles of the crystal lattice (α, β, γ). These are partially determined by the space group, as some lattice types have fixed angle parameters.
3) Unit cell ratios (rA, rB, rC). These parameters are a new addition to the schema, as recent discoveries have shown that a subset of space groups are not adequately represented by cell angles alone. This is covered in Chapter 3.
4) The internal rotation of the molecule relative to the crystal lattice; this can be represented by three rotation angles (Φ_1, Φ_2, Φ_3) but ultimately the underlying representation is based on an angle-axis formulation.
5) The fractional position of the molecule within the lattice (x_f, y_f, z_f). Many space groups have degenerate fractional positions due to symmetry operations that form equivalent subspaces in the lattice. However, the mathematical representation of these subspaces is inconsistent, and so this parameter may or may not be constrained depending on the space group (Hahn, 2002).
6) Torsional angles of flexible bonds in the molecule. This is dependent on the chemical diagram of the molecule.

Some additional things to consider as part of the schema are the ranges of acceptable

values for different schema elements. In particular, the cell angles of the unit cell are constrained, both from the perspective of different space groups, and from a mathematical standpoint. The full range of cell angles spans the open interval (0°, 180°), but in actuality, the sum of the three cell angles cannot exceed 360 degrees, nor can they violate certain properties as defined in Foadi and Evans (2011), otherwise, the unit cell effectively becomes "imaginary". In general, this particular set of rules only applies to triclinic, monoclinic, and rhombohedral lattice types, since in other lattices cell angles are typically fixed. A further constraint on cell angles deals with the concept of the reduced unit cell: many lattice types can have degenerate combinations of unit cell angles, cell lengths, and molecule rotations. According to section 9.2 in the ITC handbook (Hahn, 2002), the preferred angles for reduced cells are bounded such that all cell angles are between 60 and 120 degrees; choosing unit cells angles in this range essentially eliminates degeneracy and prevents the formation of bizarre unit cells that are extremely thin.

Volume Filtering

The volume filter serves to reduce the search space for the GA to a reasonable set of structures, which is important both for reducing the number of optimizations, as well as having a nontrivial impact on the local optimization process. In general, the reason for the volume filtering is to eliminate structures which are mostly empty space, or otherwise poorly packed. In such cases, spurious energies can be given which give poor representations of the packed solid state, as has been found in the case of glycine. Discussed in Chapter 4, the zwitterionic form of glycine is unstable in poorly packed configurations and changes protonation state accordingly, but in the solvated or packed state, intermolecular forces stabilize the energy substantially (Lund et al., 2015).

The filter works by starting with an estimate of the likely single molecule volume for the molecule of interest. The model used is the ARH model (Beaucamp et al., 2007),

which relies on a semiempirical method to calculate the density of the candidate molecule, from which a volume estimate can be given. When applied to a representative population of structures, the average error of the model is approximately 2%, with most structures within 30% of the expected density. Consequently, any structure that fitcell generates and passes to the volume filter, if within a range of +/- 30% of the volume estimate is expected to be a valid structure. In practice there may be advantages to expanding beyond a 30% threshold, however, as a consequence of using local optimization this is difficult to parameterize, because a tightly packed structure may be difficult to optimize.

Practical Considerations for CSP with MGAC

All in silico experiments require varying amounts of computing resources. In MGAC1 the resources required for predicting molecules are relatively minimal for today's computing resources, with computations taking on the order of hours to complete using multiple cores. With the introduction of DFT-D based methods, this computational requirement increases substantially, by a factor of at least 1,000. For direct comparison, the optimization (or minimization) of a molecule the size of histamine using molecular mechanics takes a fraction of a second on a single core, whereas using DFT-D on the same molecular system takes on the order of 10 min on a 16-core compute node. Since a basic requirement of CSP is the evaluation of many structures, it can be expected that using DFT-D as the sole energy calculation method will be very costly relative to molecular mechanics.

For MGAC, experiments are very different in cost depending on the method. A typical MGAC1-CHARMM run on Sandy-bridge era compute nodes (E5-2670, 16-core, 2.6 GHz), comprising 250 generations, with 50% replacement and a population size of 30, spanning 14 space groups, could be expected to take around 150 core hours total for a molecule the size of histamine. Generally, these calculations are performed in multiples of

ten to improve sampling, so for complete search using MGAC1-CHARMM could take up to 1500 core hours, for an estimated total of 530,000 structure evaluations. An MGAC1-QE run on the same hardware and same molecule, however, takes substantially longer. An MGAC1-QE run similar to that presented in Chapter 5, with 10 generations, population replacement of 3 times the population size, in a single space group, with a population size of 90, in a single space group without duplication, takes 40,000 core hours by itself, for at most 3000 evaluations. It can be very easily seen that a full prediction using statistical sampling of the best 14 space groups, would take a minimum of 560,000 core hours to complete using this methodology, which is a substantial amount of resources requiring the use of a national supercomputing center to be even feasible. Furthermore, because of poor scaling in Quantum Espresso, such a prediction would likely take closer to 1 million core hours to complete. Therefore, there is a strong imperative to streamline the MGAC process to permit higher quality predictions using DFT-D, while preserving the capability to sample multiple space groups and keeping CPU costs lower.

Concluding Remarks

The most recently published results from an MGAC1 are from 2008 (Kim et al., 2009), where a large set of molecules, the Karamertzanis-Price dataset, were used as the basis for testing. The K-P dataset is a set of structures designed to represent molecules of pharmaceutical interest; it contains a few polymorphs and co-crystals (Karamertzanis and Price, 2006). Of the 22 structures tested, 16 matches were found, but the ranking of the matches were extremely varied, ranging from 1 to 1162. The results of these experiments essentially highlight the issues of using generalized force fields; although in some cases high accuracy could be obtained, the sporadic nature of the ranking severely complicated the prediction process. Furthermore, several instances of bias were identified that prevented the successful prediction of structures. This provides the impetus to use a higher

level of theory, as that would presumably correct the bias issues and ranking problems presented by using CHARMM and the GAFF.

In the next chapter, the exploration of using DFT-D as the energy and optimization source for MGAC is discussed, demonstrating the successful reranking of datasets from the 2008 predictions (Kim et al., 2009) using Quantum Espresso (Giannozzi et al., 2009). In Chapter 3, the formulation of a new algorithm for the fitcell routine is discussed, as well as the addition of new schema elements to take into account the ratios of unit cell lengths. Results based on the successful implementation of this algorithm are presented in Chapter 4, where the three atmospheric pressure polymorphs of glycine were found in their native space groups. Previously unpublished results on the prediction of histamine are presented in Chapter 5. In the sixth chapter, the theoretical basis for a new set of space group schema elements enabling structures from different space groups to be crossed with each other is outlined, along with other important innovations towards crystal structure prediction in multiple space groups. Chapter 7 includes a summary of preliminary results for the theoretical work presented in Chapter 6.

CHAPTER 2

OPTIMIZATION OF CRYSTAL STRUCTURES OF ARCHETYPICAL PHARMACEUTICAL COMPOUNDS: A PLANE WAVE DFT-D STUDY USING QUANTUM ESPRESSO[1]

Albert Lund,[a] Anita M. Orendt,[b] Gabriel I. Pagola,[d]

Marta B. Ferraro,[d] and Julio C. Facelli [b,c]

[a]Department of Chemistry, [b]Center for High Performance Computing,

and [c]Department of Biomedical Informatics, University of Utah,

155 South 1452 East Room 405, Salt Lake City, UT 84112-0190, US

[d]Departamento de Física, Ifiba (CONICET), Facultad de Ciencias

Exactas y Naturales, Universidad de Buenos Aires,

Ciudad Universitaria, Pab. I (1428), Buenos Aires, Argentina

Abstract

Previously, it was shown that crystal structure prediction based on genetic algorithms coupled with force field methods (called MGAC[2]) could consistently find experimental structures of crystals. However, inaccuracies in the force field potentials often resulted in poor energetic ranking of the experimental structure, limiting the usefulness of the method. In this work, dispersion corrected density functional theory is

[1] Adapted with permission from *Cryst. Growth Des.*, **2013**, *13* (5), pp 2181–2189, **DOI:** 10.1021/cg4002797. Copyright 2013 American Chemical Society.

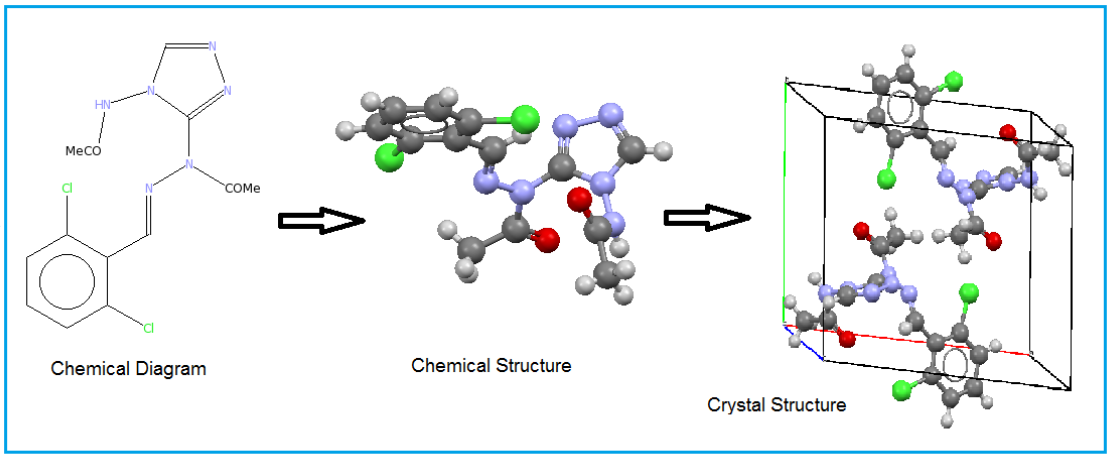
[2] In this chapter, MGAC generally refers to MGAC1, but was published before the inception of the MGAC1/MGAC2 notation. For the remainder of this chapter it is left as it was originally published.

employed to correct the results of those experiments, using the software package Quantum Espresso. Proper running parameters were established for application with MGAC, and it is shown here that the variable cell optimization of experimental structures will reproduce the experimental structure with high accuracy (RMS < 0.5) for a large set of archetypical pharmaceutical compounds. We show that using electronic structure theory-based methods greatly enhances the energetic ranking of structures produced by MGAC-CHARMM, such that the experimental match is found with a high degree of accuracy.

Introduction

Over the last several decades there has been much effort towards the goal of being able to readily and reliably predict, by computational methods alone, the crystal structure of a molecule based only on its chemical diagram (Day, 2011, 2012; Kendrick et al., 2011; Lehmann, 2011; Price and Price, 2011). The process to do this is shown in Scheme 1. The ability to do so has far reaching implications in many areas. On a basic science level, the accomplishment of this goal can lead to an understanding of the principles that control crystal growth. More practically, the ability to successfully predict crystal structures based on computation alone will have an impact in many industries, including pharmaceutics, agrochemicals, pigments, dyes and explosives.

The current status of CSP can be evaluated by the performance of the participants in the periodic blind tests that have been organized by the Cambridge Crystallographic Data Centre (Lommerse et al., 2000; Motherwell et al., 2002; Day et al., 2005, 2009; Bardwell DA et al., 2011). There have been five blind tests since 1999, the latest held in 2011, and we have participated in the last four with our MGAC (Modified Genetic Algorithm for Crystals and Clusters) package (Bazterra et al., 2002a, 2002b, 2004, 2007; Kim et al., 2009). MGAC is capable of doing CSP for any space group, any number of molecules per asymmetric unit, and is able to deal with conformational flexibility of the



Scheme 2.1: Overview of the Crystal Structure Prediction (CSP) process.

molecule being studied in order to explore the entire crystal energy landscape. The generation of trial crystal structures is completed utilizing genetic algorithms but when using the current version of MGAC, which relies on the use of the CHARMM (Brooks et al., 1983; MacKerell et al., 1998) molecular mechanic program using the Generalized Atomic Force Field (GAFF) (Wang et al., 2006) for the energy evaluation of the trial structures, the ranking of the structures is not always reliable do to deficiencies of GAFF.

The results of the last two blind tests showed the advantage of using dispersion corrected density functional theory (DFT-D) (Grimme, 2004, 2006; Grimme et al., 2010) to both generate a molecule specific tailored force field that is thereafter used to generate trial structures and to reorder a subset of the trial structures in search of the lowest energy crystal structures (Neumann and Perrin, 2005; Neumann, 2007, 2008; Neumann et al., 2008; Kendrick et al., 2011). There have been two other approaches utilizing first principle calculations applied to crystal structure prediction of organic crystals recently presented in the literature (King et al., 2011; Zhu et al., 2012). These results do lend promise to using DFT-D methods to completely replace molecular mechanics as the method of choice for the evaluation of the energies of the trial crystal structures in CSP. This has been thought to be computationally unfeasible; however, with recent advances in computer technology and availability, we believe the time has come to explore this option.

Quantum Espresso (QE), (Giannozzi et al., 2009) www.quantum-espresso.org, is a set of computer codes to perform electronic structure calculations based on density functional theory, plane waves, and pseudopotentials that is capable of calculating the energy and performing local optimizations on crystal systems using DFT-D. Its primary application is determining the band structure and other properties of semiconductors and other solid state materials, but while it can also be used to examine the energies of and optimize organic crystal structures, to the authors' knowledge, no comprehensive study of its performance has been reported in the literature.

The data set chosen for a comprehensive study of the performance of QE for the energy evaluation of organic crystal structures in conjunction with the MGAC generation of trial structures is the Karamertzanis and Price (K&P) data set (Karamertzanis and Price, 2006) which was initially selected as a proxy representation for characteristic molecules of pharmaceutical interest and subsequently was used by us to test the reliability of MGAC crystal structure predictions (Kim et al., 2009). This set, shown in Figure 2.1, contains molecules which represent a variety of pharmaceutically relevant functional groups, as well as five compounds that present experimentally determined polymorphs, and three co-crystal systems. In our previous work, we attempted to predict 22 of these structures using MGAC, and were successful in obtaining structures for 16. However, in the majority of these successful predictions, the rank of the best match was often well outside of the expectations of blind test criterion. We attributed this to inaccuracies in the force field, as well as bias introduced by CHARMM optimizations. In five of the six cases where a match to the experimental structure was not found, potential failures of the GAFF to properly handle the intermolecular interactions were identified.

In this work, full crystal optimizations of the crystal structures in the K&P data are completed using the DFT-D method found within QE. The parameters necessary to reproduce experimentally determined structures for each of the K&P molecules without trading performance are determined. In addition, QE is used to do full crystal optimizations on several MGAC derived initial populations as well as do re-ranking of several sets of the lowest energy MGAC-CHARMM structures from previous runs on these systems.

Methods

All calculations were performed using version 5.0.1 of Quantum Espresso using the vcrelax option, which allows for optimization of the unit cell parameters along with all

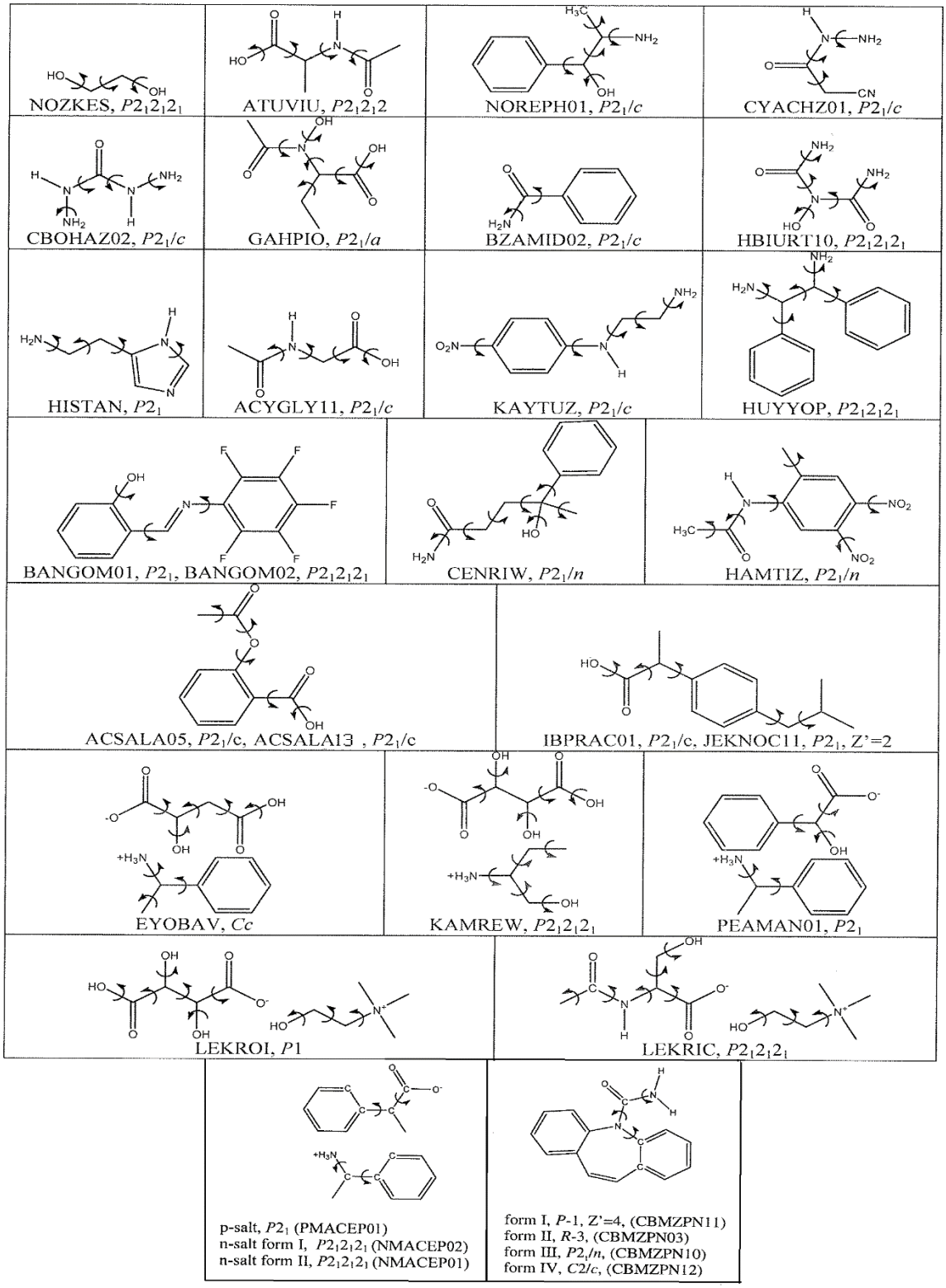


Figure 2.1**:** Molecules in the Karamertzanis and Price (Karamertzanis and Price, 2006) data set. Also shown are the CSD Reference codes, the space group, and the important dihedrals to consider in the search of the crystal system.

atomic coordinates. Calculations were performed on either 12-core nodes (two socket, 6 core Xeon X5660 processors, 2.8 GHz) with 24 GB RAM or on 8-core nodes (two sockets, quad core Xeon E5462 processors, 2.8 GHz) with 16GB RAM. Calculations were performed on a single node, which was found to give the best hardware utilization, except in one case where wall time limits imposed in the batch system necessitated the use of two nodes. All calculations to determine run times for the crystal structure optimizations were performed on the 12-core nodes. The QE parameters explored in this study are discussed in the next section. Intermediate processing of input and output files was conducted using custom Python scripts. RMS values were computed using the Solid Form Crystal Packing Similarity method in Mercury CSD, using 15 molecules for comparison and ignoring hydrogen atoms (Chisholm and Motherwell, 2005). The current version of MGAC which uses CHARMM (Brooks et al., 1983; MacKerell et al., 1998) to optimize crystal structures was used to generate an initial population of 30 trial crystal structures and the final populations used in this study.

Results and Discussion

Determination of parameters. A series of preliminary calculations were performed on four representative molecules from the K&P data set to determine the optimal conditions for the calculations. The parameters varied, along with the range explored and the final parameter chosen for use, are given in Table 2.1. Along with these parameters, different pseudopotentials were explored. A two-step process of selecting these parameters was used. First, each parameter listed was varied independently. After selecting a subset of values from each range, a grid test was performed for those variable ranges to assess if any interdependencies existed between parameters. The results were compared by RMS to the experimental structures and simulation time.

Initial calculations determined that ultrasoft pseudopotentials gave better results

Table 2.1**:** Parameters explored to determine the optimal conditions for QE optimizations.

Parameter	Description	Range	Final
conv_thr	Minimum error threshold for self-consistency	10^{-4} to 10^{-12}	10^{-7}
ecutwfc	Kinetic energy cutoff (Ry); ecutrho, the kinetic energy cutoff for charge density and potential functions, was always 10 times the value of ecutwfc	30 to 70	55
forc_conv_thr	Force threshold for structural relaxation	10^{-1} to 10^{-4}	10^{-2}
etot_conv_thr	Energy threshold for structural relaxation	10^{-2} to 10^{-6}	10^{-3}
k-points	Grid of discrete points used in integration during DFT-D calculation	1x1x1, 2x2x2, and 4x4x4	2x2x2

than norm conserving ones. After a search of existing pseudopotentials for use in QE, three different ultrasoft pseudopotentials combinations were chosen: (1) the Vanderbilt (Vanderbilt, 1990) PBE (Perdew et al., 1996a, 1996b) pseudopotential (Van-PBE), which was available for all elements in the molecules in our test set; (2) a mix of the RRKJUS (Rappe et al., 1990) PBE pseudopotential (RRKJUS-PBE), which exists for all elements needed for the test set except F and that was used in conjunction with the Van-PBE on the F; (3) the Vanderbilt BLYP (Becke, 1988; Lee et al., 1988) pseudopotential (Van-BLYP) which does not include the F so the evaluation did not include any of the fluorine containing molecular systems of the dataset.

The results show that the components that most directly affect RMS are the choice of pseudopotential, the number of k-points, and the energetic cutoffs (ecutwfc, ecutrho). As the choice of the self-consistency threshold (conv_thr) had little effect on the RMS, the threshold was set to correspond to an energy of 0.13 J/mol, which should allow polymorphs to be energetically distinguished (Yu et al., 2005). The convergence thresholds for vcrelax (forc_conv_thr, etot_conv_thr) were also found to have minimal impact on the final RMS, so these terms were set to favor shorter simulation times. While higher energy cutoffs did result in lower RMS values, a cutoff of 55 Ry was chosen as a good balance between the RMS and the time for the calculation to complete.

It was determined that pseudopotentials utilizing the BLYP cross-correlation function produce the best RMS values overall. However, as a QE-compatible pseudopotentials utilizing BLYP is not available for F, an element commonly used in medicinal chemistry for the synthesis of pharmaceutical compounds, the combination of the RRKJUS-PBE for C, H N, O, S and Cl along with the Van-PBE for the F was selected, as it gave the second best RMS results among the three tested when including molecules containing F.

A grid of 2x2x2 k-points was found sufficient for accurate calculations; lowering

this constraint severely decreases the quality of results, whereas increasing grid size provides no clear benefit for systems in the volume range we tested. Note that the actual number of k-points used in the simulation is adjusted internally by QE to reflect the dimensions and symmetries of the unit cell; where the unit cell is not accurately sampled by a 2x2x2 grid, the grid is automatically expanded.

Optimization of experimental crystal structures. Once the optimum choice of parameters was determined, full local crystal optimizations using the experimental structure as starting initial one were performed on each of the systems shown in Figure 2.1. The only structure for which the vcrelax calculation was not completed was CBMZPN03, an exceptionally large crystal structure with 18 molecules (or 1584 valence electrons) in the unit cell.

The results of these calculations are shown in Table 2.2. In all cases where the simulation completed, QE returned a final structure which gave an RMS < 0.5 Å, with 15 out of 15 molecules aligned when compared to the experimental structure. The RMS difference between the QE full optimization and the experimental structure ranged from 0.056 Å to 0.459 Å, with a median RMS of 0.196 Å. A typical match, using the case of NOZKES, is shown in Figure 2.2, on the left. These results imply that the experimental structure is at least close to local minima of the QE energy hyper surface and that for unknown structures MGAC-QE most likely will find structures with similar proximity to the experimental ones, provided that the GA generates structures in its proximity. This is a strong indication that the MGAC-QE combination could be successful for global optimization of crystal structures.

The simulation time is primarily affected by the number of valence electrons (which correspond to the Kohn-Sham states in the DFT method), the number of k-points, and the number of vcrelax iterations that are performed. From a theoretical standpoint, simulation time scales linearly with number of electrons when normalizing against vcrelax

Table 2.2**:** Summary of QE results for the K&P dataset (Karamertzanis and Price, 2006). Included is the information on the size of the unit cell and the time for the optimization. The energy reported is that on a per molecule basis and the RMS is based on a 15 molecule match with the experimental structure.

		Molecule/system info				**Metrics**			
CSD Ref Code	Space group	Natoms[1]	no. of e-[1]	N[1]	Kpt	VC	T(Hr)	Energy (Ry)	Rms[2]
ACSALA05	$P2_1/c$	84	272	4	4	14	2.6	-239.6195421	0.097
ACSALA13	$P2_1/c$	84	272	4	4	25	4.7	-239.6193637	0.170
ACYGLY11	$P2_1/c$	60	304	4	2	23	1.4	-169.5440140	0.188
ATUVIU	$P2_12_12$	72	208	4	4	31	3.1	-183.3139758	0.128
BANGOM01	$P2_1$	52	208	2	2	32	2.3	-450.5943046	0.196
BANGOM02	$P2_12_12_1$	104	416	4	2	31	10.6	-450.5914761	0.331
BZAMID02	$P2_1/c$	64	184	4	4	25	2.0	-139.8490878	0.201
CBMZPN03	*R*-3	324	1584	18	12	?		Too Complex	
CBMZPN10	$P2_1/n$	120	352	4	2	40	9.5	-256.7059809	0.232
CBMZPN11	*P*-1	240	704	8	4	31	71.9	-256.7025145	0.182
CBMZPN12	*C*2/c	240	704	8	4	34	38.1[3]	-256.7010208	0.219
CBOHAZ02	$P2_1/c$	48	144	4	4	55	1.8	-129.8488205	0.409
CERNIW	$P2_1/n$	116	304	4	2	22	3.9	-226.8550397	0.173
CYACHZ01	$P2_1/c$	48	152	4	4	23	1.0	-131.5329627	0.063
EYOBAV	*Cc*	140	400	4	4	32	7.6	-336.3618078	0.436
GAHPIO	$P2_1/a$	88	256	4	4	40	5.9	-228.8603043	0.213
HAMTIZ	$P2_1/n$	104	360	4	4	31	11.5	-331.8894879	0.133
HBIURT10	$P2_12_12_1$	52	184	4	1	47	1.4	-184.0473440	0.459
HISTAN	$P2_1$	34	88	2	2	33	0.4	-127.2667014	0.260
HUYYOP	$P2_12_12_1$	128	328	4	4	36	11.8	-218.1767345	0.270
IBPRAC01	$P2_1/c$	132	324	4	4	24	7.7	-233.3437649	0.206
JEKNOC11	$P2_1$	132	324	4	4	42	14.5	-233.3415498	0.349
KAMREW	$P2_12_12_1$	132	384	4	4	17	6.0	-354.6493038	0.056
KAYTUZ	$P2_1/c$	96	360	4	4	34	7.4	-227.4160927	0.188
LEKRIC	$P2_12_12_1$	156	408	4	4	24	11.0	-339.3602102	0.099
LEKROI	*P*1	36	102	1	4	30	0.5	-368.3247454	0.132
NMACEP01	$P2_12_12_1$	164	392	4	4	29	18.6	-302.3179367	0.167
NMACEP02	$P2_12_12_1$	164	392	4	4	37	22.4	-302.3254026	0.225
NOREPH01	$P2_1/c$	96	240	4	4	23	4.0	-169.7442352	0.201
NOZKES	$P2_12_12_1$	40	104	4	4	19	0.3	-93.7290501	0.134
PEAMAN01	$P2_1$	78	210	2	4	21	2.6	-320.4332603	0.070
PMACEP01	$P2_1$	82	196	2	4	41	5.6	-302.3196545	0.234

[1]The number of atoms (Natoms), electrons (no. of e-) and molecules (N) reported are per unit cell.
[2]The RMS values are for 15 molecule comparisons without hydrogens completed using the crystal packing similarity function of Mercury.
[3]Times for this structure are for a run using two nodes.

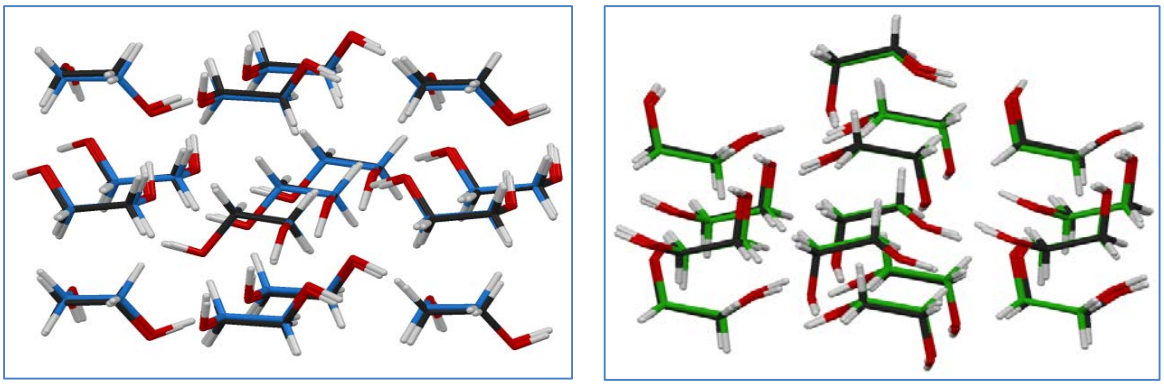


Figure 2.2**:** Results on NOKZES. The comparison on left is between the experimental and QE optimized structures (blue), whereas that on the right is between the experimental and the QE reoptimized MGAC-CHARMM match (green). Note orientations are different due to the choice of molecules used in the comparison that is made by the comparison program.

iterations, but in practice, some variation exists due to differences in convergence times for each self-consistent step. From a technical standpoint, systems with high numbers of valence electrons (>300 in our approximation) are probably not within the realm of feasible use with CSP on commonly available hardware, due to the large number of simulations performed in the search.

Test of initial MGAC populations. In order to further test the validity of using QE for the energy evaluation in MGAC, we did full crystal optimizations on initial MGAC populations of 30 crystal structures of a given space group for three of the K&P molecules, namely ATUVIU ($P2_12_12_1$), BANGOM01 ($P2_1$), and IBRAC01 ($P2_1/c$). Due to the variations in the starting structures, the number of iterations needed, the time for the optimization varied among the 30 crystal structures for each molecule. However, in each case the average time was not significantly different than the time for the optimization required when starting from the experimental structure. The average run time for the 30 structures was 2.1 h for ATUVIU, 2.5 h for BANGOM01 and 7.6 h for IBRAC01. As expected, the diversity of the initial population for each structure was maintained after optimization, with a spread of optimized energies between 0.03 and 0.05 Ry (approximately 9 and 15 kJ/mol). In every case the optimized experimental structure (Table 2.2) was lower in energy that any of the locally optimized structures obtained from the members of this initial population. Figure 2.3 shows the BANGOM01 initial population with the energies before and after local optimization, ranked in order of pre-optimization energy. Note that the energies do not decrease in a uniform fashion during optimization; this is highlighted by the fact that the structure ranked number 7 in the population is the lowest one after local optimization. This indicates the energy of the structure prior to optimization cannot be used as a proxy for the final rank of the optimized structure, and requires that all structures to be optimized before rank comparison.

Reranking of MGAC-CHARM final populations. Literature precedent shows that

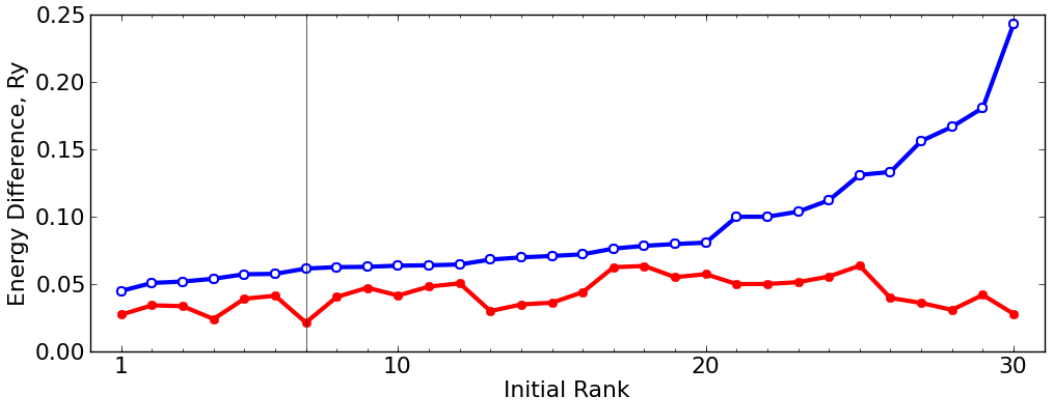


Figure 2.3: Energies for the BANGOM01 initial test population relative to the energy of the QE optimized experimental structure. The structures are ranked in order of increasing starting energy (blue points), and the best structure after optimization is indicated by the black vertical line. The order of the optimized energies (red points) does not correspond to the order of the initial single-point energies.

the use of molecular mechanics produces unreliable energy rankings. Therefore, the accuracy of the QE energy rankings should be explored. The final structures of previous MGAC-CHARMM runs for NOZKES, KAYTUZ, HBIURT10, and BZAMID02 were re-optimized using QE. In all cases, a re-ranking based on the QE final energy after local optimization placed the match to the experimental structure as one of the lowest energy structures. Table 2.3 summarizes the results obtained, including the number of structures re-optimized, the QE optimized ranking of the MGAC-CHARMM match and the RMS with respect to the experimental crystal structure for the QE structures.

For NOZKES the CHARMM-MGAC results found a match at energy rank 78 with an RMS of 0.18 Å. The first 111 crystal structures predicted by CHARMM-MGAC were re-optimized by QE and the match was found now to be the lowest energy at 0.000376 Ry (0.12 kJ/mol) relative to the energy of the QE optimized experimental structure; this structure has an RMS of 0.23 Å relative to the experimental structure (a comparison is shown in Figure 2.2, on the right). It should also be noted that two of these 111 structures, CHARMM structures ranked 68 and 74, had unphysical structures, an issue found when using CHARMM, and therefore QE calculations were not performed. The results of the QE calculations are shown in the graph in Figure 2.4, where the QE energy of each of the 111 structures is shown at the first optimization step (blue) and the optimized structure (red), with the energies being reported relative to the energy of the QE optimized experimental structure. A comparison of the red and blue lines confirms the single point QE energy of the nonoptimized MGAC-CHARMM final structures is insufficient for re-ranking, in agreement with the analysis in the previous section

Another way to view the reordering of the CHARMM-MGAC structures upon optimization using QE is shown in Figure 2.5 for the KAYTUZ results. The structure with the best match with the experimental one, which was ranked 22 with CHARMM, becomes the lowest energy structure after reoptimization using QE. This lowest energy structure

Table 2.3**:** Results of reranking MGAC-CHARMM lowest energy structures using QE optimization.

	MGAC-CHARMM results		QE re-optimization results		
Case	Rank of match	RMS	# of structures reoptimized	Rank of match	RMS
NOZKES	78	0.18	111	1	0.23
KAYTUZ	22	0.44	51	1	0.42
BZAMID02	39	0.67	53	3	0.46
HBIURT10	106	0.32	171	2	0.46

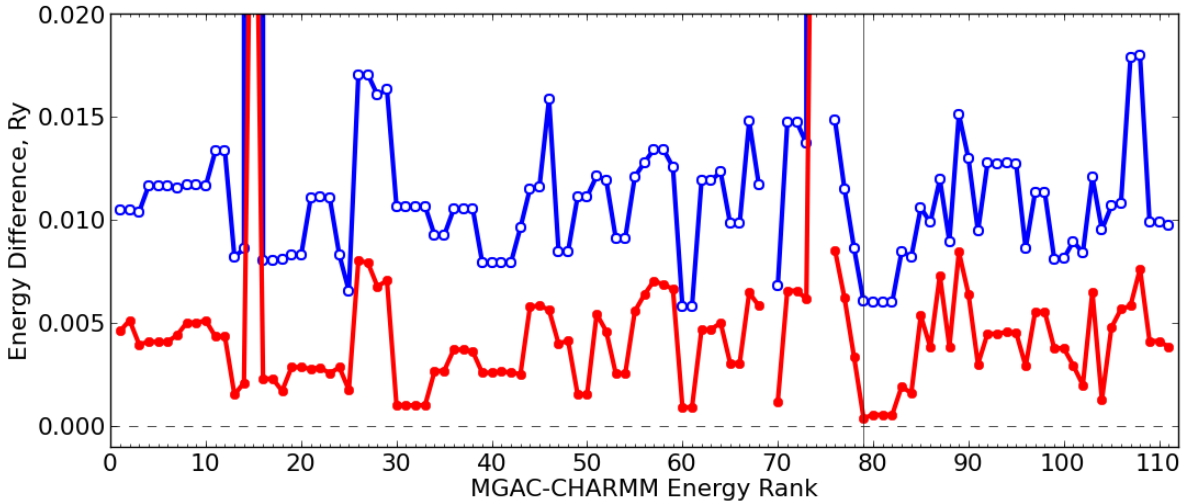


Figure 2.4**:** Results of reoptimization of MGAC-CHARMM NOZKES results with QE. The CHARMM ranking is indicated by the x axis scale. The QE energies are for a single molecule and given relative to the energy of the QE optimization of the experimental structure. The gray vertical line marks the MGAC-CHARMM match to the experimental structure.

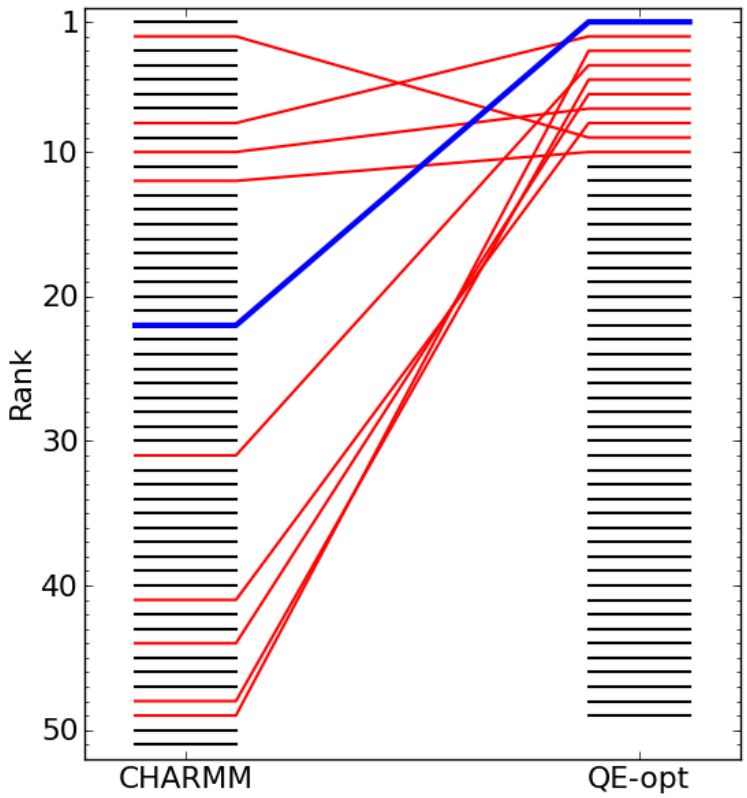


Figure 2.5**:** Reordering based on reoptimization of KAYTUZ MGAC-CHARMM crystal structures with QE. The QE column has two fewer entries due to unphysical CHARMM structures and one structure that did not converge among the 51 lowest energy structures. The blue indicates the match to the experimental structure whereas the red indicates the remainder of the top ten lowest QE energy structures.

was found to be at 0.00468 Ry (1.47 kJ/mol) relative to the energy of the QE optimized structure. The unpredictable nature of the reordering is shown by the connection of the ten lowest QE structures, shown in red and blue.

The remaining two cases tested show similar energy reranking characteristics. In the case of HBUIRT10, the structure which matched the known crystal structure was ranked 106 by the CHARMM energies, but after QE optimization it became the second lowest in energy, with both structures within 10^{-4} Ry from the QE energy of the experimental structure.

For BZAMID02 the match to the experimental structure was found to be third (0.0020 Ry or 0.63 kJ/mol relative to the QE energy of the experimental crystal structure), with two different herringbone-like motif structures coming in at lower energy, -7.1×10^{-5} Ry (-0.02 kJ/mol) and 0.00014 Ry (0.44 kJ/mol) relative to the QE energy of the experimental structure. The three lowest energy BZAMID02 structures display a high level of similarity to each other. Figure 2.6 shows cross-sections of these structures and the planar nature of the sub-lattices which results in a herringbone-like motif. The aromatic rings form interlocking pockets at the interface of each planar section (black line) such that the orientation of one planar section relative to the other is constrained to rotations of 90° or -90°. Furthermore, the orientation of the amides in each planar section relative to the other creates an additional parameter to distinguish between potential structures. This corresponds to four theoretical structures which should be very energetically similar; the top three structures correspond to three of these configurations. We were unable to find a structure resembling the fourth configuration in any of our optimized structures, and in examining the CDCC database, we were unable to locate any of these configurations except for the one corresponding to the experimental structure. The absence of other experimental configurations may have two possible explanations: 1) the nature of the planar interface may allow for all four configurations to coexist in a single crystal, which

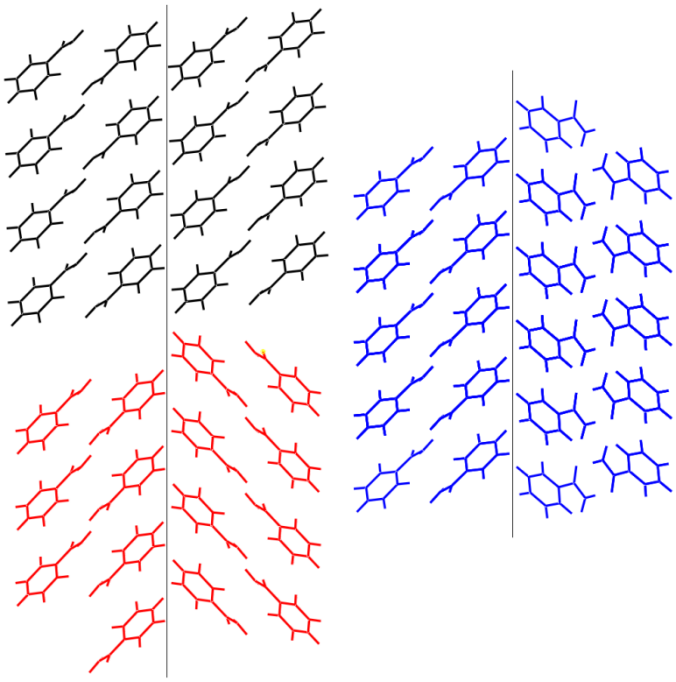


Figure 2.6**:** Cross sections of the three lowest energy structures from the BZAMID02 reranking. The black structure is the experimental match, ranked third. Red is an inversion of the top ranked structure, and blue is the second ranked structure. The left side of each structure is identical in three-dimensions, aside from minor differences in unit cell parameters. Symmetry operations with respect to the plane indicated by the thin black lines produce the other half of each structure.

may be lost due to experimental averaging or during refinement, or 2) the growth of the crystal favors one set of configurations.

Conclusions

These findings are consistent with other work where electronic structure optimizations are being used to correctly rank structures produced by force field approaches. We have shown that this method is equally applicable to our GA-based method, and that DFT-D optimization can produce high quality structures consistently and effectively. Furthermore, our system can identify and energetically distinguish highly similar structures, as seen in the case of BZAMID02 and HBIURT10. We fully expect that integration of MGAC and QE will provide a useful tool to the scientific community where quality CSP software is not freely available.

Acknowledgements: The support and resources from the Center for High Performance Computing at the University of Utah is gratefully acknowledged. MBF and GIP acknowledge the support from the University of Buenos Aires and the Argentinean Research Council.

CHAPTER 3

AN IMPROVED METHOD FOR BUILDING CRYSTAL STRUCTURES FROM GENETIC ALGORITHM BASED SCHEMA IN CRYSTAL STRUCTURE PREDICTION

A fundamental step in the process of generating structures in the Modified Genetic Algorithm for Crystals (MGAC) is the transformation of the genetic schema into a three-dimensional structure. This is important for two reasons: 1) The formation of a crystal structure from a schema does not have fixed unit cell lengths, which are dependent parameters that are constrained by the shape and orientation of the molecules in the unit cell (Bazterra et al., 2002a); 2) A volume filter is applied to the population to restrict the search space to reasonable structures, and so the volume of the unit cell must be minimized as much as possible. The second reason is also important when using Quantum Espresso because the cost of a volume minimizing step in QE is much higher than simply minimization based on steric hindrance due to the energy and force calculations involved in QE minimization. In MGAC1, suboptimal volume optimization was not a problem because CHARMM could perform optimization steps very quickly, but in MGAC1-QE, when the evaluation method was switched to DFT-D, this was a significant issue due to the much longer calculation times. The design of the fitcell routine is also important because a poor design can result in undesirable bias, as will be shown later in this chapter. Also, as was mentioned in Chapter 1, new parameters were added to the genetic schema for MGAC, so the rational for that addition is presented in this chapter. In the course of performing predictions with MGAC1-QE on glycine, a peculiar tendency in the structure

generation was identified that caused problems with the prediction of Gamma glycine. Gamma glycine is in space group P_{31}/P_{32} (Boldyreva et al., 2003), a space group with a primary screw axis, that is, an axis where the rotations of molecules in symmetry operations are constrained to rotations about one axis, with translations only occurring along that same axis. Careful inspection of the glycine populations showed a tendency to favor an elongated primary axis, which was suspicious. Further analysis was performed by tightening the volume constraints, and it was confirmed over a population of more than 100 individuals that there was a strong preference for structures with an elongated primary axis, a bias which should not exist. This was shown to be true in P_{31}/P_{32}, as well as P_{21}.

A mathematical analysis of this issue revealed a deficiency in the schema formulation with respect to the unit cell lengths. It is generally true that the unit cell lengths are constrained by the shape and orientation of molecules contained within the unit cell. However, some space groups need an additional term to take cell length ratios into account, an issue which is most effectively demonstrated by comparison of two extreme types of cells in P_{31}/P_{32}. In P_{31}/P_{32} there are two forms of unit cell, the thin elongated form and the flat wide form, shown in Figure 3.1. All other cells can be formed as an approximate ratio of these two structures. In each unit cell the molecules are labelled 1, 2, and 3, corresponding to their respective symmetry operations as defined in the ITC tables (Hahn, 2002); because the only nontrivial operation is a screw axis rotation, each molecule rotates by $2\pi/3$ radians about the central axis, and is translated along the central axis by one third of the unit cell length relative to the previous symmetry element. If we consider stacking (or occlusion) of molecules along the main screw axis, we see that there are two different arrangements of molecules: the elongated unit cell, in which the interaction pattern along the primary axis is -1-2-3-1-2-3-, and the flat unit cell, where there are three separate stacking patterns along the primary axis of the form -1-1-1-,

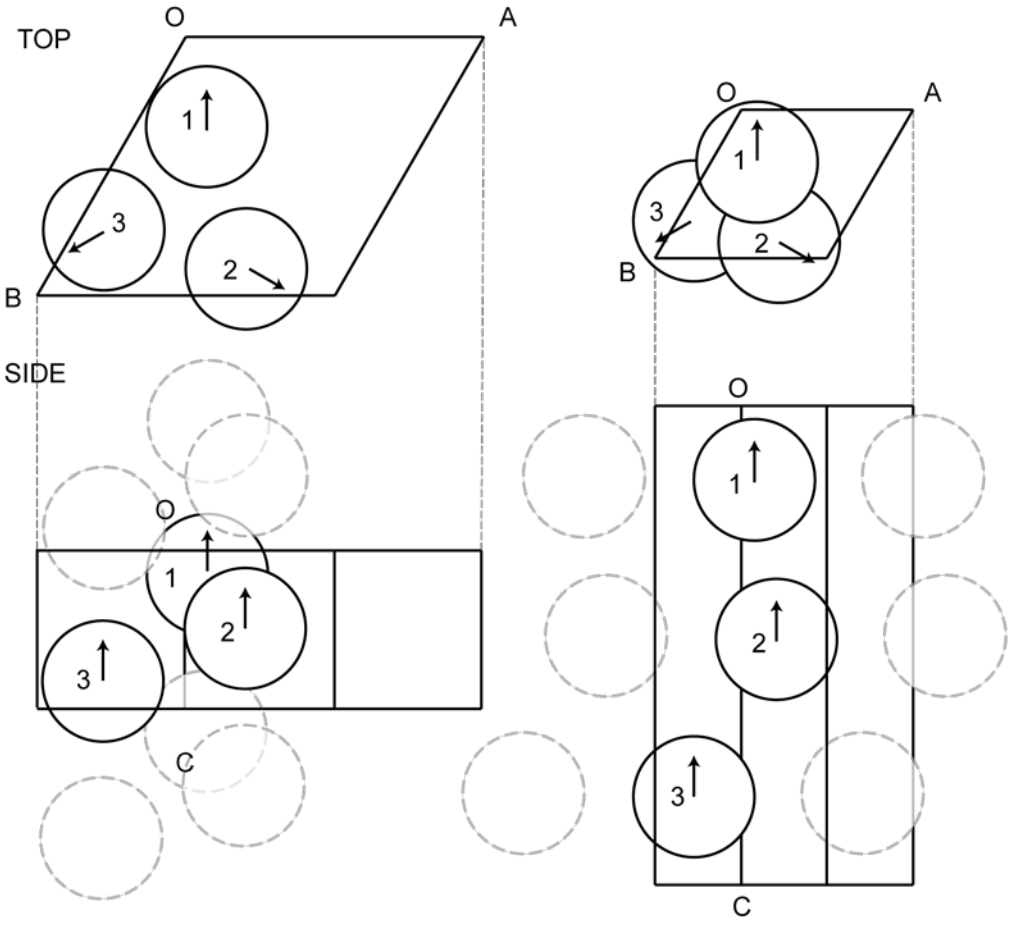


Figure 3.1: Hypothetical flat (left) and elongated (right) unit cells in space group P_{31}. The arrows indicate relative orientations of each molecule, which are represented by the spheres. In the side view, extensions of the lattice are represented by dotted circles. In the flat form, symmetry equivalent molecules form rods, whereas in the elongated form the symmetry equivalent molecules form sheets or planes.

-2-2-2-, and -3-3-3-. Taking into account full translational symmetry, the distinguishing factor is that symmetry equivalent molecules form planar sheets in the flat form, and rods in the elongated form. Because the symmetry elements of the space group constrain rotations to only one axis, it is impossible to pick a different set of cell parameters such that these two separate unit cells can be represented independently in the absence of a term representing the unit cell ratios. The MGAC1 schema coupled with the structure generation algorithm (named fitcell) suffers from this deficiency by favoring the elongated form.

Because the addition of new schema elements required modification to the fitcell algorithm, fitcell has been rewritten in the more recent versions of MGAC. Furthermore, the MGAC1 fitcell was heavily dependent on CHARMM to perform certain operations; this dependency needed to be removed to achieve the goal of having MGAC be completely open source. The remainder of this chapter will comprise two sections: in the first, the original fitcell in MGAC1 will be described along with some additional issues needed to fully understand our methods. In the second section, the new version of fitcell will be described in detail, along with potential pitfalls and improvements that could still be implemented in the design of the algorithm.

MGAC1 Fitcell

The original fitcell algorithm as derived from the MGAC1 source code[3] follows:

1. Prior to the generation of the unit cell, the dihedrals angles are implicitly applied and the rotations are performed on the molecules. Symmetry-based rotations are also applied to individual molecules where applicable.

[3] The original description of the fitcell algorithm was written in Spanish by a former member of the Facelli group (V. Bazterra, PhD Dissertation, University of Buenos Aires), so the algorithm is presented here in English. Because changes to the algorithm might have been made since the original deposition of the fitcell algorithm, the algorithm shown is derived from the implementation in the most recent versions of MGAC1.

2. The unit cell parameters are checked and constrained based on the lattice type of the respective space group. If the unit cell factor (the component of the volume calculation excluding cell lengths) is below 0.1, then it is rejected. The cell lengths are set to a very large value, dependent on molecule size.
3. The molecules are placed in the unit cell and positioned based on the fractional symmetry positions, taking into account the position component of the schema. Rotations or inversions resulting from symmetry elements of the space group are also applied.
4. Once the molecules are placed, a step-wise scaling process occurs. Through the effects of steps 2 and 3, each molecule should be contained in a box that matches the lattice shape. The box is scaled until every atom of the molecule is completely contained by the box, plus an additional buffer volume defined by the distcell parameter.
5. Once the scaling is complete, the volume of the unit cell is calculated.

Generally speaking, the majority of the original fitcell was designed well, aside from some issues with the order of operations which resulted in some computational inefficiencies. Moreover, because MGAC1 was previously successful in making predictions (Bazterra et al., 2002a; Kim et al., 2009), it indicates that under the right conditions this algorithm worked. However, step 4, the volume minimization step, creates a bias that results in problems for a substantial number of searches.

Figure 3.2 illustrates the issue caused in this step of the original fitcell. The outer box represents the asymmetric unit cell of an arbitrary space group, with the dashed line representing the buffer volume around the molecule, and the ellipse representing the shape of the molecule. Because the molecule is completely encapsulated in the asymmetric unit cell, it naturally excludes a large number of valid structures, including but not limited to all molecules that have effective Z'=0.5. Put differently, because the molecule cannot

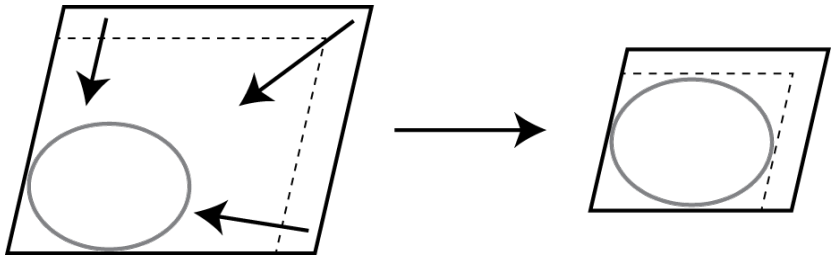


Figure 3.2: The final volume minimization step of the MGAC1 fitcell algorithm. The cell lengths of the unit cell are minimized until a thin volume encapsulates each molecule, the boundary of which is denoted by the dashed line. In space groups with multiple molecules this boundary is established around the asymmetric unit cell. In many structures deposited in the CSD database, molecules will cross the unit cell boundary, highlighting the disadvantage of using this algorithm, because it unnaturally excludes those valid structures.

cross the boundary of the unit cell –or– the asymmetric unit, configurations that form valid space groups by centering molecules at one of the vertices of the unit cell are excluded from the search. A simple cursory inspection of the CSD database (Allen, 2002) further reveals that a significant number of crystal structures have this property, and so this is an undesirable trait of the original fitcell algorithm.

A secondary problem is the treatment of volume restriction in this case. Because volume restriction must happen before optimization and energy evaluation, it is impossible to perform any kind of volume reduction. The presence of the buffer zone, although tunable, leads to volume inflation of the unit cells produced by fitcell. Although this is technically allowable, it can lead to some issues during optimization. As mentioned before, the presence this buffer area will increase the optimization steps required because of the increased volume reduction needed, which is especially problematic in Quantum Espresso. In QE this can also lead to the “radial fft” error (see source code, Giannozzi et al., 2009), which is a problem with volume changes impacting the accuracy of performing Fourier transforms on point meshes, requiring troublesome restarts of QE to preserve calculation quality.

Some additional issues that were present were problems with consistency in rotations and dihedral application. In some cases, the orthogonality of rotation matrices was not preserved, leading to instances where the shape of the molecule could be skewed. In both the cases of dihedrals and rotations, operations were applied somewhat inconsistently, in that zeroing operations were not performed consistently to validate the operations being output. Also, because these operations were not incorporated in the fitcell algorithm efficiently, there were several optimizations identified that could reduce workload. The combination of these factors led to problematic bias issues and other problems when the integration of QE with MGAC1 was undertaken. The next section discusses a number of changes that were implemented in MGAC2 to overcome these

issues.

MGAC2 Fitcell

The algorithmic steps in the new fitcell are detailed as follows (and shown in Figure 3.3):

1. Any previous fitcell operations are undone through the removal of symmetry equivalent molecules.
2. The unit cell lattice type is enforced, with mathematically invalid unit cells being rejected as detailed in Foadi and Evans (2011). (Invocation of this check before dihedral applications is preferable for optimization reasons).
3. The center of mass of each molecule is calculated based on atomic positions and the coordinates of the molecules are adjusted to be centered about the origin in Cartesian space.
4. For each molecule in the asymmetric unit cell, dihedral modifications and rotations are applied. Since the previous configurational state of the molecule might be unknown, these steps include a zeroing step to arbitrary angles and rotations. The connectivity of each molecule after dihedral modifications is tested for violation of steric hindrance before rotations are applied.
5. The space group operations are collected, and the symmetry based rotations are applied to copies of the molecule. Fractional positions based on the schema and space group operations are assigned at this point, but they are not applied to the actual atomic coordinates.
6. A supercell of 3 x 3 x 3 unit cells is generated from the symmetrized and rotated unit cell. This supercell is given unit cell lengths equal to the ratios of the unit cell lengths.

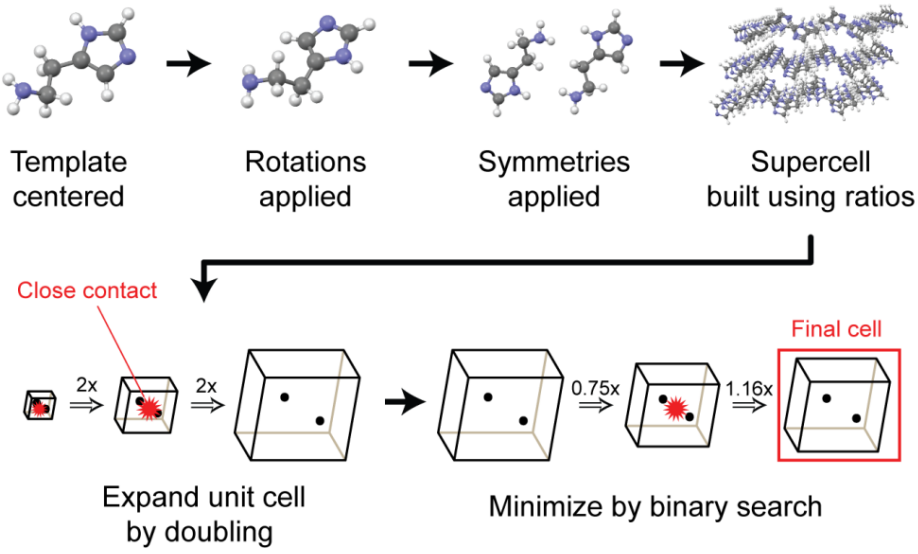


Figure 3.3: The MGAC2 fitcell process. In the initial steps the molecule is adjusted so the center of mass is at the origin in Cartesian space, followed by adjustment of the dihedral angles and rotation of the molecule. Following that, the symmetric copies of the molecule are generated and 3 x 3 x 3 supercell is built. The cell lengths of the supercell are doubled until there are no inter molecular contacts (represented by the red starburst). The cell lengths are then minimized over the range specified by the two cell lengths using a binary search until there are no contacts and the change in distance is less than 0.125 angstroms.

7. The unit cell size is increased by doubling the unit cell lengths until all atoms interatomic distance shorter than the combined Van der Waals radii (Bondi, 1964; Rowland R., 1996) of both atoms plus some buffer distance (typically 0.1 angstroms).
8. Once all molecules are no longer touching each other, the unit cell size is decreased using a binary search; the length of the unit cells is decreased or increased by half the previous change in length until all molecules are no longer touching and the change in all cell lengths is less than 0.125 angstroms.

The primary difference between this fitcell and the previous version, aside from the introduction of ratios, is that the molecular positions are not constrained by the asymmetric unit cell, but instead only by steric hindrance between molecules. This allows for much improved packing properties and removes the issues presented in the previous method. With this new algorithm, most bias issues are removed and unit cells for glycine and histamine that were previously unobserved in MGAC1 predictions have been readily produced.

A second set of improvements is the approach to dihedrals and rotations; anywhere a rotation matrix is used in the algorithm, a stabilization method is used to verify the orthogonality of the matrix so that skew operations do not happen. Furthermore, the application of rotations and dihedrals happens in a logical order so that calculations are not duplicated unnecessarily, leading to improved efficiency. A caveat to this is that, although the matrices are stabilized, the method used still allows a limited amount of numerical drift. However, because of the structural optimization steps performed prior to energy evaluation, the effects of this drift are eliminated in the calculation of the final energies.

Although this new fitcell is improved in many respects, there are new unresolved issues that need to be addressed in the new fitcell. In particular, the treatment of co-

crystals and systems with Z' > 1 is not handled in an effective manner in this new fitcell algorithm. The introduction of extra molecules makes finding efficiently packed structures much more difficult and computationally expensive. The essential step that needs to be taken in this endeavor is the construction of a globular asymmetric complex of molecules based on genome properties. A first approximation to the correct means of achieving this follows:

1. The order of molecular placement is established via arbitrary genome encoding. Each molecule is then placed in the globular unit based on that order.
2. The first molecule is set to its zero rotation and centered about the origin. This is the start of the globular unit.
3. For each successive molecule, dihedrals are applied, and then the molecule is rotated according to its internal rotation gene. Then, the molecule is added to the globular unit, using the position vector from the molecules gene as a basis for the position. The distance between the molecule and globular unit is increased and then decreased using the same binary method used to establish optimize volume in the main fitcell routine until the distance between the molecule and globular unit is optimized.
4. Once all molecules are placed, the entire globular unit is rotated according to the rotation matrix of the first molecule in the unit cell. This unit is then treated as a single entity by fitcell.

The proposed fitcell method, although promising, may or may not escape the issue presented in the current fitcell. Particularly troubling is the establishment of molecular placement order. It is not clear what the best way to approach this would be, for example, if the placement order should be mutable for a given molecular system or not. This is also an issue with respect to the use of the position vectors as described in step 3; can position vectors be used in the proposed way for both fractional placement in the unit cell and as a

relative position vector? These problems will need to be addressed before MGAC2 is ready to be used for more complex molecular systems.

One tempting potential method that was initially explored but abandoned is the optimization of unit cell positions and cell ratios. Optimization of both terms has proven to be problematic in implementation and effect. Attempts at optimizing molecule positions in the unit cell during fitcell produced inconsistent results and were difficult to validate as useful initially. Additionally, problems arose in certain space groups where the positions of the molecules almost universally collapsed to identical points in space regardless of starting position, leading to extreme bias. Optimizing the cell ratios led to a different set of problems; although it was still possible to overcome the bias issues of the original fitcell while optimizing cell ratios, the optimization process led to flat or elongated unit cells depending on the starting ratios. Consequently, the results of these cursory studies into positional and ratio optimizations were rejected in favor of allowing the GA to handle the optimization of those parameters.

The fitcell algorithm described in this chapter was validated by the results presented in the next two chapters. In Chapter 4, the method was shown to be successful in CSP searches for three polymorphs of glycine. Chapter 5 repeats the experiment with the flexible molecule histamine, and was also successful in predicting the natural structure.

CHAPTER 4

CRYSTAL STRUCTURE PREDICTION FROM FIRST PRINCIPLES: THE CRYSTAL STRUCTURES OF GLYCINE[4]

Albert M. Lund,[a,b] Gabriel I. Pagola,[d] Anita M. Orendt,[b] Marta B. Ferraro,[d] and Julio C. Facelli [b,c]

[a]Department of Chemistry, [b]Center for High Performance Computing, and [c]Department of Biomedical Informatics, University of Utah, 155 South 1452 East Room 405, Salt Lake City, UT 84112-0190, US

[d]Departamento de Física, and Ifiba (CONICET) Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. I (1428), Buenos Aires, Argentina

Abstract

Here we present the results of our unbiased searches of glycine polymorphs obtained using the Genetic Algorithms search implemented in Modified Genetic Algorithm for Crystals coupled with the local optimization and energy evaluation provided by Quantum Espresso. We demonstrate that it is possible to predict the crystal structures of a biomedical molecule using solely first principles calculations. We were able to find all the ambient pressure stable glycine polymorphs, which are found in the same energetic ordering as observed experimentally and the agreement between the experimental and

[4] Reprinted from Chemical Physics Letters, 626, pp 20-24, Copyright 2015, with permission from Elsevier.

predicted structures is of such accuracy that the two are visually almost indistinguishable.

Introduction

More than a decade ago Professor Desiraju published (Desiraju, 1997) a critical article identifying crystal structure prediction as one of the most important unsolved problems in computational material science and questioned if this problem could ever be solved. Since 1997 there has been much effort towards the goal of being able to readily and reliably predict, by computational methods alone, the crystal structure of a molecule based only on its chemical diagram (Bardwell DA et al., 2011; Kendrick et al., 2011; Lehmann, 2011; Day, 2012). The process to do this is depicted in Figure 4.1.

The ability to accomplish this goal has far reaching implications well beyond just intellectual curiosity. On a basic science level, this can lead to an understanding of the principles that control crystal growth, by providing accurate information on the crystal energetics necessary for any further dynamical model of aggregation. More practically, the ability to successfully predict crystal structures based on computation alone will have a significant impact in many industries for which crystal structure and stability plays a critical role in product formulation and manufacturing, including pharmaceuticals, agrochemicals, pigments, dyes and explosives (Datta and Grant, 2004).

The current status of crystal structure prediction (CSP) can be evaluated by the performance of the participants in the periodic blind tests that have been organized by the Cambridge Crystallographic Data Centre (CCDC) (Day et al., 2005, 2009; Bardwell DA et al., 2011). The results of the last two blind tests showed the advantage of using dispersion corrected density functional theory (DFT-D) (Grimme, 2004, 2006; Grimme et al., 2010) to create a tailored molecule specific force field that is used to generate trial structures and to reorder a subset of the trial structures in search of the lowest energy crystal structures (Neumann and Perrin, 2005; Neumann, 2007, 2008; Neumann et al., 2008; Kendrick et

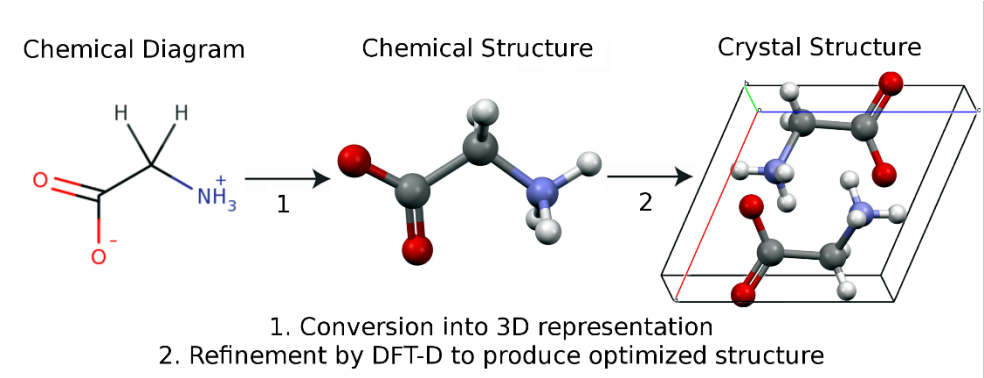


Figure 4.1. Overview of the Crystal Structure Prediction (CSP) process, which attempts to predict the structure or structures (when polymorphs exist) of a molecular entity based solely in its chemical diagram. The prediction of these structures directly from first principles has been identified as one of the greatest challenges remaining in computational molecular sciences.

al., 2011). The software used in this approach is proprietary. There has also been an attempt to utilize first principle calculations to predict the crystal structure of organic crystals (Zhu et al., 2012), in which the authors used a two-step approach optimizing the crystal structures with constrained molecular geometries in the initial stages and allowing full relaxation in the final stages. In this approach the authors used a combination of open source and proprietary software tools for the constrained and fully relaxed optimizations.

These results lend promise to using DFT-D methods to completely replace molecular mechanics and/or multistep optimization approaches as the method of choice for the evaluation of the energies of the trial crystal structures in CSP. To the authors knowledge there are no reports of any open source software capable of successfully predicting crystal structures of molecules of biomedical interest directly from first principles without using either common or tailored potentials as intermediate steps and/or multistep optimization strategies.

It is important to realize that local optimization of plausible crystal structures is not a feasible approach for CSP. We have recently demonstrated (Lund et al., 2013), using a set of drug like molecules, that local optimization using full DFT-D results in near experimental structures only when the starting point is quite close to the experimental one. Therefore, global optimization with a reliable and universal energy function is necessary for accurate CSP.

The MGAC (Modified Genetic Algorithm for Crystals) package has been developed in our lab over the last decade (Bazterra et al., 2002a, 2002b, 2004, 2007; Kim et al., 2009). MGAC is capable of doing CSP for any space group, any number of molecules per asymmetric unit, and can take into account the conformational flexibility of the molecule both at the local and global optimization levels. This allows an efficient, GA (Genetic Algorithms) based, global exploration of the crystal energy landscape. The previously released versions of MGAC relied on the use of the CHARMM (Brooks et al., 1983;

MacKerell et al., 1998) molecular mechanics program using the Generalized Atomic Force Field (GAFF) (Wang et al., 2006) for the energy evaluation and local minimization of the GA trial structures.

Previously, we used the set of molecules present in the Karamertzanis and Price (K&P) paper (Karamertzanis and Price, 2006), to demonstrate the capabilities of the MGAC-CHARMM program (Kim et al., 2009). These results demonstrated that the implementation of the GA in MGAC was effective and was always able to find the correct experimental structures provided that the GAFF potential energy represented the experimental energy landscape with sufficient fidelity. However, the matches to the experimental structure ranged from rank 1 to rank 1182 in terms of energy, highlighting the second issue with the use of the generic force field, namely the unreliability of the energy ranking.

Our more recent work (Lund et al., 2013) has demonstrated that when using Quantum Espresso (QE) to locally optimize the experimental structures in the K&P set the calculated local minima structure compares well with the experimental structure in all 32 of the molecules, with RMS differences ranging from 0.056 to 0.459 Å. This implies that for unknown structures, an approach which couples the use of MGAC with energy evaluation using QE will be successful in finding the "true" experimental structures.

In this article we report the results of our unbiased searches for glycine polymorphs obtained using the global GA search implemented in MGAC coupled with local optimizations and energetics provided by QE (MGAC1-QE). To our knowledge here we demonstrate for the first time that it is possible to predict the crystal structure of a molecule of biomedical interest, glycine, using solely first principles calculations (DFT-D) of the crystal energetics without using any intermediate steps, such as constructing special interatomic potentials, reordering the structures found by the search algorithm and/or using multistep search strategies with nonuniform approximations for the energy

calculations. The only difference in the calculations presented here and a complete blind CSP search is that we only performed searches in the known space groups of each of the three stable polymorphs of glycine.

Methods

Using the existing MGAC framework we have integrated the QE calculation of the energy and local optimizations into the framework as well as reworked the way in which the initial populations are selected and how the genetic algorithms were implemented (MGAC1-QE). A full account of the technical and computational details of the integration of QE into the MGAC framework will be presented in detail elsewhere, along with the documentation and instructions on how to use the software that we will make available as an open source tool.

Glycine's biological interest, relatively small size and polymorphic characteristics make it a good case to demonstrate the ability of MGAC1-QE to predict the crystal structures of biomedical relevant compounds. Glycine is a precursor to the synthesis of proteins, a building block to numerous natural products, and provides the central C_2N subunit of all purines. It is a relatively small, semi rigid molecule, for which polymorphism is well established in the literature. The existence of polymorphism is critical to demonstrate the usefulness of MGAC1-QE to successfully predict crystal structures of biomedical interest for which the existence of polymorphism is prevalent (Datta and Grant, 2004).

Glycine has three room temperature and atmospheric pressure polymorphs: α-glycine ($P_{21/c}$) (Aree and Bürgi, 2012), β-glycine (P_{21}) (Tumanov et al., 2008), and γ-glycine (P_{31}/P_{32}) (Boldyreva et al., 2003) (stability order: γ-glycine > α-glycine > β-glycine), as well as two high pressure polymorphs, δ-glycine (high pressure of the β-glycine form) (Tumanov et al., 2008), and ε-glycine (the high pressure form of the γ-glycine form)

(Boldyreva et al., 2005). For the purpose of comparison of our results we used the following glycine reference structures from the Cambridge Structural Database (CSD): GLYCIN98 for α-glycine (Aree and Bürgi, 2012), GLYCIN71 for β-glycine (Tumanov et al., 2008), and GLYCIN33 for γ-glycine (Boldyreva et al., 2003). These three experimental structures were locally optimized using the QE vc-relax option, which allows for optimization of the unit cell parameters along with all atomic coordinates, with the same QE parameters used in our previous work (Lund et al., 2013). In all cases the experimental structures converged to local minima in close proximity to the experimental structures. The QE energies for these local minima structures are $E_{\alpha\text{-glycine}}$ = -147,662.07 kJ/mol, $E_{\beta\text{-glycine}}$= -147,659.78 kJ/mol, and $E_{\gamma\text{-glycine}}$ = -147,663.10 kJ/mol, which reproduce the experimental stability order: γ-glycine > α-glycine > β-glycine.

Following these preliminary tests we conducted unbiased global searches for crystal structures in the following space groups, with a number of molecules per unit cell given in parenthesis: $P_{21/c}$ (4), P_{21} (2) and P_{31} (3). All calculations were performed using a population size of 120 individuals, a replacement rate of 1.0 per generation, and the searches were run for 50 generations. The probability of an individual being mutated was 0.01, and the probability of a crossover occurring between two individuals was 1. The selection method was a roulette wheel, using linear scaling of the energy, with the lowest energy structure having the largest selection probability. The optimization parameters for the QE optimization were again identical to those used by Lund et al. (2013) The DFT functional used was the Perdew-Burke-Ernzerhof generalized gradient approximation (Perdew et al., 1996c). The dispersion correction method selected was the semiempirical D2 method proposed by Grimme as implemented in Quantum Espresso (Grimme, 2006). The self-consistency threshold was set to 10^{-7} Ry and the plane wave cutoff energy was set to 55 Ry per the recommendation of the pseudopotentials creators. The pseudopotentials used for glycine were the Rappe-Rabe-Kaxiras-Joannopoulos-Ultrasoft pseudopotentials

provided at the QE website, http://www.quantum-espresso.org/.

Calculations were performed on a LINUX cluster using six 16-core nodes (2 x 8-core Intel Xeon E5-2670 processors clocked at 2.60 GHz), with 64 GB memory per node and Mellanox FDR Infiniband for node interconnectivity. The total number of core hours for each run was: α-glycine: 10,238 core hours; β-glycine: 7,174 core hours; and γ-glycine: 9,518 core hours. Therefore, the total number of core hours used for these three searches was 26,930, which represent a total elapsed time of approximately 12 days.

Results and Discussion

The results of the analysis of the populations generated by the MGAC-QE runs described above are presented in Figure 4.2. This figure presents, as suggested by Price (Price, 2009), the distribution of the energies of crystals in the MGAC populations as function of their volume. As expected when polymorphism is present, the plot shows a great deal of crowding and the volume energy pairs of the different polymorphs are not well separated (Price, 2009). This clustering of the three polymorphs reinforces that glycine is a challenging case for CSP, and therefore a stringent test for the MGAC-QE method. From the figure it is also apparent that the structures found by MGAC-QE (solid symbols) for each of the symmetry groups studied here closely match the experimental ones (hatched symbols) corresponding to the most stable polymorphs, in the same space group. Notably, in some initial generations we observed structures where the protonation state of glycine was altered and the non-zwitterionic form was adopted. This is made possible by the unconstrained optimization algorithm in QE. These structures were typically much higher in energy (by >80 kJ/mol) than structures remaining in the zwitterionic form, and were therefore eliminated rapidly from the population. The conclusion drawn from this is that one must be careful to identify low energy structures where the protonation state (and in general, bonding state) might be altered.

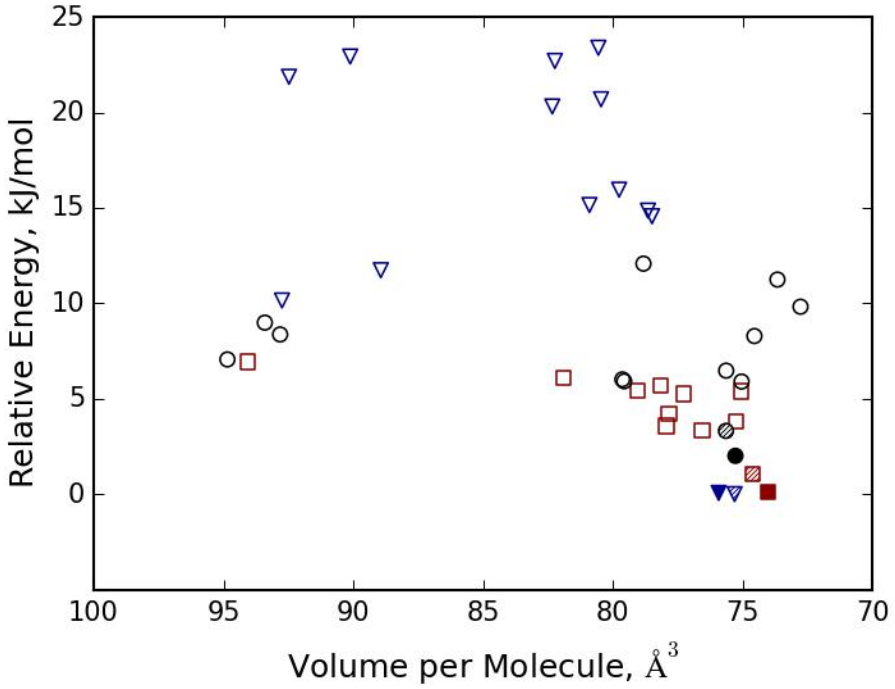


Figure 4.2. The distribution of the energies of crystals in the MGAC-QE populations as a function of the molecular volume. The crystals structures corresponding to the $P_{21/c}$, P_{21}, and P_{31} GA runs are represented by squares, circles, and triangles, respectively. The hatched markers correspond to the experimental structures and the solid ones to the lowest energy structures found by MGAC-QE.

In Table 4.1 the crystallographic parameters and the calculated energies of the best structures found by MGAC-QE for each of the space groups considered here along with the RMS between them and the corresponding experimental structures are given. The results in Table 4.1 show an excellent agreement between the MGAC-QE predicted structures and the experimental ones; the agreement is apparent in both in the cell parameters as well as the RMS difference between the experimental and predicted structures. The RMS values can be compared to the RMS values observed when comparing different experimental structures of the same polymorphs reported in the CSD; for instance, the RMS between α-glycine structures GLYCINE89 and GLYCINE17 is 0.026 Å, for β-glycine structures GLYCIN74 and GLYCINE25 is 0.114 Å and for γ-glycine structures GLYCIN65 and GLYCIN15 is 0.07 Å.

The energies of the MGAC-QE predicted structures follow the experimental stability order: $E_{\gamma\text{-glycine}} < E_{\alpha\text{-glycine}} < E_{\beta\text{-glycine}}$, with α-glycine and β-glycine 70 J/mol and 1,950 J/mol, respectively, less stable than γ-glycine. These values can be compared with recent values from the literature[34] of 962 J/mol and 1,506 J/mol, respectively, obtained using the DFT method plus many body dispersion correction and zero point energy corrections (PBeh + MBD +ZPE) (Marom et al., 2013).

The graphical comparison between the experimental and the best MGAC-QE structures is presented in Figs. 4.3-4.5. This comparison does not require additional discussion, as it is apparent that the agreement is of such quality that the two structures are almost indistinguishable.

In conclusion, using MGAC-QE we were able to find each of the ambient pressure stable polymorphs of glycine when searching in their corresponding space group. The match to the experimental structure was the lowest energy structure found in each of the three searches. The polymorphs encountered by MGAC-QE are energetically ordered in agreement with experimental results and the comparison of the experimental and

Table 4.1. Comparison of the energies and geometries of the α-glycine, β-glycine, and γ-glycine structures found by MGAC-QE with the reference experimental structures.

Polymorph		SPG	Energy [d]	Cell Parameters [e]						RMS [f]
				a	b	c	α	β	γ	
α-glycine	MGAC-QE	$P_{21/c}$	-147,663.00	5.0517	11.7146	5.7965	90.0	120.3102	90.0	0.097
	Exp[a]			5.0874	11.7817	5.4635	90.0	112.0530	90.0	
β-glycine	MGAC-QE	P_{21}	-147,661.12	5.6840	6.0727	5.0305	90.0	119.8711	90.0	0.199
	Exp[b]			5.3880	6.2760	5.0905	90.0	113.1200	90.0	
γ-glycine	MGAC-QE	P_{31}	-147,663.07	6.9166	6.9166	5.4983	90.0	90.0	120.0	0.087
	Exp[c]			7.0383	7.0383	5.4813	90.0	90.0	120.0	

[a] Structure GLYCIN98 (10 K) from (Aree and Bürgi, 2012).
[b] Structure GLYCIN71 (room temperature) from (Tumanov et al., 2008).
[c] Structure GLYCIN33 (room temperature) from (Boldyreva et al., 2003).
[d] Energy in kJ/mol for the lowest energy found by MGAC-QE in the corresponding space group.
[e] Crystallographic axis in Å, cell angles in degrees.
[f] Computed using the Solid Form Crystal Packing Similarity method in Mercury CSD with 15 molecules for comparison and ignoring hydrogen atoms (Chisholm and Motherwell, 2005).

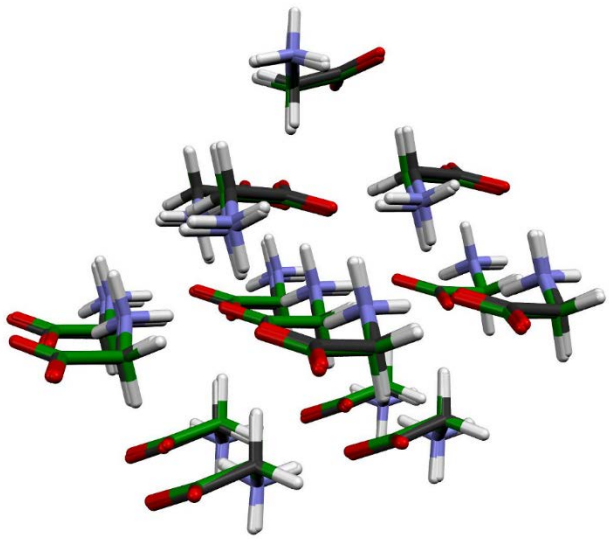


Figure 4.3. Comparison of the experimental structure of the α-glycine CSD structure GLYCIN98 from (Aree and Bürgi, 2012) (black) with the lowest energy structure found by MGAC-QE in the $P_{21/c}$ space group (green).

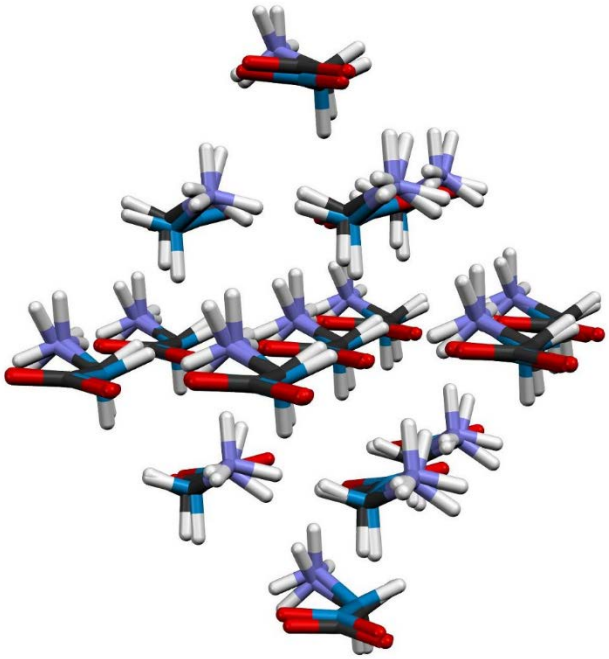


Figure 4.4. Comparison of the experimental structure of the β-glycine CSD structure GLYCIN71 from (Tumanov et al., 2008) (black) with the lowest energy structure found by MGAC-QE in the P_{21} space group (blue).

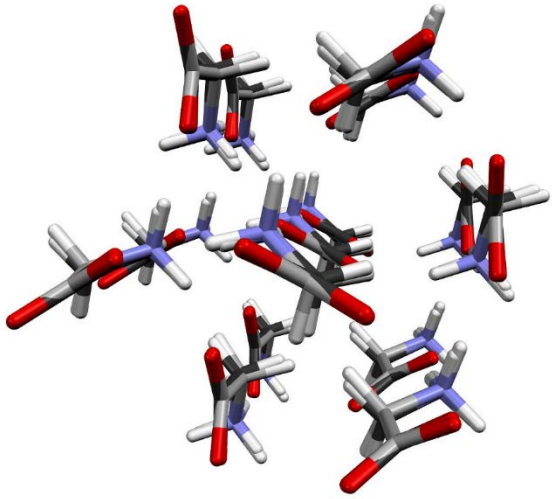


Figure 4.5. Comparison of the experimental structure of the γ-glycine CSD structure GLYCIN33 from (Boldyreva et al., 2003) (black) with the lowest energy structure found by MGAC-QE in the P_{31} space group (white).

predicted structure is of such accuracy that the two are visually almost indistinguishable. When the success of MGAC-QE is compared with the results for glycine in Zhu et al. (2012), it becomes apparent that allowing the full relaxation of both molecular and crystal structural parameters as well as using a single approach for the calculation of the crystal energies at all stages of the global optimization is critical for successful CSP. However, there is already enough evidence in the literature that current functional and dispersion correction lattice energies may not be adequate for all crystals, particularly disordered ones, therefore new DFT approaches may be needed to address those systems.

The computer times required by the calculations reported here are significant, but manageable. Computer times for larger systems will be a significant challenge, but we are confident that we will be able to greatly improve performance once we better understand the optimal GA parameters like population size, replacement, and number of generations, and are able to make use of emerging computer technologies like GPU accelerators. A truly blind test of the method, exploring most common space groups and/or using searches in *P1* with different number of molecules per unit cell is the next goal. The exact search strategies will be defined by studies that are underway in our laboratory to establish the most efficient search protocols for blind test CSP. The results of this exploration will be used to participate in the current sixth CSP blind test, and our results will be presented at the 2015 Cambridge Crystallographic Data Centre meeting in the fall of 2015.

Conclusions

The results presented here show that it is possible to predict the crystal structures of molecules of biomedical interest from first principles without using any intermediate potentials, energy reordering strategies and/or step wise optimization strategies. With these results we believe that we can answer Professor Desiraju's question with an unquestionable **yes!** Crystal structures can be predicted from first principles and with

existing computational resources and appropriate optimization of our methods, CSP can become a standard tool for material design.

Author Contributions: All authors contributed to the design of the project, analysis of the data, and to the preparation of the manuscript. AML and GIP contributed to the MGAC-QE code development. AML conducted the calculations reported on glycine.

Author Information: Correspondence and requests for materials should be addressed to Professor Julio C. Facelli at the University of Utah: Julio.Facelli@utah.edu. The source code for MGAC-QE will be made available by an open source mechanism once it is sufficiently stable for wide distribution. None of the authors declare any competing financial interest regarding the contents of this paper.

Acknowledgements: Computer resources were provided by the Center for High Performance Computing at the University of Utah and the Extreme Science and Engineering Discovery Environment (XSEDE), supported by NSF grant number ACI-1053575. MBF and GIP acknowledge the support from the University of Buenos Aires and the Argentinean Research Council. For information on the MGAC software, please contact the corresponding author.

CHAPTER 5

DEPOSITION OF HISTAMINE PREDICTION RESULTS

Histamine is a small molecule with three degrees of internal freedom which participates in a number of important physiological and biological processes (Lopez, 2002; Haas et al., 2008; Leurs et al., 2009). It is also included in the Karamertzanis-Price dataset as an example of a pharmaceutically relevant molecule, and so its use as a molecule for validating CSP has been established (Karamertzanis and Price, 2006). In this short study, two independent predictions of histamine are presented to establish the predictive power of MGAC1-QE, to complement the results of Chapter 4.

Methods

Two independent MGAC1-QE runs were performed on histamine in the native space group P21 (CSD designation HISTAN) (Bonnet and Ibers, 1973). The parameters for the GA were set to a population size of 90, with replacement three times the size of the population at each generation (Lund et al., 2015). The volume constraints were set to -50% to 200% of the estimated system volume. The maximum generation cutoff was set to 200. The optimization parameters for QE optimization and energy calculations were set to the same parameters used in Lund et al. (2013). For all optimizations, a maximum of 70 optimization steps were allowed to complete.

Results

Both prediction runs were successful in obtaining the correct structure of histamine. For unknown reasons the first prediction run experienced an error after the

fourth generation, and did not produce any valid structures after that point, however, both predictions were able to generate a matching structure within the first two or three generations. Importantly, in the other prediction run, which completed ten generations, the highest six ranked structures all matched the experimental structure of histamine. Volume-energy plots of both prediction runs are presented in Figure 5.1, where the expected funnel is clearly shown for both structures. The energies of the best structure from runs one and two are -167,063.35 kJ/mol and -167,069.05 kJ/mol, respectively, which are in good agreement with the energy calculated from optimizing the known experimental structure, which is -167,069.34 kJ/mol. Table 5.1 gives the unit cell information for each of best structures and for the experimentally determined histamine structure; the agreement of unit cell parameters is very high for the second run, which is expected, and the parameters for the first run are also quite good, given the low refinement quality. This can be visually verified for both structures in Figure 5.2, which shows an alignment of the best structure from each run with the experimentally determined histamine.

Conclusion

With these results it is apparent that MGAC1-QE is able to predict the structure of histamine. This is a marked improvement over MGAC1, which although successful in predicting histamine, was unable to directly find the structure of histamine in the native space group with Z'=1 in the unit cell. The success of this prediction also supports the assertions in Chapter 1 that a higher level of theory is a valid means of calculating energies for highly flexible molecules. Generally, this can be interpreted as meaning that when the space group is known, MGAC is more than capable of making successful predictions. Given this knowledge, the next major problem is to address the problem of handling full blind test systems, including multiple space groups. In the next chapter, an exploration of this

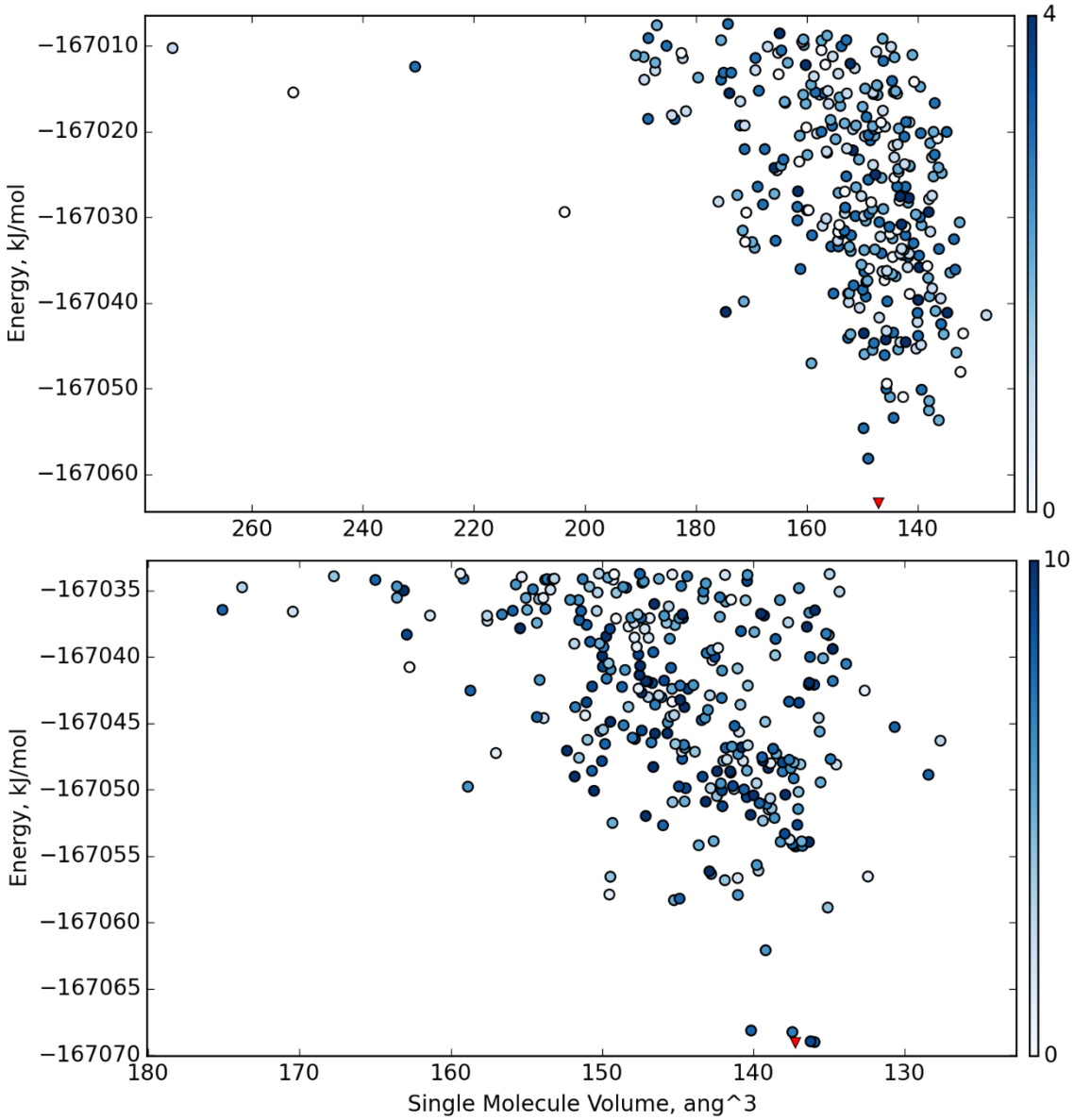


Figure 5.1: The volume energy plots of the first (top) and second (bottom) MGAC1-QE prediction runs. In both graphs the color of the data point indicates the generation that the structure evolved. The inverted red triangle represents the lowest energy structure in both plots.

Table 5.1: Parameters for the predicted structures compared against the experimentally determined histamine structure. HISTAN data is from Bonnet and Ibers (1973).

Structure	A (Å)	B (Å)	C (Å)	Beta (deg)	Energy (kJ/mol)	RMS
HISTAN	7.249	7.634	5.698	104.96	-167,069.34	-
Exp 1	7.158	6.803	6.221	103.76	-167,063.35	0.506
Exp 2	7.219	7.087	5.519	103.58	-167,069.05	0.256

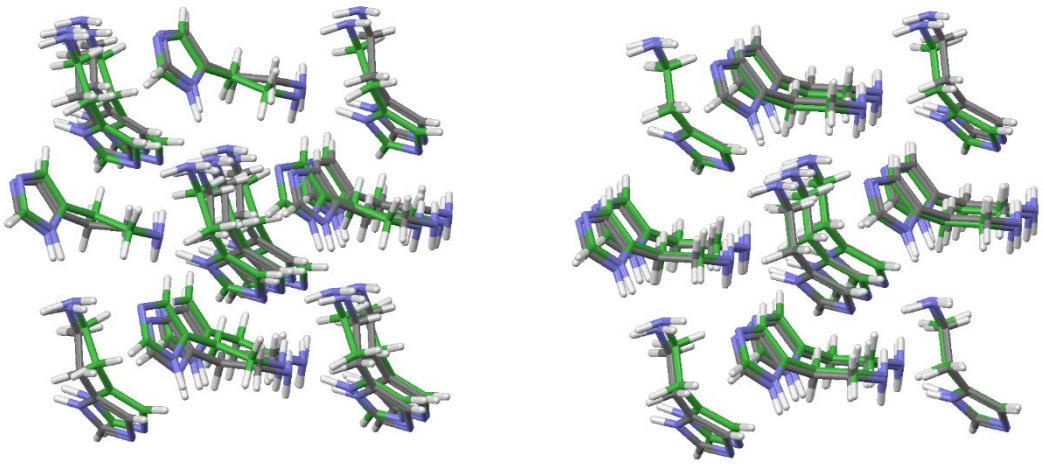


Figure 5.2: Alignments of the predicted structures against the experimentally determined structures of histamine, with the first (RMS 0.506) and second (RMS 0.256) predictions on the left and right, respectively. Grey structures correspond to the experimental structure of histamine, while the green structure represents the predicted structure.

problem area is presented and improvements and additions to the MGAC algorithm are given in the design of a new algorithm for CSP, called MGAC2.

CHAPTER 6

MGAC2: A NEW ALGORITHM FOR CRYSTAL STRUCTURE PREDICTION

Abstract

A new version of the Modified Genetic Algorithm for Crystals is presented in this chapter. The new MGAC algorithm has been enhanced and modernized for use with the density functional theory software Quantum Espresso, and to take advantage of modern computing architectures. As discussed in Chapters 4 and 5, multiple polymorphs of glycine and the experimental structure of histamine, respectively, were successfully predicted using MGAC1-QE. Despite the success of those predictions, it became apparent that there were several problems with the design of MGAC1-QE that resulted in inefficient use of computational resources. Furthermore, it was determined that the genetic algorithm in use had never been thoroughly refined because of the relatively low computational cost of using CHARMM as a fitness evaluator.

In this new algorithm presented here (Figure 6.1), these issues are addressed, and a new genome representation that eliminates the need for multiple independent searches in single space groups is presented. Several changes to the design and use of genetic operators are outlined, and a number of technical changes are discussed. Finally, a variable population size strategy for steady state genetic algorithms that will allow for much better scalability and use of computation resources is presented.

Introduction

In 2002 the Facelli group developed the Modified Genetic Algorithm for Crystal and Clusters (MGAC) to answer the call to solve the problem of Crystal Structure

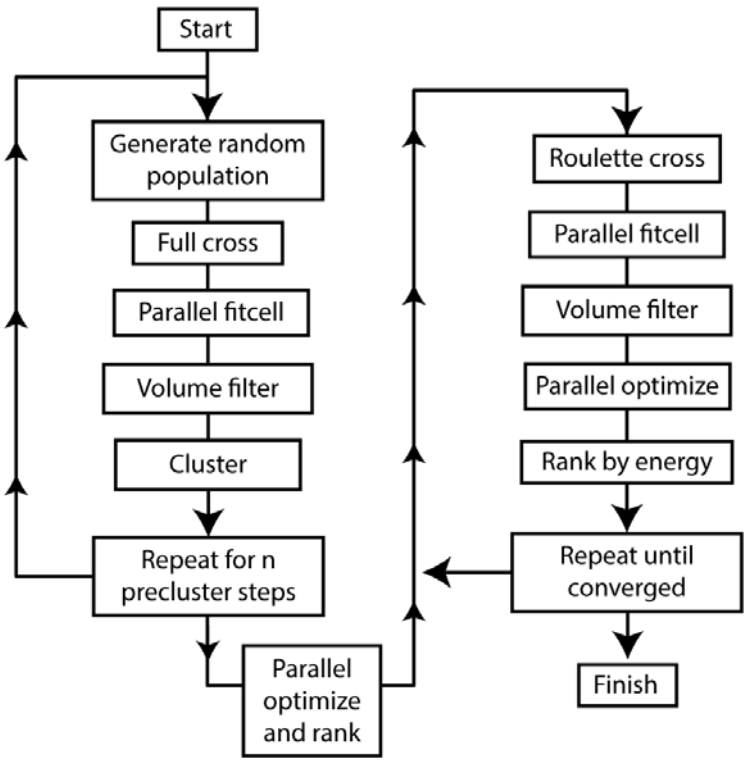


Figure 6.1: The complete MGAC2 algorithm layout. The algorithm goes through two phases; in the first phase, an initial population of structures is generated through a clustering process (left column). Once this population has been generated the population is evaluated using Quantum Espresso. In the second phase, a step-wise elitism model is used to iteratively refine the population.

Prediction (CSP) for small organic molecules. As highlighted in other sections of this dissertation, CSP has importance for the pharmaceutical and explosives industries (Datta and Grant, 2004; Deschamps et al., 2008), as has been highlighted by the NSF Assistant Director for Mathematical and Physical Sciences (Crim, 2014). To date MGAC has been used to make predictions on a variety of molecules, and has made predictions as part of the periodic blind tests held by the Cambridge Crystallographic Data Centre (CCDC) (Bardwell DA et al., 2011). Results have been varied; earlier predictions using MGAC1 relied on CHARMM and the General Amber Force Field, and had high variability in prediction quality, and suffered from bias issues as a consequence of using the GAFF universal empirical potential.

In Chapter 2 the research investigating the viability of dispersion-corrected density functional theory (DFT-D) for use in MGAC was outlined. The tests with Quantum Espresso demonstrated that DFT-D was viable for use in calculating structure energies and provided a significant improvement over CHARMM and the GAFF. Based on those results, a new implementation of MGAC1 using Quantum Espresso, called MGAC1-QE in this discussion, was developed. This implementation was used to successfully predict multiple polymorphs of glycine, as well as histamine, with excellent quality results (see Chapters 4 and 5). The success of these predictions lends credence to the viability of using DFT-D as the sole energy calculation method in CSP.

Despite the success of those results, it was determined that there were a number of overarching issues that would preclude the wide utilization of the current MGAC1 algorithm. First, although the glycine and histamine predictions were successful, they were of limited value because the searches were performed using the *a priori* knowledge of the native space groups for each of the glycine polymorphs and histamine. In MGAC1 the protocol for predicting structures relied on sampling the 14 most common space groups among known crystal structures as characterized, in the CCDC Crystal Structure

Database (Allen, 2002). This statistical reliance would have been deficient in a true blind test of glycine, as the Gamma polymorph is in space group P_{31}/P_{32} (Boldyreva et al., 2003), which is not counted among the 14 most common space groups. Since there are 230 space groups, an individual search in all possible space groups is computationally not feasible, so being able to search all space groups effectively became a high priority for the MGAC research effort. Furthermore, several technical deficiencies were identified in MGAC1-QE that needed to be addressed. In particular, MGAC1 was never designed to handle long running calculations like those performed when using QE, which require substantially more computing time and resources compared to the original CHARMM-based method, and a revision to the restart mechanism to handle intermediate QE optimizations was deemed necessary. MGAC was also designed prior to the advent of commodity multi-core computing hardware, meaning a number of subtasks in the MGAC algorithm could be handled much more efficiently by taking advantage of these architectures. Starting in 2015 the development of a new algorithm was commenced, to be eventually used in a completely new version of MGAC (hereafter referred to as MGAC2), which would be designed to address the above issues, and to incorporate innovations to improve the efficiency of the MGAC structure predictions. The following sections address each of these topics.

Multiple Space Group Schema

A primary concern when designing a genetic algorithm schema is finding the minimum number of degrees of freedom that can be used to represent the phenotype of interest. The MGAC1-QE schema for a single molecule in the unit cell has approximately 12+n schema elements, comprising six elements for the unit cell angles and ratios, six elements for the rotation and position of the molecule in the unit cell, and n elements one for each flexible torsion angle in the molecule. A final term exists to represent the space group, but in MGAC1 this was not implemented as a crossable schema element. Using this

term by itself is impossible because there are 230 crystallographic space groups to sample, which would require a very large population size to properly represent this large search space. At a minimum, at least 20 individuals would be required from each space groups, leading to population sizes in the thousands, which is computationally untenable when using QE.

One initial idea considered was to eliminate the use of higher order space groups altogether, solely using variable numbers of molecules in the P1 space group (Z'=1, 2, 3, 4, 6, 8, etc.), with each molecule possessing its own individual schema. Some initial tests using MGAC1-QE to investigate this approach proved to be highly problematic, however, as it quickly became apparent that even for three or four molecule cases, constructing valid crystal structures was difficult to handle for the genetic algorithm. The reason for this is that by eliminating space groups from the basic representation of the schema, the genetic algorithm was responsible not only for finding volume and energy minimized solutions, but also for solving the basic symmetry operations of each space group. Framed differently, the addition of each new molecule to the crystal system would add 6+n new degrees of freedom (for the position, rotation, and internal flexibilities of the molecule). Ultimately, this is a waste of computational resources, so a more sophisticated method was required.

Being faced with this difficulty, a thorough examination of the 230 crystallographic space groups was commenced by studying the International Tables of Crystallography (ITC handbook) (Hahn, 2002), which is considered the authoritative manual on space group mathematics. Section 4.3 of the ITC handbook organizes the space group symbols based on their underlying mathematical features, as well a number of extended symbol types that deal with degenerate cell settings. A full discussion of this section is out of the scope of this dissertation, but it is important to note that the tables are organized in logical sections that lend themselves well to a genetic algorithm schema. In the ITC handbook

space groups are first ordered by lattice system (e.g., triclinic, monoclinic, tetrahedral), and then by fundamental point group. A full discussion of crystallographic group theory can be found elsewhere (Tinkham, 2003), but essentially, whereas there are infinite point groups, only a subset of 32 point groups are valid when considering translational symmetry. These can be further reduced to 12 point group classes and 5 axis numbers, of which almost every combination is valid.

Besides the crystallographic point groups, there are also other features that distinguish the different space groups within each class. A prominent feature is the face centering type of the space group, which deals with the placement of lattice points in the unit cell. There are five fundamental face centerings, which combine with different lattice types to form 14 different Bravais lattices. Importantly (and unlike the crystallographic point groups), there are significant exclusions in the lattice type/face centering combinations. For example, in monoclinic lattices, only primitive (P) and base-centered (A/B/C) types are allowed, whereas in orthorhombic lattices in addition to the primitive and base-centered types, body- and face-centered types (I and F, respectively) are also allowed. Additional crystallographic features also serve to distinguish groups from each other; the presence of glide plane and screw axis elements are highly important features that are possible in certain point groups that could potentially be used, and an operator for axis order that essentially represents handedness in relevant space groups is also present. These final three terms, however, are highly dependent on the lattice type and are not distributed in any consistent fashion. Summarizing the space group elements identified as potential schema elements:

1. Point group classes (7+5 options):

   C_n, C_{nv}, C_{nh}, S_n, D_n, D_{nd}, D_{nh} + T, T_h, O, T_d, O_h

2. Axis numbers (5 options): n(axis)=1,2,3,4,6
3. Face centerings (5 options): P, A/B/C, I, F, R

4. Axis order (inconsistent options)
5. Glide plane operations (inconsistent options)
6. Screw axis operations (inconsistent options)

These features of the system of crystallographic space groups provide a basis that can be used for designing a new schema for multiple space groups. Furthermore, the use of this basis results in a maximum of six degrees of freedom that are able to represent all possible space groups with a much smaller population.

In the design of the MGAC2 schema, all six elements are used, but the axis order, glide and screw operations are condensed into a single element, leading to a total of four parameters in the GA to represent the symmetry group. Since every combination of point group class and axis number, with the exception of D_{4d} and D_{6d}, form crystallographically valid point groups, these elements are excellent for use in the GA schema. The other four elements, however, do not map consistently between the different point groups. Of the elements 3-6, only face centering is given a unique GA schema element, because almost all point groups have multiple face-centerings. The remaining three are rolled into a single parameter, which is a variable expression gene that maps differently based on point group and face centering. This means that different subclasses of point group and face centerings will have variable groupings. For example, C_{2h} in the P-centering has four possible space groups, whereas D_{4h} in the P-centering group has seventeen possible space groups.

It is quite reasonable to assume that this implementation will result in bias issues, and make some space groups more difficult to access by virtue of the imbalance between space group types. However, this problem is superseded by the statistical distribution of space groups as deposited in the CSD, which is also highly uneven (Allen, 2002):

1. $P_{21/c}$ and P_{-1} space groups dominate almost 75% of all known crystal structures
2. Point class C_{nh} (mostly $P_{21/c}$) comprises 47% of all known crystal structures,

followed by S_n (25%, mostly P_{-1}), with the remaining point group classes comprising 3-8% each.

3. Axis order n=2 comprises 93% of all structures, while remaining axes are 1-2% each.
4. Groups T, T_h, O, T_d, O_h comprise <0.5% of all structures.
5. Most face centerings are P (85%), followed by A/B/C (11.5%), with F, I, R being 1-2% each.

Because of this unequal distribution, the fact that subclasses are not evenly distributed is much less of a problem than would be otherwise. It also provides an impetus for excluding some space groups and providing a boost to certain subtypes. A very easy target are the high order tetrahedral and octahedral space groups; because these are very complex space groups having many symmetry elements (24-196) and because they are so under represented, it is quite logical to have the capability to exclude them from the GA schema. Bias favoring axis order n = 2 can also be built into the algorithm, as well as minor favoritism for C_{nh}. By manipulating the distribution of genomic parameters, the exploration of different space groups can favor those space groups with high presence in the CSD, but care must be taken not to bias too heavily in favor of those space groups at the expense of the others.

As an additional restriction on the GA schema, MGAC2 also needs to be capable of limiting space groups based on the number of symmetry elements. The reason for this is because QE scales quadratically with system size: as an example, a space group with four symmetry elements will take sixteen times as long to complete a single point energy calculation relative to a space group with one symmetry element. This quadratic scaling is extremely problematic for space groups with eight or more symmetry elements, which comprise more than 70% of all space groups, because it creates a serious imbalance in the

potential calculation time between two differing space groups. [5] This sets a practical limit on what space groups can be searched, but the capability to search all space groups needs to be implemented. This is especially true if new methods are developed which permit the search of higher order space groups at lower cost than QE are developed in the future, or if the use of computational accelerators can be made effective with QE.

In initial tests of this schema some issues presented themselves. MGAC2 was originally conceived to use a population of structures in a mixed pool of space groups with a reasonably high population size (n = 300). However, this was deemed impractical fairly quickly, because of the complexity of generating higher order space groups. The observed behavior of the mixed space group population is that simple space groups, especially P_1 and P_{-1}, come to dominate the population very quickly. In low order space groups, the unit cell can be minimized very easily to approximate the shape of one or two molecules. For higher order space groups, the position of the molecule in the unit cell is more likely to impact the volume minimization process due to inversion symmetry elements, which are more prevalent in higher symmetries. This makes it fundamentally more difficult to generate valid structures in the high order space groups, leading to significant bias towards low order space groups in a mixed space group population.

To prevent this, an approach was adopted where different space groups are sorted into individual bins, to prevent the premature loss of space group diversity. Essentially, this is a bookkeeping trick, where the best structures from each space group are maintained and crossed, but not all space groups are considered at all times. In order to reduce computational burden, only the top 10-15 space groups participate in crossing operations, along with a random population to permit continued searching in all space

[5] Among the 230 space groups, 58 space groups have less than eight symmetry elements, 63 space groups have eight elements, and the remaining 109 space groups have more than eight elements. In a true blind test prediction, space groups up to and including eight element groups need to be considered to realistically cover the search space; going beyond eight elements is constrained only by the availability of computing resources (Hahn, 2002).

groups. After evaluation of the candidate structures and sorting, the bins are sorted by the energy of the lowest structure in each bin, potentially leading to drastic reordering of bins and allowing space groups that are disfavored early in the search to take higher precedence later. A potential disadvantage of this method is that different subpopulations will not be refined at equal rates, but this is outweighed by the advantages of being able to search all space groups simultaneously,

An important point about using binned space groups is that the crossing algorithm requires some careful design. Since no predictions can be made about the refinement process, it cannot be determined if crossing two partially refined populations with different space groups will be able to produce a set of structures with enough diversity to properly sample other space groups. By this reasoning, it makes no difference whether or not structures are crossed with a refined population or a random population. So, in this algorithm, refined populations of different space groups are not crossed with each other, but are crossed with a randomly generated population at each generation. This preserves diversity in the population, and enables superior searching of the energy hypersurface, and from a practical standpoint drastically simplifies the implementation of the algorithm. This point is discussed in more detail in the next section.

Genetic Algorithm Refinements

There are three fundamental conditions that need to be fulfilled in a GA based CSP method: 1) the refinement process must converge on a solution, 2) the process cannot converge too quickly on a solution to avoid selection of a local minimum over a global minimum, and 3) the starting population must be sufficiently unbiased so that the global minimum is not excluded from the search, and the crossing operations designed so that the global minimum is accessible. Condition 1 is satisfied by using an elitism based strategy; from the previous successful predictions with MGAC-QE, it is known that an

elitism strategy will strongly favor convergence on a solution. In fact, with strong elitism a population will converge on a solution with relative ease within 30-50 generations, using a population size between 50-100 individuals (Kim et al., 2009; Lund et al., 2015). However, elitism by itself can fail conditions 2 and 3. Based on experience with MGAC1, in order to successfully obtain a true solution, up to ten independent predictions with different random seeds are required because the elitism strategy is highly sensitive to starting conditions, and because local minima may be strongly favored due to a lack of diversity in initial candidate populations.

Although there are other methods besides elitism that could be used, the fact that the method has worked before, despite the outlined shortcomings, is an indicator that elitism is a good overall strategy for CSP. Furthermore, some limited experimentation of other methods, such as the classic genetic algorithm (Goldberg and Holland, 1988), suggests that elitism remains a superior method for CSP. Consequently, the solutions presented here are modifications to the handling of structure generation, filtering, and the addition of clustering techniques that enhance the search strategy, while retaining elitism as the primary strategy used in MGAC.

Crossing methods. The use of different crossing methodologies was investigated. In MGAC, the standard method was to generate new structures at a rate of 0.5 times the population size, per generation, or 50% replacement. The parent candidates were selected by roulette wheel, using linear scaling based on the energies of the structures to establish probabilities. In MGAC1-QE, this was adjusted so that new structures were generated at a rate of 3 times the population size per generation, which yielded good results for histamine and glycine (Chapters 4 and 5). For MGAC2 the full crossing concept was investigated to determine if this was more viable in addressing conditions 2 and 3. Full crossing is defined as crossing all structures in a population with all other structures in that population. The hypothesized value of using a full crossing method over a roulette wheel is that full

crossing can potentially search more of the energy hypersurface at each generation. Because many potential solutions can be obtained, and because all structures in the population are equally represented, identification of a broad range of minima can theoretically occur.

Although it was hoped that this technique would be useful as a general crossing algorithm, it was discovered very quickly that the structure generation using full crossing suffers from two separate issues with respect to refined populations. The primary issue was completing the evaluation of volume-restricted candidates; after two or three generations, with small population sizes between 20-50, the number of solutions that pass the volume filter begins to expand quadratically, meaning that in subsequent generations, hundreds or thousands of structures could be candidates for QE evaluation. Given the computational constraints in using QE as the energy evaluator, this was deemed highly untenable given reasonable expectations for computational resources. The second issue with full crossing deals with solution convergence when highly similar structures are present in the population. Under an elitism model, when similar structures with greater than 90% similarity are crossed, they come to dominate the population because all of their offspring have similar genes and fitness evaluations, especially if those structures are relatively low energy in rankings. Although this is a problem when using the roulette wheel, in a full crossing this effect is much more acute because every structure is crossed with every structure, thus guaranteeing the production of multiple similar offspring in a single generation. The end effect result of this, especially if a low number of crossings have occurred, is that the solution converges on a local minimum dominated by the duplicate structures. Because of these issues, full crossing is not viable for a general crossing algorithm under elitism.

On the other hand, because so many potential solutions were found as a result of the quadratic expansion, it suggests that the full crossing does meet the hypothesis of a

providing a broader search. So, although impractical for convergence under elitism because of the computational resources required, the full crossing method has utility, particularly in the formation of initial and otherwise nonrefined populations. This is especially true for the multiple space group schema, where the initial discovery of structures is complicated by the addition of the space group schema elements. For the general crossing method, however, a reversion to the roulette wheel was determined to be most practical, using the same protocol as in MGAC1-QE.

Mutation versus migration. Typically, mutation is used to allow populations to escape stagnation, that is, local minima, potentially converging on a superior minimum, and broadening the search space. However, mutation is generally inefficient unless the mutation rate is high enough, and it is unclear how often a mutation operation will result in a structure that passes the volume filter. An alternative to using mutation is the addition of random new structures, which is equivalent to migration in terms of genetic algorithm theory. In fact, this was implemented in MGAC1: if an insufficient number of structures generated from crossing were able to pass the volume filter, new structures were generated until enough structures passed the filter to complete the population. This process is an iterative brute force process, which is inherently inefficient, and does nothing to incorporate refined structures into the random generation process. Furthermore, with the incorporation of the multiple space group schema, wasted effort from random mutation and brute force searches becomes much costlier, because random space group transitions will almost certainly result in a structure that cannot pass the volume filter.

In MGAC2, the appropriate strategy to add diversity to the population is through generating a random population at each generation step. This random population is individually crossed with the space group sorted populations, and with itself, as illustrated in Figure 6.2. Importantly, the crossing method used in the random/random crossing can be a full crossing, whereas for the other crossings, a standard roulette wheel should be

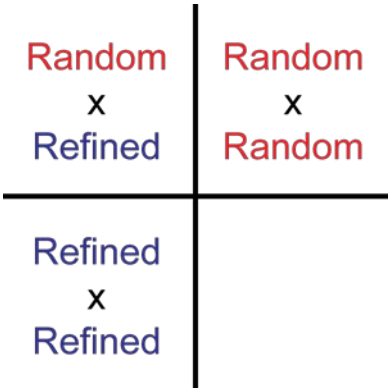


Figure 6.2: The crossing diagram of MGAC2. Three crossing are performed: random with random, refined with random, and refined with refined, once per space group. This strikes a good balance between convergence and population diversity.

used, treating the random population as having a uniform probability distribution. This hybridization of crossings fulfills all three conditions for good crossing; strong convergence can occur because of the underlying elitism, while the constant introduction of unrefined genes via migration prevents early convergence and broadens the sampling of the energy hypersurface at all generations.

Initial population construction and clustering. A final and important point to be considered in the algorithm is the generation of the initial population. Since GAs are highly sensitive to starting conditions, some precaution to generate a genetically diverse starting population is desirable. Therefore, some means of clustering and structure comparison is necessary. In the context of MGAC, a cluster is a set of closely related structures, where structure similarity is determined by application of a distance function. A cluster can then be represented by the lowest energy structure in that set, or by a singular structure that captures the essential properties. However, because MGAC deals with a high complexity problem, extreme care must be taken to choose an appropriate clustering algorithm. Since this is a genetic algorithm, an effective means of comparing structures is through measuring the genetic distance between structures. Essentially, when two structures are compared, if they share significant number of genes, then they are considered part of the same cluster. However, since most of the GA parameters are actual physical properties, the concept of similarity needs to be addressed in terms of distance thresholds. A step is defined as a normalized difference between two parameters, where each parameter is defined by the physical characteristic of that parameter. For example, when considering cell angles, the primary measure is degrees, so the step size might be 10 degrees. Cell positon is expressed in fractional coordinates, so the step size might be a unit less fractional difference of 0.1. Of each of the parameters, the only one with special consideration is the molecular rotation. Since there is no local reference frame for molecule rotation, the difference between rotation matrices is measured by calculating the

axis-angle rotations of each matrix relative to an arbitrary vector, and then measuring the angle difference of both the axis and angle components, taking care to respect singularities and axis inversions. Once the parameters have been normalized to generic “steps,” the distance between structures can be measured.

Several methods of measuring distances were explored; for a more comprehensive discussion see (Cha and Srihari, 2002). In the interest of simplicity, the standard Euclidean, Taxi-Cab, and Chebyshev (or max metric) distances were explored for use in the MGAC2 algorithm, for building the initial population, and potentially for intermediate structure generation. After some investigation, a hybrid method that uses the Chebyshev distance (which only uses the maximum parameter difference) and the average of the remaining parameter differences was selected for the distance measure. The justification for this choice is that many structures are highly similar in terms of average genetic distance, but often there is a single parameter that acts as an outlier, which has a large effect on the genetic distance function, but has minimal effects when actually comparing the physical structure. This method was also favored over the Euclidean distance, because many assertions about this distance measure break down due to the high dimensionality. The clustering method used with this distance measure was a hard cutoff method, where structures are only considered as being in the same cluster if they are within a maximum distance of the first structure in the cluster, as illustrated in Figure 6.3. This clustering method was tested by generating a set of approximately 20,000 structures using a small clustering cutoff, and then reclustering the set of structures using higher cutoffs to see how sensitive the clustering method was. In all cases, the fitcell algorithm was used to limit structures to only those that pass the volume filter, meaning that the subset of valid structures of the full search space was being explored. Figure 6.4 shows a plot of the number of clusters as the cluster cutoff was increased from 1 to 3 steps, as well as the number of interconnected regions, or cluster groups. The cluster group is defined as a set

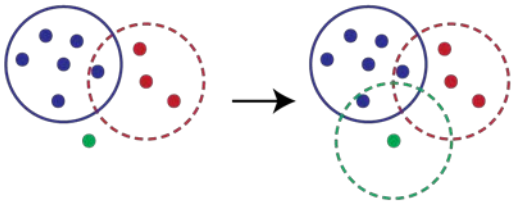


Figure 6.3: The inferior clustering method used in the initial implementation, represented in two dimensions. The blue cluster was the first to form, while the red cluster was second. The structure contained by both the blue and red cluster is considered part of the blue cluster because comparison occurs with the blue cluster first in the algorithm. A hypothetical new structure (green dot) is outside of both cluster, so a new cluster is formed. All three clusters form a group because they are close enough together to overlap. In multiple dimensions, this is problematic because connectivity between clusters can be established through multiple dimensions.

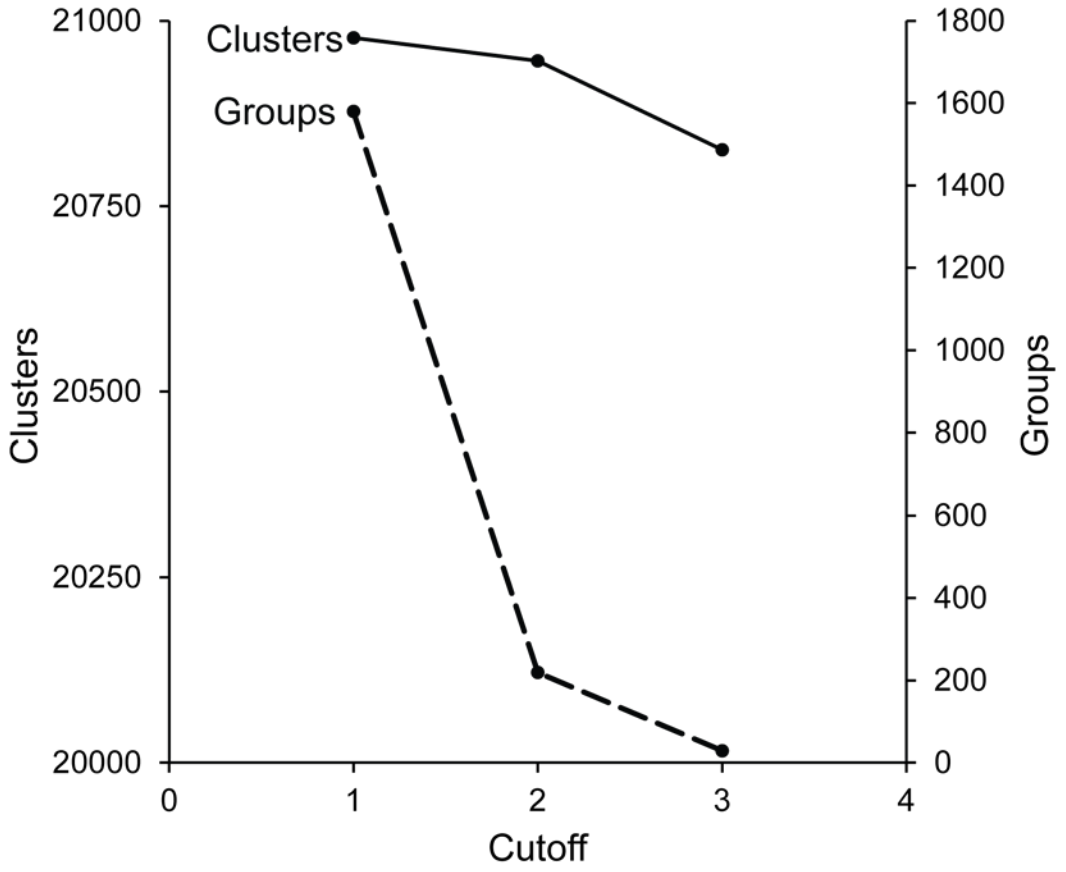


Figure 6.4: Effects of raising cluster cutoffs. A set of approximately 20,000 clusters was generated with a low cutoff of one step. The cutoff was increased to determine if a higher cutoff would produce fewer structures, but the decrease in cluster count was quite low. The number of interconnected cluster groups decreased significantly by using a higher cutoff, but because of the impossibility of evaluating more than 20,000 structures to sort based on cluster groups this method of clustering was abandoned.

of clusters where every cluster in the group touches another cluster, and all clusters can be related to the others through that adjacency. Tripling the cutoff distance did not reduce the number of clusters significantly, implying that the cluster algorithm is not very sensitive to this parameter. Furthermore, because the number of groups decreased substantially, it implies that the set of volume limited structures are highly interconnected through the distance function as a consequence of high dimensionality, and possibly that all volume limited structures are related to each other. It also clearly demonstrates that there is a very large set of possible structures, which although expected, presents the difficulty of too many structures to evaluate.

A different approach to the clustering algorithm was also explored. Based on the previous results, using classic distance measurements effectively fail because of the high dimensionality of the system. Instead, for each parameter in the genome, the distance between the two structures for that parameter is measured and normalized as before; if the distance between those two parameters is below a predetermined threshold, then the structure is considered identical for that parameter. The genetic distance is then defined as the number of identical parameters divided by the total number of shared parameters; a structure with 30% similarity to another structure will have one third of the parameters structure is considered identical for that parameter. The genetic distance is then defined as the number of identical parameters divided by the total number of shared parameters; a structure with 30% similarity to another structure will have one third of the parameters within each respective cutoff. Figure 6.5 illustrates the genetic measure using alphanumeric strings. For the purposes of the MGAC2 algorithm, a structure is considered as being part of the same cluster if it shares more than 30% similarity with another structure; if a structure is part of an existing cluster it is rejected, thus guaranteeing that no two clusters have more than 30% similarity with each other. This essentially allows for the creation of a wide diversity of structures that permits wide sampling of the energy

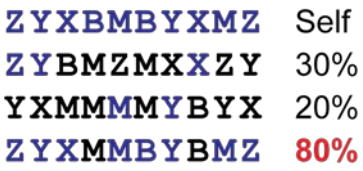


Figure 6.5: Measuring the genetic distances between arbitrary strings. The first string is compared against the remaining strings; if the letter in one position is the same in both strings, then that contributes to the similarity score.

hypersurface by maximizing the genomic space sampled in the initial population. A test of this algorithm was performed using the same framework as the previous clustering test, but using an iterative approach where at each stage new structures were generated and clustered. Figure 6.6 shows a plot of the number generations against the number of clusters, performed in several different space groups using this final clustering algorithm; the data show that this clustering method generates a low number of representative structures much more effectively, and that increasing iterations of the clustering generations is effectively bounded by the existing clusters. Consequently, this clustering algorithm is suitable for use in the MGAC2 algorithm for generating the initial population.

Computational Changes

One of the underlying issues with MGAC1 was the parallelization scheme. MGAC1 was never designed to be used with long-running energy calculation software like QE, and it was targeted at systems with low numbers CPU cores, in a fully parallel distributed system using MPI. MGAC2 has been redesigned to make use of modern computing hardware, which has become highly multicore and substantially more suitable to parallelism, in order to properly distribute a number of tasks in parallel that could be completed much more efficiently than MGAC1 is capable of. In addition, the adoption of the multiple space group schema also prompted some changes to how QE evaluations are distributed across computing resources. In MGAC1-QE, since all prediction experiments were performed in single space groups, the distribution of resources was straightforward, because compute times could be expected to be relatively similar for any individual in the population. In MGAC2, because multiple space groups of different sizes will be evaluated in the same generation, the scheduling of resources becomes greatly complicated. In addition to this, the restart mechanism of the software needed some changes to accommodate long running QE jobs.

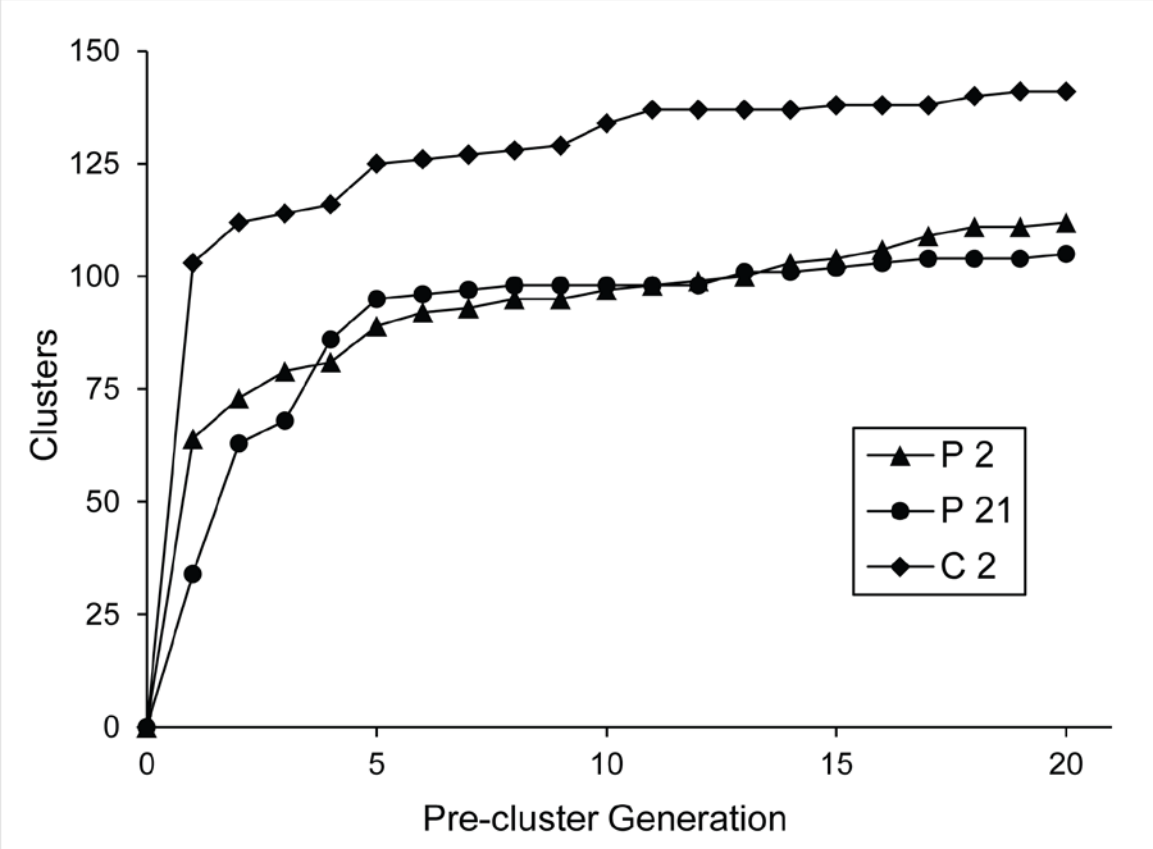


Figure 6.6: The number of structures generated using the improved measurement clustering for an arbitrary sampling of space groups. The number of clusters increases very quickly in the first few generations, but reaches an inflection quickly as the genome sampling becomes saturated.

Figure 6.7 shows the primary algorithmic changes in MGAC2 to improve parallelism and robustness. In MGAC1, the fitcell algorithm, which as discussed in Chapter 3 is used to minimize the volume of the unit cell, was performed in a coupled fashion with the energy evaluation; fitcell would be applied to a structure, and then immediately evaluated if it passed the volume filter. This operation was parallel distributed across each candidate population, resulting in severe inefficiencies due to the handling of MGAC1 processes.[6] In MGAC2, the fitcell and energy evaluation steps need to be decoupled to make better use of parallelism, especially when tens of thousands of structures need to be evaluated in forming the initial generation. As shown in Figure 6.7, fitcell is now performed independently from the evaluation step, parallel distributed across nodes and threads. This allows for much greater efficiency in generating structures and allows for more decision making to be made about the quality of structures ahead of time, which is important for scheduling work in the evaluation step. As mentioned before, Quantum Espresso computation time scales quadratically with system size. Since the multiple space group method is a significant part of this design, structures can now take substantially different amounts of time to complete. In order to partially normalize compute times, QE needs to be distributed across different numbers of cores based on the number of space group operations. In MGAC2, this is implemented linearly, so that the number of cores is proportional to the number of symmetry operations. Ideally quadratic scaling should be used, but for some space groups this becomes prohibitive because sufficient compute resources are typically not available for such a high computation cost. In addition to that, QE suffers from scaling to higher numbers of cores because of increased communication overhead, putting an effective technical limit to the degree to which QE can be parallelized.

[6] In MGAC1 active processes were scheduled in a “one process per core” basis, with multiple processes per compute node; in MGAC1-QE, this was changed to “one process per compute node”, with each node typically having sixteen cores. This means that during the fitcell process, only 1 out of 16 cores was being used; because in some instances fitcell was being run repeatedly for a long time (hours) this resulted in some glaring inefficiencies.

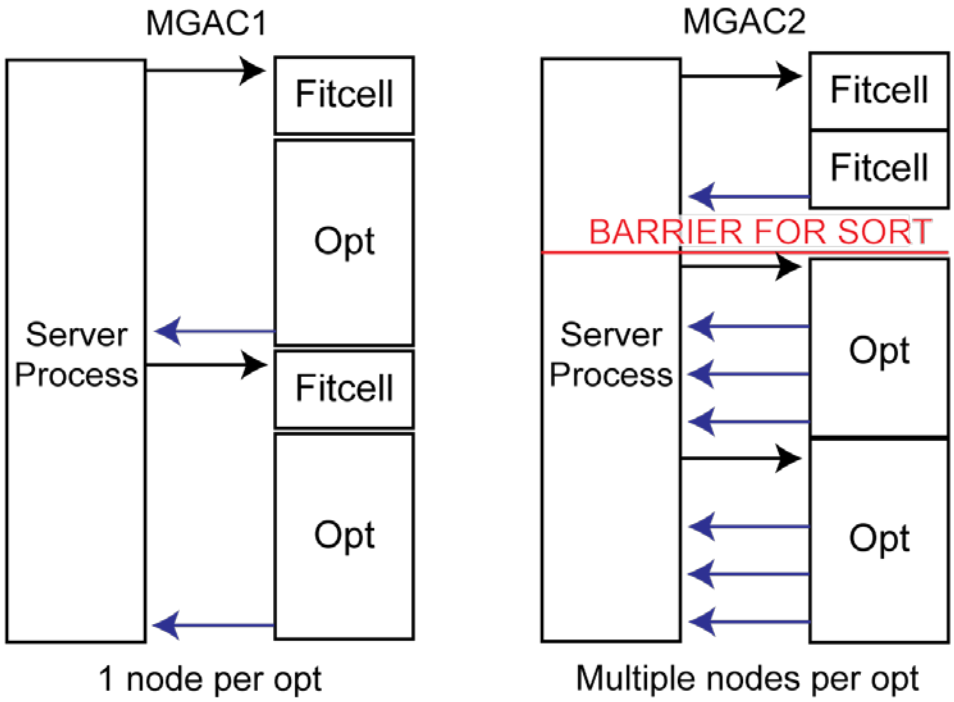


Figure 6.7: Technical changes the parallel distribution of fitcell and optimizations. The server process is responsible for all high level genetic algorithm functions, whereas clients are responsible for fitcell and optimization functions. In each diagram, arrows represent information transactions between nodes. In MGAC1, fitcell and optimization/evaluations were performed in sequence on each worker node. In MGAC2, each worker node can perform multiple fitcells in parallel, taking advantage of the now common multicore architecture found in high performance computing. Once all fitcell operations have been performed, the volume minimized structures can be sorted and potentially evaluated. During the optimization/evaluation phase, intermediate structure information can be sent back to the server process to be saved, so that if the prediction run terminates early, the prediction can be restarted with minimal loss of work.

Importantly, prior to evaluation but after fitcell, the candidate structures are sorted in descending order of number of symmetry elements; this makes the scheduling of structure evaluations much simpler and more predictable. Incorporated with this change is the change to the restart mechanism. The MGAC2 algorithm is designed to permit more communication during QE evaluations since optimizations can take a long time, allowing partial optimizations to be saved on each increment of the QE optimization.

Steady-State Algorithm

One additional algorithmic change for MGAC2 is the implementation of a steady-state genetic algorithm as an alternative operating mode, complimentary to the step-wise mode described above. A steady-state GA removes the concept of discrete generations from the GA, and instead continually evolves a population until convergence (i.e., stagnation) is achieved. There is precedent for the use of the steady-state GA in similar problems in solid-state materials research (Bhattacharya et al., 2013, 2015; Scheffler, 2014), so the option of using this method is a potentially important innovation for MGAC2. Furthermore, a steady-state implementation of MGAC2 would solve the resource scheduling problem presented by the multiple space group schema, by allowing for more variability in the compute times (although the implementation of scheduling in the step-wise method would remain in the algorithm design).

Figure 6.8 shows a flowchart of the proposed steady-state GA. In this model, the processes of structure generation and optimization are decoupled in a desynchronized way. The population size is also allowed to be variable in size, instead maintaining two lists of structures, for optimized and nonoptimized structures. Each of the two halves of the processing algorithm then work independently, with the optimization workers carrying most of the heavy workload, while the light worker (left in the figure) handles structure management. The removal of the step-wise component primarily solves part of the

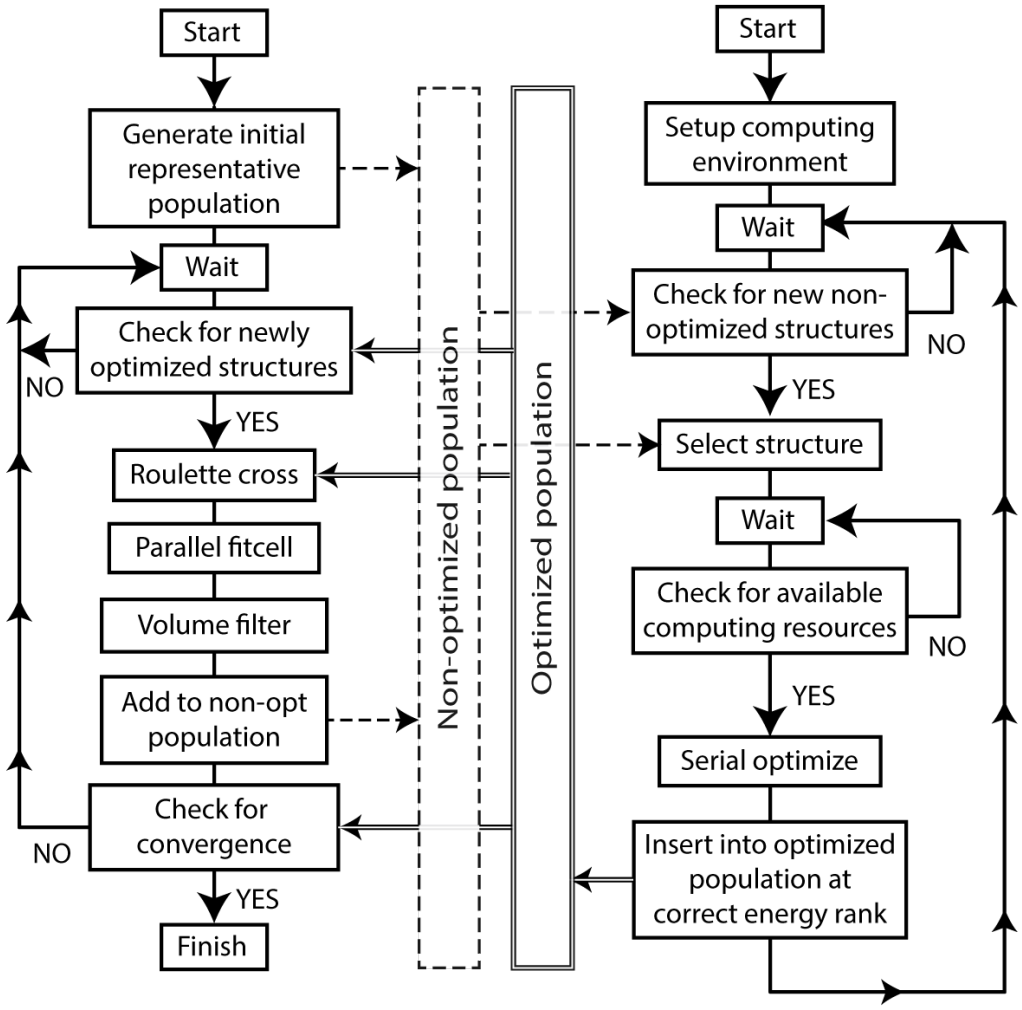


Figure 6.8: A flowchart for a steady-state GA implementation. In the middle are two populations, a nonoptimized and an optimized structure list. These are collections of all structures to be potentially evaluated, and those that have already been evaluated. On the left side is the structure management workflow, which handles the creation, fitcelling, and convergence checking for the algorithm. On the right is the workflow which handles the optimization, evaluation, and ranking of the structures in the optimized population. The important feature is that this algorithm does not operate in a step-wise fashion, but that both workflows continuously integrate new solutions into both population lists, guaranteeing a constant flow of work to keep the optimization queues full.

scheduling issue, where at the end of each generation, there are unused resources as optimizations finish at different times and no new work is available to fill the queue. In a steady state mode this means that the unpredictability of the structure optimization is softened substantially, allowing for new work to be constantly generated and performed continuously. A disadvantage to this method is that it somewhat complicates the structure generation; because there is a large variable population, the selection of structures becomes problematic because there are so many structures to choose from, it becomes more difficult to select structures that meaningfully contribute to the diversity or convergence of the population. One way to get around this is to adjust the weighting strategy used in the roulette wheel, to balance out poorly ranked structures with highly ranked structures in such a way that progress can be made towards convergence, without having to rely on the removal of structures.

Conclusion

The innovations proposed here greatly enhance the CSP capabilities of MGAC. In particular, the advancement of the multiple space group schema overcomes one of the largest obstacles in CSP, which is the proper sampling of all space groups. It is expected that this will allow for truly blind tests to be performed using MGAC2, which will be a major step forward to solving the problem of CSP. A full implementation of MGAC2 will hopefully follow soon after the publication of this dissertation. The modernization of MGAC will greatly simplify the ability to perform experiments and allow for even more scientific improvements to be made, because the tools to perform solid CSP will be available. Some future problems to be addressed with MGAC2 are the improvement of the fitcell algorithm to include multiple molecular systems. As shown in the recent sixth blind test by the CCDC, co-crystals are an important area of research, and particularly important to the formulation of many compounds. The improved fitcell discussed in Chapter 3

discusses some potential ideas to handle this, but additional features and research will need to be performed to implement variable stoichiometries and co-crystal formulations in the MGAC algorithm, which represents the next major frontier for CSP. However, it is expected that MGAC2 will make great strides towards solving the problem of CSP in the near future for single-molecule systems. In the next chapter some results using an implementation of MGAC2 will be described.

CHAPTER 7

DISSEMINATION OF HISTAMINE RESULTS USING THE MGAC2 ALGORITHM

In the previous chapter, the MGAC2 algorithm was described. Here we report limited experimental work performed to validate some key points of the algorithm. In this chapter, an implementation of MGAC2 is tested and results are presented on a prediction of histamine in the native space group. Some additional findings about volume control are also discussed.

Methods

A single MGAC2 prediction was performed on histamine in the native space group P21, as per the previous validation run using MGAC1-QE, (Bonnet and Ibers, 1973). The population size was set to 60, with population replacement set at 2.5 times the population size. A cluster similarity of 30% was used for the preclustering stage, with 50 pre-clustering steps. The volume constraints were set to -30% to +30% of the estimated system volume, and then expanded to -30% to +100% on further review. All Quantum Espresso parameters were set to the same values used in Lund et al., (2013). All QE optimizations and energy calculations were limited to either 70 optimization steps, or a maximum run time of one hour.

Results

In this prediction two structures were identified after 24 generations that correspond to the experimental structure. These two structures are of low quality in

comparison to the results presented in Chapter 5, in that the energies and cell parameters of the MGAC2 predicted structures do not match precisely with the experimental parameters. Table 7.1 shows the cell parameters for histamine and the MGAC2 predicted structure. Comparison of the energy values reveals a 11.69 kJ/mol energy difference, suggesting a partial optimization towards the true global minimum, or a local minimum that is similar to the global minimum. This is reflected also in the cell parameters, where differences in cell lengths up to 0.5 angstroms are observed on all three axes. When comparing the structures using Mercury, a complete match of 15/15 molecules is not obtained for these molecules, instead obtaining 7/15 with RMS=0.622, but visual inspection of the structure overlays (Figure 7.1, left) reveals that the structures are fundamentally the same barring differences in cell axis lengths. To determine if incomplete optimization was the culprit, a final optimization was performed on the best structure from the prediction run. The results reveal that the structure was indeed only partially optimized. An additional 134 QE optimization steps were required to obtain a nearly perfect match, with 15/15 molecules matching with an RMS of 0.251. These results are highly comparable to the results obtained using MGAC1-QE, and highlight the importance of completely optimizing structures.

Notably, the volume energy plot (Figure 7.2) shows a difference in behavior between MGAC1-QE and MGAC2. Compared with Figure 5.1, the distribution of structure energies in each generation (as highlighted by the coloring of the markers) in MGAC2 is much more ordered, with earlier generations having higher energies relative to later generations. This is partly due to the preclustering step, which can preclude the generation of lower energy structures because of the effects of clustering. On the other hand, because the minimization process was truncated in this experiment (due to the one hour time limits), the higher degree of ordering might be an artifact of incomplete optimization. On the other hand, because the final structures are so strongly clustered at

Table 7.1: Cell parameters for the MGAC2 histamine prediction, the fully optimized structure from that prediction (Full opt), and from the 10-generation prediction using MGAC1-QE from Chapter 5. HISTAN data is from (Bonnet and Ibers, 1973).

Structure	A (Å)	B (Å)	C (Å)	Beta (deg)	Energy (kJ/mol)	RMS
HISTAN	7.249	7.634	5.698	104.96	-167,069.34	-
MGAC2	6.894	8.097	5.160	102.14	-167,057.65	7/15, 0.622
Full opt	6.989	7.511	5.355	101.92	-167,068.05	15/15, 0.251
MGAC1-QE	7.219	7.087	5.519	103.58	-167,069.05	15/15, 0.256

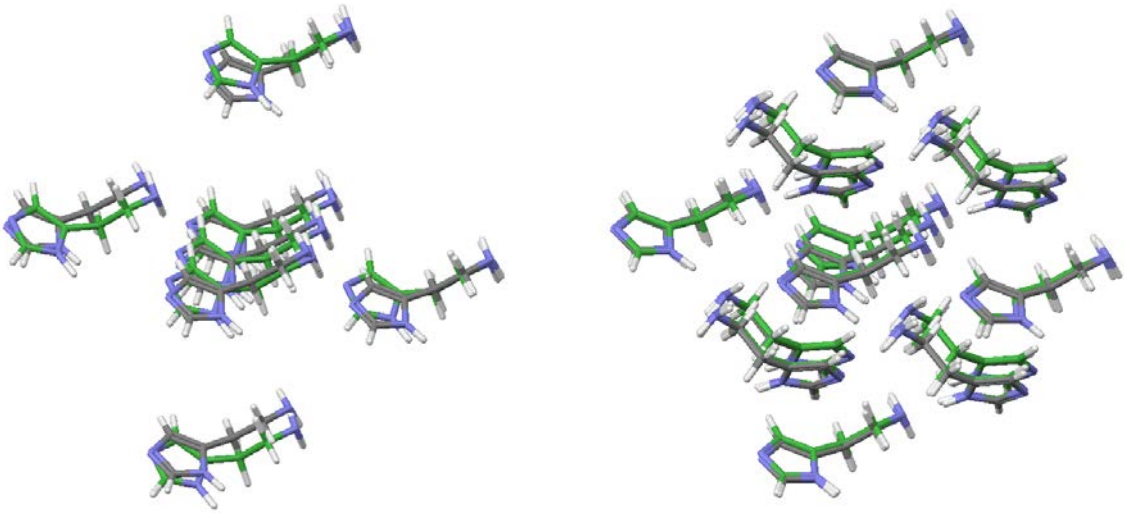


Figure 7.1: Structure overlays for the best structure obtained from the MGAC2 prediction run. Green molecules denote the predicted structure while grey indicates the reference structure HISTAN. On the left is the final structure as produced by MGAC2 (7/15, RMS=0.622), and on the right is the fully optimized version of that structure (15/15, RMS=0.251). In the left, note that the alignment is poor and that the symmetry equivalent molecules are missing from the alignment. With the fully optimized structure all symmetry elements are present, and the alignment between molecules is substantially better.

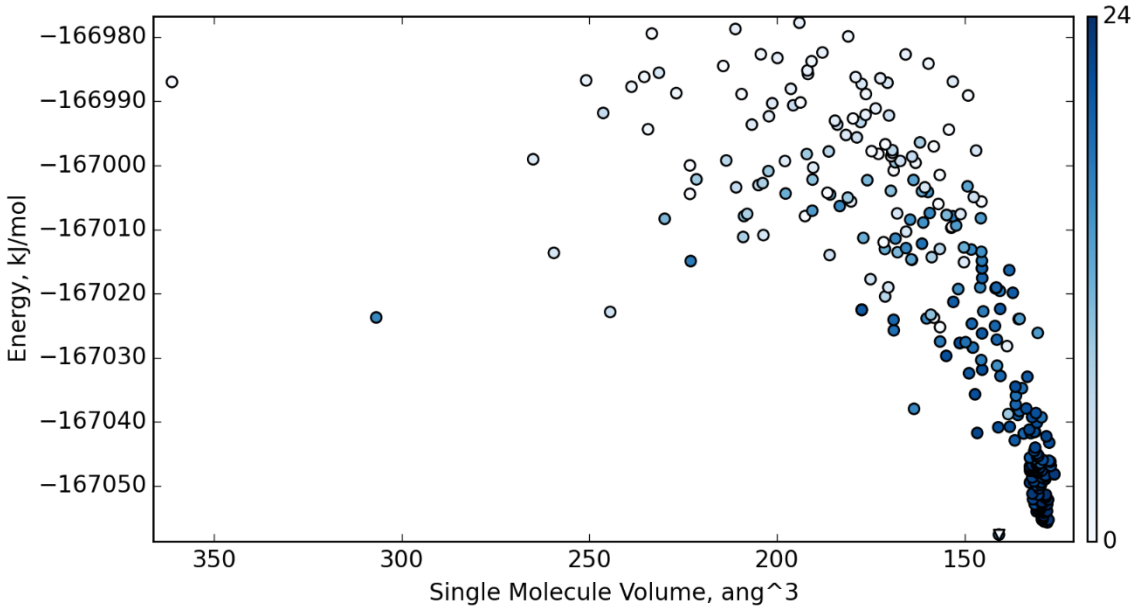


Figure 7.2: Volume-energy plot for the MGAC2 histamine run, without final optimization. The color of the marker denoting the generation as indicated by the color bar on the right. The energy funnel is well defined; note that as the number of generations increases, the funnel becomes narrower.

the bottom of the energy funnel, there is a strong implication that MGAC2 behaves more consistently with structure generation relative to MGAC1-QE, indicating that the clustering approach provides a significant improvement to prevent the loss of diversity prematurely.

As mentioned in the Methods section, the volume constraints were originally set very tightly, but then expanded to include a higher maximum range. In some earlier testing, the +/- 30% range used produced structures, but failed to converge on a solution in a reasonable manner. The fact that expanding this volume tolerance permitted the identification of an effectively correct structure highlights the importance of selecting proper volume constraints and distance parameters in the unit cell construction. However, this is at odds with the need to constrain volumes, since the volume optimization in QE is much costlier than using fitcell to optimize the volume. Therefore, more exploration of parameterization is needed to find the optimal conditions for volume control of candidate structures.

A comparison of timings follows. To complete 24 generations, this prediction took approximately 50,500 core hours to complete about 3,000 QE evaluations and 10,000 fitcell minimizations. For comparison, the 10 generation MGAC1-QE run took approximately 39,120 core hours to complete 1,845 QE evaluations and an estimated 4,000 fitcell minimizations. On average this means that MGAC1-QE evaluations took 21.5 core hours per evaluation, versus MGAC2 which took 16.8 core hours per evaluation, a 4.7 CPU hour difference. This difference could potentially be problematic, depending on the distribution of long optimizations. Therefore, in terms of efficiency of single space group tests, MGAC2 is not any more efficient than MGAC1-QE, when comparing time spent on evaluations. Any difference between the two algorithms lies strictly in the statistical likelihood that a solution is found in a set of predictions, as well as the efficiency difference in multiple space group predictions.

The data presented demonstrate the correct implementation of MGAC2 in the single space group case. This strongly implies that a full search across all space groups will work correctly, because of the bin sorting process that is implemented in MGAC2. Consequently, it is expected that in the coming future MGAC2 will be shown to be suitable for complete blind test searches using only the chemical diagram of the target molecule. This will represent a great achievement for crystal structure prediction.

REFERENCES

Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr., Sect. B Struct. Sci.* **2002**, *58*, 380–388.

Aree, T.; Bürgi, H.-B. Dynamics and Thermodynamics of Crystalline Polymorphs: α-Glycine, Analysis of Variable-Temperature Atomic Displacement Parameters. *J. Phys. Chem. A* **2012**, *116* (30), 8092–8099.

Bardwell D.A.; Adjiman C.S.; Ammon H.L.; Arnautova Y.A.; Bartashevich E.; Boerrigter S.X.M.; Braun D.E.; Cruz-Cabeza A.J.; Day G.M.; Della Valle R.G.; et al. Towards Crystal Structure Prediction of Complex Organic Molecules - a Report on the Fifth Blind Test. *Acta Cryst.* **2011**, *B67*, 535–551.

Bauer, J.; Spanton, S.; Henry, R.; Quick, J.; Dziki, W.; Porter, W.; Morris, J. Ritonavir: An Extraordinary Example of Conformational Polymorphism. *Pharm. Res.* **2001**, *18* (6), 859–866.

Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. Modified Genetic Algorithm to Model Crystal Structures. I. Benzene, Naphthalene and Anthracene. *J. Chem. Phys.* **2002a**, *116* (14), 5984–5991.

Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. Modified Genetic Algorithm to Model Crystal Structures. II. Determination of a Polymorphic Structure of Benzene Using Enthalpy Minimization. *J. Chem. Phys.* **2002b**, *116* (14), 5992–5995.

Bazterra, V. E.; Ferraro, M. B.; Facelli, J. C. Modified Genetic Algorithm to Model Crystal Structures. III. Determination of Crystal Structures Allowing Simultaneous Molecular Geometry Relaxation. *Int. J. Quantum Chem.* **2004**, *96*, 312–320.

Bazterra, V. E.; Thorley, M.; Ferraro, M. B.; Facelli, J. C. A Distributed Computing Method for Crystal Structure Prediction of Flexible Molecules: An Application to N-(2-Dimethyl-4-5-Dinitrophenyl) Acetamide. *J. Chem. Theory Comp.* **2007**, *3*, 201–209.

Beaucamp, S.; Mathieu, D.; Agafonov, V. Optimal Partitioning of Molecular Properties into Additive Contributions: The Case of Crystal Volumes. *Acta Crystallogr., Sect. B: Struct. Sci.* **2007**, *63* (2), 277–284.

Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.

Bhattacharya, S.; Levchenko, S. V; Ghiringhelli, L. M.; Scheffler, M. Stability and Metastability of Clusters in a Reactive Atmosphere: Theoretical Evidence for Unexpected Stoichiometries of MgMOx. *Phys. Rev. Lett.* **2013**, *111* (13), 135501.

Bhattacharya, S.; Sonin, B. H.; Jumonville, C. J.; Ghiringhelli, L. M.; Marom, N. Computational Design of Nanoclusters by Property-Based Genetic Algorithms: Tuning the Electronic Properties of $(TiO_2)_n$ Clusters. *Phys. Rev. B* **2015**, *91* (24), 241115.

Boldyreva, E. V.; Ivashevskaya, S. N.; Sowa, H.; Ahsbahs, H.; Weber, H.-P. Effect of Hydrostatic Pressure on The γ-Polymorph of Glycine. 1. A Polymorphic Transition into a New δ-Form. *Z. Kristallogr.* **2005**, *220*, 50–57.

Boldyreva, E. V; Ahsbahs, H.; Weber, H.-P. A Comparative Study of Pressure-Induced Lattice Strain of α-and γ-Polymorphs of Glycine. *Z. Kristallogr.* **2003**, *218* (3), 231–236.

Bondi, A. Van Der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68* (3), 441–451.

Bonnet, J. J.; Ibers, J. A. Structure of Histamine. *J. Am. Chem. Soc.* **1973**, *95* (15), 4829–4833.

Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comp. Chem.* **1983**, *4*, 187–217.

Cha, S.-H.; Srihari, S. N. On Measuring the Distance between Histograms. *Pattern Recognit.* **2002**, *35* (6), 1355–1370.

Chisholm, J. A.; Motherwell, S. COMPACK: A Program for Identifying Crystal Structure Similarity Using Distances. *J. Appl. Crystallogr.* **2005**, *38* (1), 228–231.

Crim, F. F. Presented at the NSF Mathematical and Physical Sciences Advisory Committee Meeting, April 3, 2014. https://www.nsf.gov/attachments/130168/public/MPSAC_Presentation_Crim_April_2014.pdf (accessed Jan 1, 2015).

Datta, S.; Grant, D. J. W. Crystal Structures of Drugs: Advances in Determination, Prediction and Engineering. *Nat. Rev. Drug. Discov.* **2004**, *3* (1), 42–57.

Day, G. M. Current Approaches to Predicting Molecular Organic Crystal Strucutres. *Crystallogr. Rev.* **2011**, *17* (1), 3–52.

Day, G. M. Crystal Structure Prediction. In *Supramolecular Chemistry: From Molecules to Nanomaterials*; Gale, P. A., Steed, J. W., Eds.; John Wiley & Sons: New York, 2012; pp 2905–2925.

Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; van Eijck, B. P.; et al. A Third Blind Test of Crystal Structure Prediction. *Acta Crystallogr., Sect. B: Struct. Sci.* **2005**, *61* (5), 511–527.

Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J. S.; Della Valle, R. G.; Venuti, E.; Jose, J.; et al. Significant Progress in Predicting the Crystal Structures of Small Organic Molecules - a Report on the Fourth Blind Test. *Acta Crystallogr., Sect. B: Struct. Sci.* **2009**, *65*, 107–125.

Deschamps, J. R.; Parrish, D. A.; Butcher, R. J. Polymorphism in Energetic Materials. *2008 NRL Review;* Naval Research Laboratory, Washington DC, **2008**, 71–77.

Desiraju, G. R. Crystal Gazing: Structure Prediction and Polymorphism. *Science.* **1997**, *278* (5337), 404–405.

Domingos, P. A Few Useful Things to Know about Machine Learning. *Commun. ACM* **2012**, *55* (10), 78–87.

Falkenauer, E. *Genetic Algorithms and Grouping Problems*; John Wiley & Sons, Ltd.: West Sussex, England, 1998.

Fitzpatrick, J. M.; Grefenstette, J. Genetic Algorithms in Noisy Environments. *Mach. Learn.* **1988**, *3* (2-3), 101–120.

Foadi, J.; Evans, G. On the Allowed Values for the Triclinic Unit-Cell Angles. *Acta Crystallogr., Sect. A* **2011**, *67* (1), 93–95.

Foltz, M. F.; Coon, C. L.; Garcia, F.; Nichols, A. L. The Thermal Stability of the Polymorphs of Hexanitrohexaazaisowurtzitane, Part I. *Propellants, Explos. Pyrotech.* **1994**, *19* (1), 19–25.

Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; et al. QUANTUM ESPRESSO: A Modular and Open-Source Software Project for Quantum Simulations of Materials. *J. Phys. Condens. Matter* **2009**, *21* (39), 395502.

Goldberg, D. E.; Holland, J. H. Genetic Algorithms and Machine Learning. *Mach. Learn.* **1988**, *3* (2), 95–99.

Grimme, S. Accurate Description of van Der Waals Complexes by Density Functional Theory Including Empirical Corrections. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (5), 1787–1799.

Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104–154123.

Haas, H. L.; Sergeeva, O. A.; Selbach, O. Histamine in the Nervous System. *Physiol. Rev.* **2008**, *88* (3), 1183–1241.

Holland, J. H. Genetic Algorithms and the Optimal Allocation of Trials. *SIAM J. Comput.* **1973**, *2* (2), 88–105.

Issa, N.; Karamertzanis, P. G.; Welch, G. W. A.; Price, S. L. Can the Formation of Pharmaceutical Cocrystals Be Computationally Predicted? I. Comparison of Lattice Energies. *Cryst. Growth Des.* **2009**, *9* (1), 442–453.

Jones, W.; Motherwell, W. D. S.; Trask, A. V. Pharmaceutical Cocrystals: An Emerging Approach to Physica Property Enhancement. *MRS Bull.* **2006**, 31, 875–879.

Karamertzanis C. C., P. G. . P. Ab Initio Crystal Structure Prediction - I. Rigid Molecules. *J. Comput. Chem.* **2005**, *26*, 304–324.

Karamertzanis, P. G.; Price, S. L. Energy Minimization of Crystal Structures Containing Flexible Molecules. *J. Chem. Theory Comput.* **2006**, *2* (4), 1184–1199.

Kendrick, J.; Leusen, F. J. J.; Neumann, M. A.; Van De Streek, J. Progress in Crystal Structure Prediction. *Chem. - A Eur. J.* **2011**, *17* (38), 10736–10744.

Kim, S.; Orendt, A. M.; Ferraro, M. B.; Facelli, J. C. Crystal Structure Prediction of Flexible Molecules Using Parallel Geneic Algorithms with Standard Force Field. *J. Comp. Chem.* **2009**, *30*, 1973–1985.

King, M. D.; Blanton, T. N.; Misture, S. T.; Korter, T. M. Prediction of the Unknown Crystal Structure of Creatine Using Fully Quantum Mechanical Methods. *Cryst. Growth Des.* **2011**, *11*, 5733–5740.

Larionov, L. V. Polymorphism and Initiation of Explosives. *Combust. Explos. Shock Waves* **1997**, *33* (5), 605–610.

Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B Condens. Matter Mater. Phys.* **1988**, *37* (2), 785–789.

Lehmann, C. W. Crystal Structure Prediction - Dawn of a New Era. *Angew. Chem. Int. Ed.* **2011**, *50*, 5616–5617.

Leurs, R.; Chazot, P. L.; Shenton, F. C.; Lim, H. D.; de Esch, I. J. P. Molecular and Biochemical Pharmacology of the Histamine H(4) Receptor. *Br. J. Pharmacol.* **2009**, *157* (1), 14–23.

Liberti, L. Introduction to Global Optimization. *Lect. Ec. Polytech. Palaiseau F* **2008**, *91128*, 12.

Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; et al. A Test of Crystal Structure Prediction of Small Organic Molecules. *Acta Crystallogr., Sect. B: Struct. Sci.* **2000**, *56* (4), 697–714.

Lopez, J. C. Histamine's Comeback? *Nat Rev Neurosci* **2002**, *3* (2), 84.

Lund, A. M.; Orendt, A. M.; Pagola, G. L.; Ferraro, M. B.; Facelli, J. C. Optimizatrion of Crystal Structures of Archetypical Pharmaceutical Compounds: A Plane-Wave DFT-D Study Using Quantum Espresso. *Cryst. Growth Des.* **2013**, *13*, 2181-2189.

Lund, A. M.; Pagola, G. I.; Orendt, A. M.; Ferraro, M. B.; Facelli, J. C. Crystal Structure Prediction from First Principles: The Crystal Structures of Glycine. *Chem. Phys. Lett.* **2015**, *626*, 20–24.

MacKerell, A. D.; Brooks, J. B.; Brooks III, C. L.; Nilsson, L.; Roux, B.; Won, Y.; Karplus, M. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. R., Schreiner, P. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F., Eds.; John Wiley & Sons: Chichester, U. K., 1998; pp 271-277.

Marom, N.; DiStasio, R. A.; Atalla, V.; Levchenko, S.; Reilly, A. M.; Chelikowsky, J. R.; Leiserowitz, L.; Tkatchenko, A. Many-Body Dispersion Interactions in Molecular Crystal Polymorphism. *Angew. Chemie Int. Ed.* **2013**, *52* (26), 6629–6632.

Miller, G. R.; Garroway, A. N. *A Review of the Crystal Structures of Common Explosives. Part I: RDX, HMX, TNT, PETN, and Tetryl*; Naval Research Laboratory: Washington DC, 2001.

Morissette, S. L.; Almarsson, O.; Peterson, M. L.; Remenar, J. F.; Read, M. J.; Lemmo, A. V; Ellis, S.; Cima, M. J.; Gardner, C. R. High-Throughput Crystallization: Polymorphs, Salts, Co-Crystals and Solvates of Pharmaceutical Solids. *Adv. Drug Deliv. Rev.* **2004**, *56* (3), 275–300.

Morokoff, W. J.; Caflisch, R. E. Quasi-Monte Carlo Integration. *J. Comput. Phys.* **1995**, *122* (2), 218–230.

Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; et al. Crystal Structure Prediction of Small Organic Molecules: A Second Blind Test. *Acta Cryst.* **2002**, *B58*, 647.

Nagy, Z. K.; Fujiwara, M.; Braatz, R. D. Modelling and Control of Combined Cooling and Antisolvent Crystallization Processes. *J. Process Control* **2008**, *18* (9), 856–864.

Neumann, M. A. Crystal Structures of Moderately Complex Organic Molecules Are Predictable. *24th European Christallographic Meeting, Micro Symposium 14, Advanced computational methods in structural chemistry*. Marrakech, Morocco 2007, pp 11H00–H11H20.

Neumann, M. A. Tailor-Made Force Fields for Crystal-Structure Prediction. *J. Phys. Chem. B* **2008**, *112* (32), 9810–9829.

Neumann, M. A.; Perrin, M. A. Energy Ranking of Molecular Crystals Using Density Functional Theory Calculations and an Empirical van Der Waals Correction. *J. Phys. Chem. B* **2005**, *109*, 15531–15541.

Neumann, M. A.; Leusen, F. J. J.; Kendrick, J. A Major Advance in Crystal Structure Prediction. *Angew. Chem. Int. Ed.* **2008**, *47*, 2427–2430.

Niederreiter, H. Low-Discrepancy and Low-Dispersion Sequences. *J. Number Theor.* **1988**, *30* (1), 51–70.

Perdew, J. P.; Burke, K.; Wang, Y. Generalized Gradient Approximation for the Exchange-Correlation Hole of a Many-Electron System. *Phys. Rev. B Condens. Matter* **1996a**, *54* (23), 16533–16539.

Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996b**, *77* (18), 3865–3868.

Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for Mixing Exact Exchange with Density Functional Approximations. *J. Chem. Phys.* **1996**, *105* (22), 9982–9985.

Price, S. L. The Computational Prediction of Pharmaceutical Crystal Structures and Polymorphism. *Adv. Drug Deliv. Rev.* **2004**, *56*, 301–319.

Price, S. L. Computed Crystal Energy Landscapes for Understanding and Predicting Organic Crystal Structures and Polymorphism. *Acc. Chem. Res.* **2009**, *42* (1), 117–126.

Price, S. L. Why Don't We Find More Polymorphs? *Acta Crystallogr. B. Struct. Sci. Cryst. Eng. Mater.* **2013**, *69* (Pt 4), 313–328.

Price, S. L.; Price, L. S. Computational Polymorph Predicion. *Solid State Charact. Pharm.* **2011**, 427–450.

Rappe, A. M.; Rabe, K. M.; Kaxiras, E.; Joannopoulos, J. D. Optimized Pseudopotentials. *Phys. Rev. B* **1990**, *41* (2), 1227-1230.

Rowland R.; Taylor, R. Intermolecular Nonbonded Contact Distances in Organic-Crystal Structures - Comparison With Distances Expected From Van-Der-Waals Radii. *J. Phys. Chem* **1996**, *100*, 7384–7391.

Scheffler, M.; Ghiringhelli, L.; Levchenko S.; Bhattacharya S. Efficient Ab Initio Schemes for Finding Thermodynamically Stable and Metastable Atomic Structures: Benchmark of Cascade Genetic Algorithms. *New J. Phys.* **2014**, *16* (12), 123016.

Sobol, I. M. On Quasi-Monte Carlo Integrations. *Math. Comput. Simul.* **1998**, *47* (2), 103–112.

van de Streek, J.; Neumann, M. A. Validation of Experimental Molecular Crystal Structures with Dispersion-Corrected Density Functional Theory Calculations. *Acta Cryst.* **2010**, *B66*, 544–558.

Tinkham, M. *Group Theory and Quantum Mechanics*; Courier Corporation, 2003.

Tumanov, N. A.; Boldyreva, E. V; Ahsbahs, H. Structure Solution and Refinement from Powder or Single-Crystal Diffraction Data? Pros and Cons: An Example of the High-Pressure β′-Polymorph of Glycine. *Powder Diffr.* **2008**, *23* (04), 307–316.

Vanderbilt, D. Soft Self-Consistent Pseudopotentials in a Generalized Eigenvalue Formalism. *Phys Rev B* **1990**, *41* (11), 7892–7895.

Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247–260.

Yu, L.; Huang, J.; Jones, K. J. Measuring Free-Energy Difference between Crystal Polymorphs through Eutectic Melting. *J. Phys. Chem. B* **2005**, *109*, 19915–19922.

Zhu, Q.; Oganov, A. R.; Glass, C. W.; Stokes, H. T. Constrained Evolutionary Algorithm for Strucutre Prediction of Molecular Crystals: Methodology and Applications. *Acta Cryst.* **2012**, *B68*, 215–226.

Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's Paradox. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (1), 20–22.

International Tables for Crystallography, Volume A: Space-Group Symmetry, Fifth.; Hahn, T., Ed.; Kluwer Academic Publishers for the International Union of Crystallography: Dordrecht/Boston/London, 2002.