ISTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

DEVELOPMENT OF A NOVEL EVOLUTIONARY ALGORİTHM SPECIALIZED FOR CRYSTAL STRUCTURE PREDICTION OF MOLECULAR SYSTEMS: MCaSP-Evo

M.Sc. THESIS

Denizhan TUTAR

Department of Computational Science and Engineering

Computational Science and Engineering Programme

DECEMBER 2019

ISTANBUL TECHNICAL UNIVERSITY ★ INFORMATICS INSTITUTE

DEVELOPMENT OF A NOVEL EVOLUTIONARY ALGORİTHM SPECIALIZED FOR CRYSTAL STRUCTURE PREDICTION OF MOLECULAR SYSTEMS: MCaSP-Evo

M.Sc. THESIS

Denizhan TUTAR
(702161003)

Department of Computational Science and Engineering

Computational Science and Engineering Programme

Thesis Advisor: Prof. Dr. Adem TEKİN

DECEMBER 2019

Denizhan Tutar, a **M.Sc.** student of ITU Informatics Institute student ID **700261003**, successfully defended the **thesis** entitled " **DEVELOPMENT OF A NOVEL EVOLUTIONARY ALGORİTHM SPECIALIZED FOR CRYSTAL STRUCTURE PREDICTION OF MOLECULAR SYSTEMS: MCaSP-Evo**", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Prof. Dr. Adem TEKİN** ....
İstanbul Technical University

Jury Members : **Prof. Dr. F. Aylin SUNGUR** ....
Istanbul Technical University

Doç. Dr. H. Hakan GÜREL ....
Kocaeli University

Date of Submission : 15 November 2019
Date of Defense: 13 December 2019

To my family,

FOREWORD

The global optimization became one of the most essential methods to achieve satisfactory solutions in many challenging real world problems, and evolutionary algorithms have a special place among global optimization methods. The success of an implementation of those methods often depends on the appropriate combination reliable algorithmic setup with domain specific insight. The study presented in this thesis is an implementation of evolutionary algorithms for crystal structure prediction using preexisting codes generated by our research group for the sections other than optimizer, and intuitions about the field are considered in the selection and tuning phase of algorithmic components and design of problem specific operators in order to set a starting point for the development of a cutting-edge tool in this field.

The resulting implementation achieved those goals to a large extent, but the necessity of further improvements is clearly seen in order to tackle a considerable part of the problems targeted by the researchers in the crystal structure prediction field. This thesis includes the implementation steps and underlying intuitions, also aimed to point out the potential directions of progress in this thesis by summarizing the related literature and presenting the implementation.

The current level of success would not be achieved without the guidance of my advisor Prof. Dr. Adem Tekin, the support of whole research group and especially previous works of Samet Demir that this study partly relied on, and without the support of my family and friends that provided the most precious motivation source.

December 2019 Denizhan TUTAR

TABLE OF CONTENTS

ABBREVIATIONS

CASPESA	**:** Crystal Structure Prediction via Simualted Annealing
CMA-ES	**:** Covariance Matrix Adaptation – Evolution Strategies
CSP	**:** Crystal Structure Prediction
DFT	**:** Density Functional Theory
DFT-SAPT	**:** DFT-Symmetry-Adapted Perturbation Theory
DNA	**:** Deoxyribonucleic acid
EA	**:** Evolutionary Algorithm
EC	**:** Evolutionary Computation
EDA	**:** Estimation of Distribution Algorithms
ES	**:** Evolutionary Strategies
FFCASP	**:** Fast & Flexible Crystal Structure Prediction
GA	**:** Genetic Algorithms
GGA	**:** Generalized Gradient Approximations
HF	**:** Hartree-Fock

LIST OF TABLES

Page

LIST OF FIGURES

DEVELOPMENT OF A NOVEL EVOLUTIONARY ALGORİTHM SPECIALIZED FOR CRYSTAL STRUCTURE PREDICTION OF MOLECULAR SYSTEMS: MCaSP-Evo

SUMMARY

Crystal structure of a solid material determines most of the properties of a material, including but not limited to density, mechanical properties, electrical and electronic abilities. Therefore, computational discovery of novel materials and computational examination of known materials is significantly dependent on the information about crystal structure. Observational information of crystal structure, such as transmission electron microscopy images or other high-powered microscopy outcomes, is essential for this purpose; but it is not reliable or totally absent for many substances of interest, such as new material candidates or formations beyond our observation abilities. Hence, prediction of crystal structure serves as an indispensable tool for many research areas as computational materials design or research activities to gain insight about solid state physics, materials, or even biomolecular activities. The lack of successful computational prediction tools were/are one of main obstacles for scientific development of related research and development efforts.

Despite its central role and primary importance in material science, a successful and generally applicable theory leading predictions is absent; and computational crystal structure prediction (CSP) was arguable for even being possible. Until the fast development of computational methods for CSP in the first decade of our millennium, it was extensively considered that crystal structures are unforeseeable and related attempts were doubted due to computational complexity of the task and unsatisfactory performances of used computational chemistry tools. These limitations are mostly overcome, and CSP studies achieved important goals. On the other hand, the development of the field is still ongoing, and CSP applications are not yet as functional as they are expected to be in the close future.

The general definition of CSP can be given as the prediction of stable crystal structure at given conditions, whereas finding less stable or metastable structures may also be an important interest.

The general procedure of CSP includes investigation of target molecule(s) and ions, production and tests of candidate solutions and detailed examination of selected structures. The most popular approach for evaluation of candidate solutions is calculation of free energy, as it is one of the most important characteristics for stability of crystal structure; and the most common approach for production of candidate solutions is heuristic or metaheuristic routines implemented in the global optimization of free energy, or coverage methods such as random search or Sobol sequences.

Most challenges in the field emerge in special applications such as investigation of flexible molecules, large molecules, combinations of different molecules and ions, or large systems. Also, those differ according to the particle type present in the examined systems, i.e. between molecular and atomic system or organic and inorganic targets, so that the distinction among specialized CSP methods is noteworthy. First part of the

first chapter outlines history of CSP problem in the first section, which is followed by the second section that provides essential information about the definition and major challenges of CSP.

The major method for CSP remains minimization of free energy, whereas some other properties such as nucleation rate or molecular dynamics analysis may also be useful for selecting among crystal structures with similar free energies. The accuracy and reliability of energy calculation is crucial for the optimization performance, on the other hand, high costs of most successful energy calculations create a trade-off condition and make the selection of energy method dependent on the problem and resources at hand.

Total energy calculation methods are classified as *ab initio* methods, which are the most computationally expensive and reliable ones. DFT is a special case of *ab initio* methods with lower cost. Semi-empirical approximations cheaper than DFT are also available. The fastest and most affordavle approach to calculate the energy is the empirical methods such as force fields, that may exhibit sufficient accuracy when generated and used properly. These methods are summarized and explained in the second part of the first chapter.

The second chapter is dedicated to an explanatory introduction to the global optimization, by examining of its major classes with a special emphasis on nature-inspired methods, including their principal implementations for CSP problem. Other CSP procedures with different objective functions and optimization methods developed by our group are explained in the following part, followed by mentions on various artificial intelligence algorithms and other applications. Evolutionary Algorithms (EAs) are not included as the third chapter is dedicated to those.

Historical background and general structure of EAs is provided in the third chapter. Basic concepts and considerations in EA design, which guided our implementation process, are summarized and most important factors for our study are highlighted, followed by very brief mention on potential future additions to our implementation in the section of examples for popular and promising variations of EA. Among many other examples, two most promising EAs on CSP are chosen to mention in the last part, namely USPEX and GAtor.

In the last chapter, our implementation is presented by its algorithmic components and domain-specific operators, together with explanations and decisive considerations in the selection of related content. The concept of substructure discovery is especially emphasized and underlying motivations are explained, for its potential to carry our work in the future.

MOLEKÜLER SİSTEMLERİN KRİSTAL YAPI TAHMİNİ İÇİN ÖZELLEŞTİRİLMİŞ BİR EVRİMSEL ALGORİTMANIN GELİŞTİRİLMESİ: MCaSP-Evo

ÖZET

Katı bir kristal malzemenin kristal yapısı malzemeyle ilgili özelliklerin çoğunu belirler; bu özelliklere örnek olarak yoğunluk, mekanik özellikler, elektrik ve elektronik kabiliyetleri verilebilir ancak bununla sınırlı değildir. Bu nedenle bilinen malzemelerin hesapsal incelenmesi ve yeni malzemelerin bilgisayımsal keşfi önemli ölçüde kristal yapı hakkındaki bilgiye bağımlıdır. Kristal yapı hakkında güçlü mikroskoplarla elde edilen görüntüler gibi gözlemsel bilgiler bu konuda esastır; ancak yeni malzeme adayları ve gözlem kabiliyetlerimizin ötesindeki oluşumlar gibi ilgilenilen çok sayıda madde için ya tamamen ulaşılamaz durumdadır ya da güvenilir değildir. Dolayısıyla da kristal yapı tahmini, hesapsal malzeme tasarımı veya katı hal fiziği, malzeme bilimi ve hatta biyomoleküler aktivitelerin araştırmaları için olmazsa olmaz bir araç konumundadır. Başarılı bilgisayımsal tahmin yöntemlerinin eksikliği de ilgili araştırma ve geliştirme faaliyetlerinin önündeki başlıca engellerden biri olmuştur.

Malzeme bilimindeki merkezi rolüne ve büyük önemine rağmen tahminlere yön veren başarılı ve genel olarak uygulanabilir bir teori ortaya çıkamamıştır, ve bilgisayımsal kristal yapı tahmininin (KYT) mümkün olup olmadığı bile yakın zamanlara dek tartışmalı olmuştur. Bilgisayımsal KYT yöntemlerinde bu binyılın ilk onyılındaki hızlı gelişmeye kadar kristal yapının geniş ölçüde öngörülemez olduğu düşünülürdü, ve görevin hesaplama karmaşıklığı ile kullanılan hesaplamalı kimya yöntemlerinin tatmin edici olmayan başarımları dolayısıyla ilgili çabalara şüphe ile yaklaşılırdı. Bu sınırlamalar büyük ölçüde aşıldı ve KYT çalışmaları önemli hedeflere ulaştı. Öte yandan bu alanın gelişimi hala devam etmektedir ve KYT uygulamaları henüz yakın gelecekte ulaşmaları beklenen işlevselliğe ulaşabilmiş değildir.

KYT'nin genel tanımı, verili koşullar altında kararlı olan kristal yapıların tahmini olarak verilebilir; ancak daha az kararlı yapıların ve yarıkararlı yapıların bulunması da öenmli faydalar sağlamaktadır.

KYT'nin genel prosedürü hedef molekül(ler) ve iyon(lar)ın tetkikini, aday çözümlerin üretilmesini ve değerlendirilmesini ve seçilen yapılan ayrıntılı incelenmesini içerir. Aday çözümlerin incelenmesi için en gözde yaklaşım serbest enerjilerinin hesaplanmasıdır, zira bu kristal yapının kararlılığı için en önemli özelliklerden birisidir; aday çözümlerin üretilmesi için en yaygın yaklaşım ise küresel eniyileme yöntemlerinde uygulanmış olan sezgisel ve üstsezgisel rutinler veya rastgele arama ile Sobol dizileri (Sobol tarafından geliştirilmiş olan ve arama uzayının farklı bölümlerini aynı veya yakın sıklıkla örneklemeyi hedefleyen bir sayı dizisi üretim yöntemi) gibi kaplam yöntemleri olarak verilebilir.

Bu alandaki çoğu zorluk esnek moleküllerin, büyük moleküllerin, farklı moleküllerin ve iyonların kombinasyonlarını içeren veya büyük sistemlerin incelenmesi gibi özel uygulamalarda ortaya çıkmaktadır. Bu zorluklar aynı zamanda incelenen sistemlerde

var olan parçacıkların tiplerine göre, yani moleküler ve atomik sistemler veya organik ve inorganik hedefler arasında değişkenlik göstermektedir, dolayısıyla özelleşmiş KYT yöntemleri arasındaki farklılaşma dikkate değerdir. Bu çalışmanın birinci bölümünün ilk kısmında KYT probleminin tarihinden ana hatlarıyla bahsedildi, bunu KYT'nin tanımı ve başlıca zorluklarıyla ilgili temel bilgileri sağlayan ikinci kısım takip etmektedir.

KYT için en önemli yöntem toplam enerjinin minimizasyonu olarak kalmaktadır, ancak çekirdeklenme hızı veya moleküler dinamik analizi gibi başka özellikler de benzer serbest enerji düzeylerindeki kristal yapıların arasından seçim yapmak için işlevsel olabilmektedir. Enerji hesabının isabetliliği ve güvenilirliği eniylemenin performansı açısından can alıcı öneme sahiptir, ancak en başarılı enerji hesaplama yöntemlerinin yüksek maliyetleri bir ödünleşim durumu yaratmaktadır ve enerji hesaplama yönteminin seçimini probleme ve eldeki kaynaklara bağımlı hale getirmektedir.

Serbest enerji hesaplama yöntemlerinin sınıflandırılması şu şekildedir: *ab inito* yöntemler, ki bunlar hesapsal maliyetleri en yüksek ve en güvenilir olan hiçbir deneysel veri kullanmayan yöntemlerdir; DFT yöntemleri ki bunlar da *ab initio* yöntemlerin daha düşük maliyetli bir çeşididir; *ab initio* yöntemlere yarıdeneyimsel yaklaşımlar; ve son olarak en uygun maliyetlileri olup doğru şekilde üretildiğinde ve kullanıldığında yeterli isabetliliği elde etmeye imkan veren kuvvet alanları gibi deneyimsel yaklaşımlar. Bu yöntemler birinci bölümün ikinci kısmında özetlenmiş ve açıklanmıştır.

İkinci bölüm küresel eniyilemeye açıklayıcı bir girişe ve küresel eniyileme yöntemlerinin başlıca sınıflarının doğa esinli yöntemlere önemli KYT uygulamalarını da içeren özel bir yer vererek incelenmesine ayrılmıştır. Grubumuzda geliştirilmiş olup bu çalışmadan farklı amaç fonksiyonları veya eniyileme yöntemleri kullanan başka algoritmaların incelenmesi bunlardan sonraki kısımda gelmektedir, bunu da yapay zeka ve daha başka uygulamalara değinilen son kısım takip etmektedir. Evrimsel Algoritmalar (EAlar) için üçüncü bölüm ayırıldığından dolayı burada bahsedilmemiştir.

EAların tarihsel arkaplanı ve genel yapısı üçüncü bölümde verilmiştir. Bizim uygulamamıza da yön vermiş olan temel kavramlar ve değerlendirmeler bizim çalışmamız için en önemli olan yönleri vurgulanmak kaydıyla özetlenilmiştir. Bunu takiben ileri çalışmalarımızda uygulanma potansiyeli olan gözde ve gelecek vaadeden EA türlerine kısaca değinilmiştir. Son kısımda bahsedilmek üzere ise KYT alanında en ciddi başarıları elde eden iki EA, USPEX ve GAtor seçilmiştir.

Son bölümde bizim uygulamamızın sunuşu algoritmik parçaları ve alana özgü operatörleri gerekli açıklamalar ve ilgili içeriğin seçilmesinde belirleyici olan etmenlerden bahsedilmesi yoluyla gerçekleştirilmiştir. Çalışmamızı gelecekte sırtlama potansiyelinden ötürü, altyapı keşfi anlayışı özellikle vurgulanmış ve altında yatan motivasyonlar açıklanmıştır.

Bunları özet olarak belirtmek gerekirse, öncelikle yöntemin gelişim çizgisi göz önünde bulundurularak bu aşamada yerel eniyileme yöntemlerinin her aday çözüm için kullanılması yoluna gidilmemiş ve bu katkı sonraya bırakıldı. Olgunlaşmamış yakınsama problemini ortadan kaldırmak için sırasıyla mutasyon sıklığı gibi süperparametrelerin eşzamanlı kontrolü, popülasyon yakınsama kriterini sağladığında rastgele bireylerle yeniden başlatılmasını ve en iyi iki bireyin 150 nesil sonra yeniden dahil edilmesini içeren bir yeniden başlatma önlemi, yükleme (crowding) katılma (reinsertion) tasarımı, ve ada modeli paralelleştirmesi yöntemleri uygulandı. Alana özgü operatör olarak ise ilk olarak moleküllerin pozunu, yani pozisyonunu ve yönelimini belirleyen parametrelerin dağıtılmadan değiştirilmesini sağlayan bir çaprazlama (crossover) yöntemi kullanılmış, bunu en kötü moleküller arası etkileşimin iki tarafından toplamda yapının enerjisine daha kötü bir katkı sağlayan molekülün diğer tarafındaki moleküle göre pozunu değiştiren bir mutasyon operatörü takip etmiştir. Son olarak ise moleküller arasındaki iyi yani düşük enerjili etkileşimleri bozmadan çaprazlama yapmak için bu etkileşimlerin her iki tarafındaki molekülü de aynı ebeveynden ve aralarındaki göreli pozu koruyarak alan bir çaprazlama operatörü geliştirilmiştir. Bunun yaratabileceği bir yapısal meyili engellemek için özellikle iyi etkileşimler sağlayan bağları hedef alan bir mutasyon operatörü de ilave edilmiş, bunun altyapı açısından popülasyon yakınsadığında mutasyon ihtimalinin artırılması yoluyla dinamik kontrolü de sağlanmıştır.

Bunlara ek olarak, yükleme (crowding) ve göç (migration) işlemlerinde birey seçimi için kullanılan bireyler arası uzaklık hesabı da alana özgü tecrübeleri ve bilgileri hesaba katarak oluşturulmuştur ve moleküller arası mesafeleri KYT açısından anlamlı olacak bir sıra değiştirme sonrasında hesaplamakla birlikte kafes parametreleri ile moleküler parametrelerin etkisini anlamlı bir şekilde dengelemeyi hedeflemektedir.

Sonuç olarak, önemini ve farklı biçimleri incelenen sitozin molekülünün iki moleküllü kristal yapısının doğru olarak tahmin edilebildiği düzeye ulaşılmıştır.

1. CRYSTAL STRUCTURE PREDICTION (CSP) PROBLEM

1.1 Introduction

1.1.1 History of Crystal Structure Prediction

Crystal Crystal structure of a solid material determines most of the properties of a material, including but not limited to density, mechanical properties, electrical and electronic abilities. Therefore, computational discovery of novel materials and computational examination of known materials is significantly dependent on the information about crystal structure. [1] Observational information, such as transmission electron microscopy images or other high-powered microscopy outcomes, of crystal structure is essential for this purpose, but it is not reliable or totally absent for many substances of interest, such as new material candidates or formations beyond our observation abilities. Hence, prediction of crystal structure serves as an indispensable tool for many research areas as computational materials design or research activities to gain insight about solid state physics, materials, or even biomolecular activities. The lack of successful computational prediction tools were/are one of main obstacles for scientific development of related research and development efforts.

Despite its central role and primary importance in material science, a successful and generally applicable theory leading predictions is absent; and computational crystal structure prediction (CSP) was arguable for even being possible. Until the fast development of computational methods for CSP in the first decade of our millennium, it was extensively considered that crystal structures are unforeseeable [2] and related attempts were doubted due to computational complexity of the task and unsatisfactory performances of used computational chemistry tools. This general incapability for CSP was even denominated as “one of the continuing scandals in the physical sciences” by John Maddox in the famous editorial in *Nature,* in 1988 [2].

The milestone of CSP was roughly the period between 2003 and 2006 that followed first partly successful attempts in the 90s, and the subsequent developments already made this field a fruitful area. Roughly or totally unexpected results and novel materials are already obtained via CSP, e.g. the crystal structure of sodium under 2 million atmospheres pressure that is unknown for any other element that forms a transparent dielectric material was first computationally predicted via optimization techniques and approved experimentally later [1]. Development continued with independent improvements and innovations as usual, with contributions from different but related disciplines such as crystallography, material science, computational chemistry, computational sciences and computer science, and blind tests played and are playing an important role to organize top laboratories around the world in the challenge of finding experimental crystalline structure of several targets. These targets are chosen in order to indicate hot topic challenges in the field, serving as a basis for organized academic competition among research groups and companies [3].

Even though the development of CSP is expected to expand horizons in several research areas by many experts, main areas that are served at the current stage can be briefly listed as computational material discovery, computation material science, and computational drug discovery. However one should also mention practical use of CSP by medical companies for patenting all of their products' crystallization (patent protection) or finding better crystal packing of the active matter in their product, as the largest contribution to the commercial use of CSP of bio molecules, which is the main target of this thesis.

The general definition of CSP can be given as the prediction of stable crystal structure at given conditions, whereas finding less stable or metastable structures may also be an important interest. The major method for CSP remains minimization of free energy, whereas some other properties such as nucleation rate or molecular dynamics analysis may be useful for selecting among crystal structures with similar free energies.

Several branches of research is ongoing for this task: computational optimization of free energy or other properties, topological approaches, structural diagrams, data mining and machine learning approaches and so on. Many of those can be argued to be either faster or better for some extent, but computation optimization is the most non-empirical, least biased and most generally applicable method among them. As a result, most of leading methodologies in the field contain an optimization process at least at one point in the workflow; but some would optimize free energy calculation methods or molecular geometry predictions as well.

Blind tests of organic Crystal Structure Prediction Methods have been started in 1999 by the Cambridge Crystallographic Data Centre, and more than one groups submitted successful predictions for challenging targets starting from the 3rd blind test. The fourth blind test reported significant progress in CSP of small organic molecules. Results of fifth test is published with the title "Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test.". Lastly, the sixth blind test was held in 2016 and targets included interesting and challenging compounds such as a flexible molecule, a long molecule with 5 polymorphs, a two molecule system, a salt hydrate, and a large molecule that contains 6 six-membered rings. Blind tests of inorganic CSP started later in 2010, and aims to serve the same function in its field [3] [4] .

In conclusion, CSP was once impossible then a intractable problem that holds importance for several research topics and academic interests, but improvements in this area promise hope for making this challenging NP-hard (non-polynomial) problem a less formidable or even a trivial tast in the close future.

1.1.2 Challenges and hot topics in the field

Indispensable features of CSP via computational optimization consist of an optimizer and energy calculator. But widespread complications and challenges such as absence of computationally cheap and chemically accurate energy calculation methods force research teams to complicate the process, and most of the procedures employed by different equips contains generation or on the fly optimization of a cheaper energy calculation method that has sufficiently low error for the given sistem, (e.g. using force fields), crystal structure optimization using this cheaper tool and rearrangement of found candidate crystal structures via more accurate and expensive energy calculation tools, e.g. DFT. Although structure optimization of molecules are not a challenging task anymore for sufficiently small molecules to be target for CSP, often a check for structural stability and rigidity of target molecule(s) is necessary.

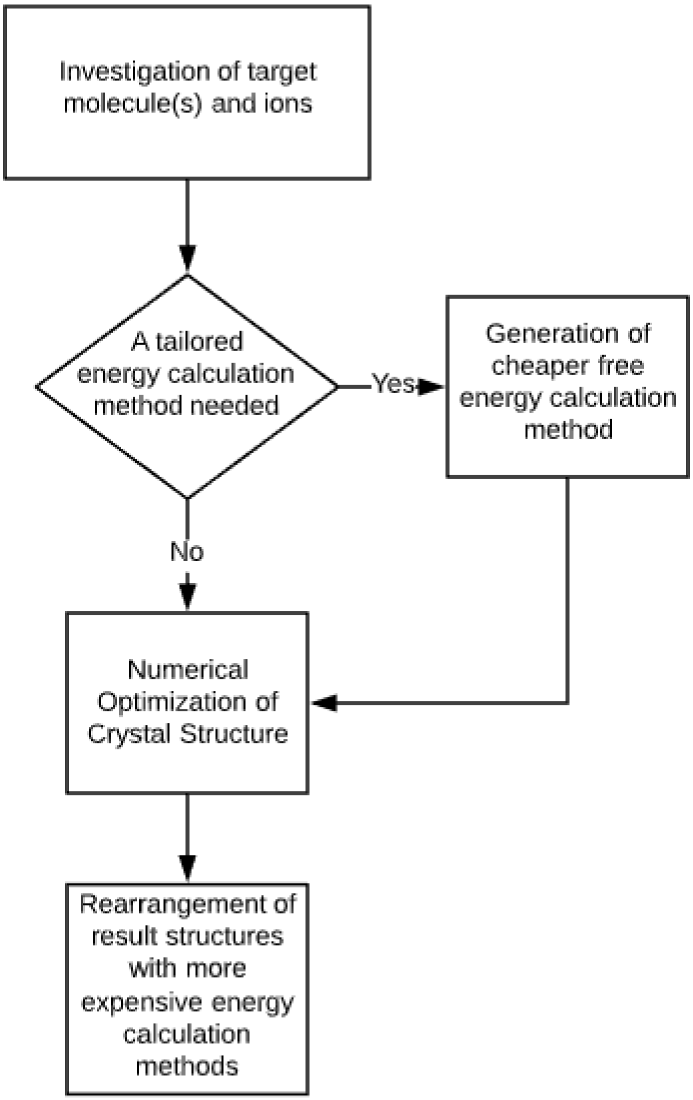
Investigation of target molecule(s) and ions
A tailored energy calculation method needed
Yes
Generation of cheaper free energy calculation method
No
Numerical Optimization of Crystal Structure
Rearrangement of result structures with more expensive energy calculation methods

Figure 1.1 : A rough representation of the general CSP procedure as a block diagram

A CSP procedure can contain more steps or can merge some of the given steps in the diagram, especially for the attempts for computation design of novel materials and in the case of on the fly optimization of cheaper energy calculation method, but the most common workflow can be described as given in the diagram. Similar to CSP procedure itself, every step of this procedure correspond to an active research area in computational chemistry or numerical optimization. In addition, contributions from seemingly unrelated areas are not uncommon, such as utilization of molecular dynamics (MD) for local optimization in the numerical optimization of crystal structure. Various challenges and hot topics arise about the other steps but we will focus on the numerical optimization of crystal structures via a given cheaper energy calculation method and not mention other important aspects (like selection or generation of the objective function or the energy calculation method employed for the numerical optimization, reranking and relaxation of found structures via more relaiable energy calculation methods such as DFT, and further investigation of optimization results) of CSP for the purposes of this work. Rest of this section will summarize why CSP is NP-hard, its special difficulties and opportunities different than any multi dimensional black box problem leading to area specific challenges and solutions in the field. Reader would be able to find some of interesting works excluded in this study in these [5-8] references.

For a wide class of problems, higher dimensionality makes the problem harder to solve, especially coverage becomes challenging if not totally impossible (Törn & Zilinskas, 1989). For the same reason, finding optimums through determination of zeroes of the function's derivative usually become unfeasible or more expensive than the optimization procedures for highly dimensional problems (Mikosch, J., & Jorge, 2006). This problem is one of the main reasons that makes development of stronger optimization algorithms crucial for CSP studies. (Hartke, 2011) For smaller problems, coverage of the space can be obtained by random walk or Sobol sequences (an example of quasi-random low-discrepancy sequences that aims to achieve an even coverage for different partitions of a unit hypercube space, first introduced by Ilya M. Sobol in 1967) and optimization can be done locally (Oganov, 2011). The more dimensional the problem, more intelligent algorithms that does not guarantee to find the global optimum but is able to find a sufficiently good local optimums are needed more. As a particle in CSP is described by 6 parameters in a 3 dimension problem, 3 parameter for the position of center of mass of the molecule and 3 to define its orientation, curse of dimensionality is especially effective for this problem. (Hartke, 2011)

Multimodality refers to the case when the objective function have more than one optimum, leading to many local optimum and one or more global optimum [9]. This forces us to employ stochastic optimization for large and complex problems [10]. A simple yet explanatory approach to understand multimodality is based on interpretation of global optimization as an Nth level local optimization: the first level of local optimization is described as finding points that have smaller value (in the case of minimization) than its neighborhood or basin, then the second level would be finding minima in this smaller search space that contains only local minima of the original search space, third level would be finding minimums among minimums of neighborhood of minimums and so on [9]. Figure 1 provides visual interpretation of this approach. From this point of view, high multimodality can be understood as the need of high levels of local optimizers in order to guarantee finding the global optimum, whereas usually efficient local optimizers are not present for even level 2 local optimization for most real world problems.

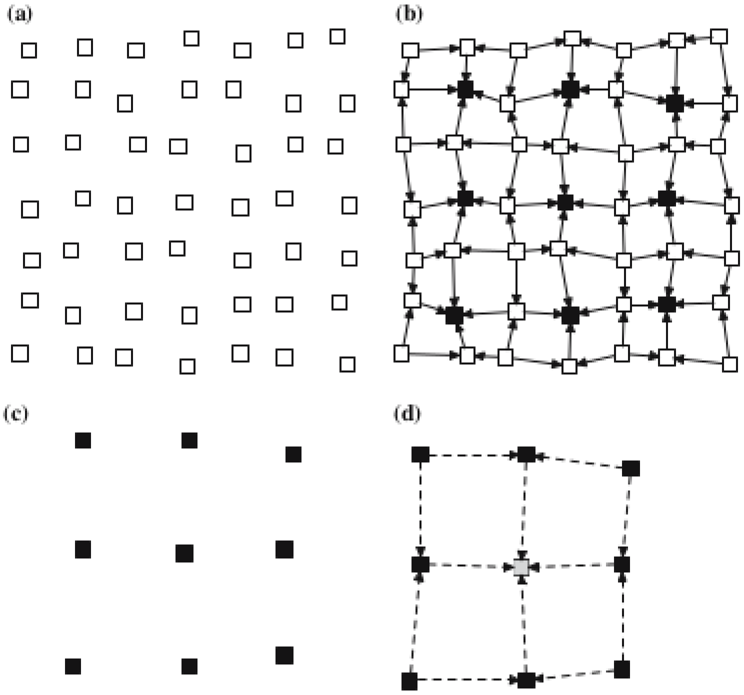

Figure 1.2 :(a) local minimizer at level 1, (b) neighborhood structure between local minimizer at level 1,(c) local minimizers at level 2, and (d) neighborhood structure between local minimizers at level 2:the grey square is the unique local minimizer at level 3 [9].

An other difficulty for CSP arises in the problems that many molecules/atoms/particles are used. Permutational complexity grows fast with the problem size, contributing largely to the total complexity. As a result, studies on multi particle systems are smaller or much less efficient than one kind of particle problems. Therefore, number of successful attempts to solve one molecule or one atoms systems with large sizes is much higher than that of multi molecular or multi atomic systems. (Hartke, 2011)

There is also considerable differences between CSP of elemental, ionic or molecular systems, as well as organic and inorganic systems. Even if they are mostly arguable, molecular systems are believed to be even more multimodal by many researchers, that leads to the increased difficulty in case of many molecules. Whereas fine constraints of molecule geometry seems to simplify the problem, elastic molecules may cause the fast energy calculation methods used in the optimization to be fairly less accurate or can increase computation cost of free energy of one candidate structure. Larger and more amorphous molecules are believed to have an energy surface with far local minima separated by larger barriers. Some molecules would be especially affected by dispersion or Van-der-Waals forces, and some will change its charge in the packing process. This kind of differences, in addition to many other distinction in energy calculations or employed heuristics, force researchers to divide their effort into more than one type of system, or specialize in one kind of system.

1.2 Methods for Energy Calculation

Quantum chemistry is in an exceptional and much better position than many other theories in contemporary scientific domains for frequently generating a quantitatively correct solution in accordance with experimental data [11].

Schrödinger's equation $(H.\Psi = E.\Psi)$ is challenging to solve because of electron repulsion term in the Hamiltonian (left most term in the Schrödinger's equation, noted with H), and three main approaches are used to tackle this problem (wave function-based methods, density functional theory (DFT), and semiempirical or empirical methods that fit a mathematical model to a dataset) by approximations and they are mainstream for different size of systems or application. The following 3 subsections of this section will explain or mention approximate energy calculation methods in 3 category: first, approximate wave function forms or quantum mechanical approach will be discussed, which are often sustainable for small systems that contain at most tens of atoms and usually used in quantum chemistry applications rather than CSP, with a special emphasis on related subjects and inspirations for CSP field. Later, DFT that is used to tackle hundreds of atoms and is sustainable for us not in optimization process but for reranking the optimization results will be resumed. Force field approach that are used in optimization will be briefly discussed in the third subsection, which briefly defines the system as atom nuclei and their interaction which is described usually with classical mechanical terms whose parameters are fitted to reliable experimental or computational data. Lastly, different approaches that may be useful will be mentioned with a particular emphasis on utilization of machine learning techniques for energy prediction as a hot topic in the field. In the rest of this introduction to energy calculation methods, all of the above will be discussed for their importance for CSP and chemistry in general with very short description of what they are but without any further detail.

Solutions and approximations to the Schrödinger equation (1926) are not only accurate for most cases, the related work paved the way for predicting conformational preferences (or geometry) of a molecule, which is very important for CSP, and explaining many chemical phenomena such as covalent bonds or auxiliary concepts like molecular orbitals [11]. Born-Oppenheimer appoximation and Hartree-Fock approaches will be briefly mentioned in the next subsection.

Although computational cost of direct solutions is far than being affordable for large systems, they cannot be cheaply repeated many times for even small systems. A fair accuracy for hundreds of atoms is needed and achieved in this field, but for example Quantum Electrodynamics (QED) calculations that are the most accurate method can be made only for the smallest molecules with a few atoms. That pushes researchers to generate much cheaper approximations or less accurate but light weight methods to calculate potential energy of a chemical system. Density functional theory (DFT) is often reliable and more affordable than methods that approximate wave function of Schrödinger equation. It is shown that interartomic interactions are dependent on density functionals in a particular way, and solution can be approximated by using the correct density functional set, where works around the density functionals compose DFT and its practical applications. They are available and quite popular for more than a half century, and development of this area is still ongoing. Many of research groups that work in CSP, including ours, use DFT for re-ranking of found crystal structures. Therefore a particular section is added for explaining DFT in a nutshell.

Structural optimization part of CSP need evaluation of energy of a candidate structure many many times, often thousands or millions of time if not more. The intensive need of computational resources to perform many DFT evaluations cannot be sustained by many researchers, therefore even cheaper methods are frequently used. A solution to that is found in the parametrization of intense calculations in the previous methods, while many uses approximations to interatomic potentials for particular systems via fitting parameters to proper data sets of similar systems. The latter approach includes force fields, that are used in this work as well as other works of our research group. A distinct section is provided for summary of this method.

On the other hand, force fields have some drawbacks and other fast methods are also interesting for this purpose. Especially those who reduces error on the fly are particularly interesting for CSP as a dynamic and comparatively young research area. A brief glance is therefore provided for this group of methods providing a few examples.

Detailed examination of energy calculation methods is out of scope of this work, related resources used in this work are suggested instead. Especially chapters 12, 13 and 14 of Chipot & Pohorille's work "Free Energy Calculations Theory and Applications in Chemistry and Biology" gives a detailed presentation of current precise methods other than DFT used in biomolecular systems, Leszczynski's "Handbook of computational chemistry" provides detailed and practical insight especially for DFT, Molecular Mechanics chapter of David C. Young's "Computational Chemistry: A Practical Guide for Applying Techniques to Real-World Problems" for force fields and Semiempirical Methods chapter for methods that replace some integrals in ab initio methods with predetermined parameters.

The following table (Table 1.1) shows used energy calculation methods for ranking the result and during optimization, together with optimization methods, in the sixth blind test of organic crystal structure prediction methods [12]. Here, the first and the second columns identify research team, while the third column denotes generation methods used for candidate crystal structures. A significant part of the generation methods are stochastic optimization methods, while many teams use random or quasi-random search with local optimization. The fourth and fifth columns give the information about how the participant groups have chosen the most probable candidates for experimental results among all the structures they found, where list one shows methods used during optimization and list two shows methods used in an optional reranking.. It is notewothy that some groups used information different than ground state energy such as vibration or critical nucleus size, but many of these are potential energy calculation methods. A major proportion of these are *ab initio* methods, particularly DFT methods, but semi-empirical methods that employ information retreated from a set of similar systems take a remarkable share in the list one submissions, due to their significantly lower computational cost. Note that most of the abbreviations signify used functional in DFT method, such as PBE, PCM, MBD, SAPT, vdW-DF, and BLYP-D3.

Table 1.1 : The optimization and energy calculation methods employed by the participant teams in sixth blind test of organic CSP.

Team	Members	Generation method	List One (L1)	List Two (L2)
1	Chadha,* Singh	MC simulated annealing	COMPASS (2.8) force field	–
2	Cole,* McCabe, Read, Reilly, Shields	CSD analogues	Fitted exp-6 potential	–
3	Day*, Bygrave, Campbell, Case, Gee, McMahon, Nyman, Pulido, Taylor, Yang	Quasi-random search (Sobol')	Atomic multipoles and exp-6	F_{vib} contributions [(XXII) and (XXV)], PCM $\epsilon = 3$ [(XXIV) and (XXVI)]
4	Dzyabchenko	Grid search	Empirical potential	–
5	van Eijck	Random search	Atomic charges, intramolecular 6-31G** energies and exp-6	–
6	Elking, Fusti-Molnar	Random generation	Empirical potential	PBE+XDM
7	de Jong, van den Ende,* de Gelder, de Klerk, Bylsma, de Wijs, Meekes, Cuppen	Random search	*q*-GRID method	Smallest critical nucleus size from kinetic MC simulations
8	Lund, Pagola, Orendt, Ferraro, Facelli*	Genetic algorithm	PBE-D2	PBE-D2 for all stages of GA search
9	Obata, Goto*	Grid search	PBE+TS	–
10	Hofmann,* Kuleshova	Random search	Fitted potential	–

Table 1.1 continue : The optimization and energy calculation methods employed by the participant teams in sixth blind test of organic CSP.

11	Lv, Wang, Ma*	Random search	optB86b-vdW	–
12	Curtis, Li, Schober, Cosburn, Lohani, Vacarro, Oberhofer, Reuter, Bhattacharya, Vázquez-Mayagoitia, Ghiringhelli, Marom*	Genetic algorithm	PBE+TS	PBE+MBD
13	Mohamed	MC simulated annealing	Atomic multipoles and exp-6	–
14	Neumann, Kendrick, Leusen	MC parallel tempering	PBE+Neumann–Perrin	Includes $Z' = 2$ structures for (XXIII) and (XXVI)
15	Sugden, Gatsiou, Vasileiadis, Adjiman,* Pantelides*	Quasi-random search (Sobol')	Atomic multipoles and exp-6	–
16	Pickard,* Monserrat, Misquitta, Needs	Random search	PBE+MBD	–
17	Jankiewicz, Metz, Podeszwa,* Szalewicz	Grid search	SAPT(DFT) fitted potential	Alternative SAPT(DFT) fitted potential
18	S. L. Price,* Hylton, L. S. Price, Guo, Watson, Iuzzolino	Quasi-random search (Sobol')	Atomic multipoles and exp-6	Different PCM treatments (all); F_{vib} for all but (XXIV)
19	Metz, Hylton, S. L. Price, Szalewicz*	Quasi-random search (Sobol')	SAPT(DFT) fitted potential	–

Table 1.1 continue : The optimization and energy calculation methods employed by the participant teams in sixth blind test of organic CSP.

20	Vogt, Schneider, Metz, Tuckerman,* Szalewicz*	Random search	SAPT(DFT) fitted potential	–
21	Zhu,* Oganov, Masunov	Evolutionary algorithm	vdW-DF	–
22	Boese	Re-ranking 10	PBE+TS and BLYP-D3	–
23	Brandenburg, Grimme	Re-ranking 18	HF-3c$_{\text{atm}}$	TPSS-D3$_{\text{atm}}$
24	Metz, Guo, Szalewicz	Re-ranking 18	SAPT(DFT) fitted potential	–
25	Hoja, Ko, Car, DiStasio Jr, Tkatchenko*	Re-ranking 18	PBE+MBD	F_{vib} contributions

1.2.1 Methods used in quantum chemistry

Solutions to the Schrödinger equation (1926) are provided by directly solving (for one electron) or approximating to the wave function. In the case that relativistic effects become important, such as molecules with very heavy atoms, Dirac theory (1928) should be used instead. An even higher accuracy can be obtained including QED corrections, yet it is rarely needed. (Leszczynski, 2012) A notable approximation to the Hamiltonian in the Schrödinger equation is Born-Oppenheimer approximation (1927), where nuclei are assumed clamped (fix in the space) first and electronic wave function is computed, then electronic energy is fed to nuclear wave function by playing the role of potential energy surface, and the total wave function is calculated as a product of electronic and nuclear wave functions. Potential energy surface notion is one of the most fruitful products for us so far; calculating nuclear positions that minimizes it provides the spatial structure of a molecule, which is directly transferred to CSP process, and the electron density distribution related with this structure displays atomic bonds and lone pairs that are also very important for many fields in chemistry [11].

Ab initio (“from the begining” in latin) calculations on programmable computers, which does not take any experimental data into account but validated by their accordance with experiments, have started to be performed at the Massachusetts Institute of Technology in 1956 [11], it was a Hartree-Fock (HF) calculation using basis set of Slater-type orbitals, and usage of Gaussian-type orbitals followed it the same year. Acquirement of a unique spectroscopic accuracy was reached only four years later. This field is benefited from the fast improvements in computer power in parallel with developments in computational algorithms like many other scientific research areas; in the current situation computational approaches to obtain molecular properties have become or are becoming faster, cheaper and more informative than experimental measurements, whereas a sufficient accuracy for most applications are possible for several hundreds or thousands of atoms, and spectroscopic accuracy can be obtained for dozens [11].

The method that started impressively accurate ab initio calculations age was HF approximations, and it is still one of the most popular methods in this field. Briefly, the main idea underlying this technique is to divide many electron wave function into simpler and computationally cheaper one electron wave functions, called orbitals, and yielding individual energy contributions that named orbital energies. Orbital concept was developed into molecular orbitals later, making sustainable computations of molecular bonds and structures, and they are also a basic element of our chemical understanding of molecules for most approximate conceptualizations of the underlying reality. HF methods always give a greater or equal result than real energy and they actually converge to a HF limit via usage of more appropriate basis sets defining wave functions, e.g. Gaussian-type orbitals, and total energy is calculated as a linear combination of them by adding a determinant in order to satisfy some quantum mechanical requirements. The procedure is basically an iterative workflow: an initial guess of one-electron orbital coefficients is chosen (usually by a semi-empirical method), and an energy is calculated, later results are used to calculate a new basis set and next step orbital coefficients and steps up to now are repeated until no important change is observed in two subsequent iterations. This method is not guaranteed to converge and expertise is often needed for obtaining convergence or detecting its non-existence. (Young, 2001)

Contemporary variety of free energy calculation methods used in biomolecular systems can be classified and understood according to their choices of 3 basic components of free energy calculations: preferred model for Hamiltonian, sampling protocol used to obtain a representative illustration of molecular configurations, and estimator of free energy difference [13]. Each component is the subject of an expertise, and in principle out of our scope for this thesis. On the other hand, some of these approaches can be insightful or inspirational for other fields such as semi-empirical energy calculation methods or CSP that lies in the scope of this thesis. Therefore we would prefer to mention the most popular methods in sampling and free energy difference estimation parts.

Sampling can be divided into two branches in reference to whether one or more states are sampled. Alternative methods for one state works can be listed as: changing the dynamics, deforming the energy surface, extending the dimensionality (note that a similar approach is present in CSP context), perturbating the forces, reducing the number of degrees of freedom (which is one of main inspirations of substructure discovery notion of my work), and multi-copy approaches. Adiabatic decoupling, staging and importance sampling are the main approaches for sampling of many state. Free energy difference estimation, on the other hand, relies to out area of interest from a more systematic point of view: they can grouped into two i) local methods that uses transition probabilities or forces and ii) global methods that takes sampling count of a given state into account or make use of energy differences [13].

Last note on this field is again on an inspiration to our field. Direction of inspiration will be explained in detail in the section 4.1, but it would be more appropriate to state here the known fact in the search of molecular configuration and especially relative spatial positions of atoms. This procedure is also basically an optimization process, local or global, in which a very interesting phenomena about coordinate system selection appears: internal coordinates and redundant internal coordinates, that are defined according to spatial positions and relation of atoms in that molecule, lead to more than 7 times less optimization cycle comparing to Cartesian coordinates (43 optimization cycle for internal coordinates versus more than 318 optimization cycle for Cartesian coordinate system) for even end-capped alanyl alanine that has only 10 atoms in its principal chain, and this great difference is often much larger for large molecules [14]. Briefly, our mutation operators are inspired from that efficiency difference between internal and Cartesian coordinates, and make use of relative poses to achieve a more effective and consistent level 2 local search ability (explained in the Figure 1.1) independent from the investigated region of the search space, therefore hopefully fasten the search or at least eliminate a structural bias of previous versions that uses Cartesian coordinates directly.

1.2.2 Density Functional Theory

In contrast with the previously seen approaches that approximate wave functions in various ways, DFT deals with challenge of solving Schrödinger's equation by solving total electron density. The basic premise of DFT is that the ground-state electron density contains all information in the ground-state wavefunction, where this information includes that for complex motions and correlations in a many-electron system; in other words, all the information about the molecule as the Hamiltonian itself. Walter Kohn was awarded with the Nobel prize in chemistry in 1998 for this development of the DFT and its theoretical proof was published in 1964 together with Hohenberg [15]. Nobel prize can be an appropriate indicator of the importance and usefulness of this theory and its reputation is very high and which is almost unrivaled for energy calculations of mildly large systems with hundreds of atoms.

Whereas many variations and improvements of DFT may rely on some other theorems or assumptions, basic theorem of DFT is as the following:

1) The electron density $\rho(r)$ determines the external potential $v(r)$.

2) For a trial density $\rho1(r)$, such that $\rho1(r) \geq 0$and integral of $\rho1(r)dr = N, E0 \leq E[\rho1]$, where $E[\rho1]$ is the energy functional. (Hohenberg & Kohn, 1964)

Functional means function of a function, as energy is a function of electron density which is a function of spatial position, or spatially dependent in other words [15].

Theoretical roots of DFT reaches until Thomas-Fermi theory (1926) which approximates electronic structure of atoms using one-electron ground-state density and insufficiently rough to describe binding of molecules. This idea is later combined with Hartree's orbital method, firstly by Slater in 50's as an unintentional invention of the simplest DFT (Young, 2011), which is followed by Hohenberg-Kohn theorem in 1964 which is the basics of DFT, providing the proof that an exact method based on ρ(r) exists in theory. The variant of DFT contemporarily in use is Kohn-Sham DFT, which differs from previous one mainly by self-consistent equations that are needed to solve for a set of orbitals whose density is described as equal to that of the system. One should also note that physical meaning and interpretation of Kohn-Sham orbitals are still debated since their eigenvalues do not correspond to the energies from photoelectron spectroscopy unlike HF orbitals that are not in accordance either.

However, DFT describes well the behavior of electrons like the others do, and globally used for this purpose [16].

Accuracy of DFT calculations are mainly threatened by the exchange-correlation (xc) energy, that must be given in terms of electron density. This small but crucial contribution can be exactly calculated for small systems, but the computational cost of the exact calculation is even higher than that of direct solutions to Schrödinger's equation [16]. This cost compel us to approximate the xc contribution which result in restricting the accuracy of the whole approach. The exact formula of density functional is not achieved, thus any functional in use can have advantages and disadvantages. (Young,2001)

The simplest approximation to density functionals, local density approximation (LDA), that reached sufficiently accurate results for chemical reactions with the valuable contribution of generalized gradient approximations (GGA) in the late 80's. Further accuracy and more general applicability was attained by hybrid methods, proposed by Becke in 90's replacing GGA exchange partly by HF exchange, yielding the most popular functional known as B3LYP. PBE, on the other hand, dominates applications on the materials, [16] as we can see in the supporting information of the aforementioned sixth blind test [12]. Also note that some of approximations to the electron density functional are developed by parameterizing functions to fit experimental data, therefore called semi-empirical versions of DFT, even if the method itself is *ab initio* in its essence. (Young, 2001)

The popularity of DFT methods are based on their lower computational cost comparing to other methods that yield similar accuracy for similar systems [16]. A remark on computational complexity therefore should be noted: Coulomb repulsion (or electron-electron repulsion) is only integrated over electron density which is a three-dimensional function, thus DFT methods scale with the cube of the electron number, which is a degree less than HF calculations that they can even over-perform sometimes for accuracy (Young, 2001) [15].

1.2.3 Force fields and other semiempirical methods

In this section we will focus on fast calculation methods for interaction energies or other interested properties. Common trait of these methods are their emphasis on reducing computational cost. First class of algorithms mentioned below achieve low cost by omitting or roughly approximating expensive integrals is HF calculations or some other accurate approach. We will only give general information and provide a number of examples to summarize these methods. Some of them are or may grow into a noteworthy alternative to the second class of algorithms, force fields; and some are potential substitutes to DFT or *ab initio* methods in applications where cost is important, such as setting molecular configuration and flexibility before optimization of molecular poses or investigating exact configuration for flexible targets on the fly in CSP. These methods are also important for determination of initial guess for *ab inito* methods. Our main focus will be on force fields, since our energy calculation method currently uses a force field, and they are currently invariably used for the examination of large systems, i.e. systems with thousands or more atoms.

Semiempirical methods are very similar to HF methods in sense of general structure, as they approximate to the solution of Schrödinger's formula by means of approximate Hamiltonian and wave function. Generally, expensive integrals are totally omitted or approximated by an interpolation or approximation of a parameterization set obtained from experiments or *ab initio* calculations including DFT results. Therefore, a general deficiency of these methods is inevitable lack of accuracy for molecules that are compellingly different than the employed database. In addition, they mostly suffer from volatile outcome and the number of properties that can be properly predicted is not comparable to previous methods. On the other hand, it is believed that they are usually less sensitive to the parametrization set than force fields. (Young, 2001)

The desired results for a semiempirical method are less widely reliable and usually less accurate than *ab initio* methods, as a result of trade-off between computational cost and quality of the outcome. They are most often parameterized for geometry and heat of formation of a molecule, sometimes for dipole moments, heats of reaction or ionization potentials. More specific properties are not uncommon, as these methods can be used to predict properties diffident than those used in the database, but accuracy can often be increased by adding property of interest into the data set. (Young, 2001)

One of the outstanding examples of these methods is the extended Hückel method. Despite not being an especially accurate nor smooth method, it is commonly used for its ability to bring low CPU times and wide applicability (nearly all of the periodic table) together, which makes it unique and very handy for inorganic modeling. Only valence electrons are considered and they are calculated by the orbital overlaps and experimental electron affinities and ionization potentials. This method is referred as tight binding calculation in some of the literature.

Other semiempirical methods are or was popular for calculation of organic molecules, such as the Austin model 1(shortly AM1) or parameterization method 3 (shortly PM3) and its transition metals extension.

Semi-*ab initio* method 1, shortly SAM1 is frequently used as a substitute for *ab initio* methods for large organic molecules which are not feasible in them. It omits less than previous semiempirical methods, but uses significantly less computational resources than *ab initio* methods.

An other special semiempirical method specialized for organic molecules is the Gaussian method (shortly G1, G2 and G3). It is developed as a response to the common observation of a systematic error for organic molecules in *ab initio* calculations. A correction equation is used to extrapolate high accuracy *ab initio* results into less accurate results, where both the extrapolation equation and its parameters are empirically determined. Many different variations are found in the literature, specialized for various aims such as less CPU time in exchange to small loss of accuracy or reliable predictions on large molecules.

Force fields, or molecular mechanics, are different from all the methods above in the sense that they do not utilize quantum mechanical terms in the definition of interaction energies. Ignoring some of the expansions, they do not take electronic motions into account directly and approximate the energy of a structure considering only the position of nuclei. They are invariably employed to extend our computation ability on systems with excessive number of atoms, unsustainable with other methods, as they consume only a little fraction of computational resources compared to *ab initio* calculations [17].

The underlying model is much simpler than *ab initio* calculations: nuclei are taken into account almost as mechanical components that interact via mechanical forces. Energy contributions of processes such as bond stretching, angle bending and torsions around a single bond are approximated with non-quantum mechanical formulae. Acceptable accuracy is obtained in some applications that utilize functions as simple as Hooke’s law [17], but potential pitfalls and important limitations should be considered when using this approach (Young, 2001). One of the vital characteristic of molecular mechanics is transferability, that enables the usage of a force field in various systems and in even much larger systems like polymers [17], which should be sought by the researchers. In force fields, atom types are widely used and for example carbon atoms are treated as different elements according to their hybridization type, functional group or being in a ring or not (Young, 2001)

The performance of the molecular mechanics is dependent on: the formula of the energy expression, the parameterization set or the data that are fitted, the optimization technique to match model’s results with data, and practical insight to use it accordingly considering its strong and weak aspects. (Young, 2001) Generally talking, the driving equation has simple elements to describe inter- and intra-molecular forces in the system. The potential energy is described as the sum of independent contributions of various phenomena [17]. These may or may not explicitly include electrostatic interactions, van der Waals forces, hydrogen bonding, or cross terms. But they make use of at least one valence term (terms in the energy expression that describe only one feature of the molecular shape), uses a single aspect of the molecular shape, (Young, 2001) which usually is computed as sum of various power of the difference between treated relative property (such as bond length) and a fitted value for its equilibrium state [17]. Bond stretching, one of the most critical and wide-spread elements for bonding inclusive force fields, is usually defined by a harmonic oscillator equation, Morse potential or Leonard-Jones or quartic potential. (Young, 2001) Van der Waals forces are especially important for organic molecules, and Lennard-Jones potential is frequently employed to calculate these. Some force fields would consider different bonds between the same atom pair separately, C-H bonds when 2 H are bonded to a C and one H bonded to a C can be an example [17].

One of the distinctive structural choice among force fields is whether they include cross terms. They are defined as descriptive elements of the energy expression about how one motion affects the other, e.g. stretch-bend terms that express how equilibrium bond lengths change with bond angles. The force fields with no cross terms tend to compensate it with sophisticated functions, and there are force fields that include as much as nine different cross terms on the other hand. (Young, 2001)

A warning should be made for a common misunderstanding about the force fields. Even if explicit terms that aim to calculate different contributions are present in the formula, parameters are fitted to the data and nor the data neither the optimizer of constants seek or make possible to calculate those individual contributions solely by corresponding terms. Therefore a molecular mechanic model should be understood as a unified approximation, none of separate terms have to mean expected contributions, even if some terms (especially bond and angle terms) are naturally sufficiently independent of others can be expressed by corresponding part of the formula [17].

Force fields are derived and used for specific applications in general, so its properties such as transferability should be interpreted accordingly. Moreover, the variety of existing force fields may seem outrageous without acknowledging this. Useful considerations are found in the literature for both selecting a force field from existing ones and tailoring one for a specific application, (Young, 2001) but both process are expected to contain trial and error approach intensively [17]. For example, force fields that are designed to express only non-bonding interaction between molecules for applications assuming rigid and non-bonding molecules such as molecular CSP, do not include bond terms and should not be used for bonding problems.

In order to provide a short list for existing force fields, we will mention the following ones: AMBER that has no cross terms but a sophisticated electrostatic term, designed for proteins and nucleic acids; CHARMM that is applied to nucleic acids, biomolecules, molecular dynamics, solvation, crystal packing, vibrational analysis, and QM/MM studies; CFF force fields that aims consistent accuracy for conformations, vibrational spectra and enthalpy, and strain energy; CHEAT that uses external atom definition for successive modeling of carbohydrates; EFF that models hyrdocarbons with three valence terms and five cross terms; MM4 and other MM force fields are general purpose that emphasize organic systems, having six valence terms and nine cross terms; OPLS for bulk liquids; UFF that models all the periodic table which is often used with an additional electrostatic term; and last but not least YETI, which is commonly used for non-bonding interactions between biomolecules and small substrate molecules such as docking applications. (Young, 2001)

1.2.3.1 Symmetry-Adapted Perturbation Theory potentials for DNA bases

Our study uses force field developed by fitting interaction energies obtained from DFT-Symmetry-Adapted Perturbation Theory (DFT-SAPT) for DNA bases. These force fields are highly accurate aiming a limited range of applications, namely cytosine [18] guanine [19] , adenine [52] and thymine [JCP 2019 Just accepted] clusters. Is is shown that they are in well accordance with various *ab initio* methods, namely DFT-SAPT and SCS-MI-MP2 and overperform AMBER force field which is frequently used for nucleic acid clusters in sense of accuracy.

Deoxyribonucleic acid (DNA) bases are known to form triplexes, quadruplexes and many other complex structures, beside Watson–Crick base pairing that is responsible for double helix structure of DNA chains in living cells. Moreover, their interactions and behavior in the presence of metals are an interesting research area for technological applications, such as organic photovoltaic tools, semi-conductors and biochip sensors; and they are observed to form 1-D and 2-D supramolecular planer networks over metal surfaces. These and other technological and scientific interests, together with structural variety of self-assembled DNA bases, drive researchers into development effort for advanced force fields for nucleic acids. Single-molecule methods are sufficiently developed for the time being for inspection of nucleic acids on metal surfaces, but their further advancement is desired to increase accuracy of Molecular Dynamics and CSP studies [18].

The first principles potential developed by Manuykan and Tekin [18] was the first attempt that aims to model the interactions in cytosine oligomers, and its ability to successfully predict cytosine cluster structures up to the hexamer case. Later, a similar study was made for guanine DNA base and its prediction ability is proved up to tetramers [19]. The missing force fields for the remaining DNA bases, adenine and thymine, were recently developed [52]. Later CSP studies carried out by Demir and Tekin [Predicting Polymorphic molecular Crystals with a Machine Learning Assisted Parallel Crystal Structure Search, in preparation] showed that these force fields can be used to predict the crystal structure of cytosine.

Functional form of these force fields is the same with previous acetylene force field [20]. (see formula below)

$$V = \sum_{i \in A}^{sites} \sum_{j \in B}^{sites} \left\{ \alpha_{ij} \exp(-\beta_{ij} r_{ij}) + \frac{C_{ij}}{(r_{ij}^{6} + c_{ij}^{6})} + f_0(\delta_0^{ij}, r_{ij}) \frac{q_i q_j}{r_{ij}} \right\}$$

In this formula, resulting potential energy is only dependent on the distance between interaction sites, and α, β, C are fitting parameters which, in cases of both guanine and cytosine, lead to 10 pair interactions and 30 fit parameters by assuming every element in the target molecules as a different site. Partial charges are electrostatic potential fitted, and f_0 Tang-Toennies damping function is employed to overcome the divergence issue of the Coulomb interaction term that arises when r approaches to 0. An other damping is added to avoid unphysically large dispersion contributions.

The Levenberg-Marquardt nonlinear weighted least squares method is used to fit the parameters. In this method, the formula below is minimized.

$$\chi^2 = \sum_{i=1}^{N} \sigma_i (y_0(x_i) - y(x_i; \alpha_{ij}, \beta_{ij}, C_{ij}))^2,$$

Here, y_0 refers to *ab initio* results of given structure, and y refers to energy value obtained by the model being optimized. Weight terms σ are calculated according to the interaction energies of each dimer: σ i was set to $1/(y0)\,2$ for$E \int 1mH(2.6\,kJ/mol)$ and $exp((1 - y0)/3)$ for $E \int \leq 1mH.$

Although exponential terms which are present in both dumping function and potential energy function increase computational cost, the performance of these force fields is superior to any similar cost approach by means of the achieved accuracy.

1.2.4 Other methods

Typical CSP procedure is a two stage process, where a cheaper energy calculation method and an optimizer is used for structural optimization phase, and obtained results are reranked and/or refined via DFT or other *ab initio* methods, as we described before. The accordance between two employed energy calculation method should be properly checked in an ideal study, but it can be unfeasible for many cases. The questionability of cheap method, e.g. force field, prompt researchers to check the success of their method on CSP or other prediction problems directly, as in Tekin’s studies [18] [19] [20]. But an other interesting approach to address this problem is to optimize energy calculation method on the fly, in parallel with structural optimization. (Hartke, 2011) A method that optimizes parameters of the employed model potential [21] and an other work in the CSP field that uses machine learning interatomic potentials [7] [22] are noteworthy. Also, some studies already start to expand the limits of CSP towards material discovery with unconstrained composition [23].

2. OPTIMIZATION MEHTODS OTHER THAN EVOLUTIONARY ALGORITHMS AND THEIR APPLICATIONS ON CSP

2.1 Summary of Optimization Methods

There is no efficient technique to solve highly dimensional multimodal objective functions for global optimum, in other words, target problems of global optimization are often unsolvable in a mathematical sense. (Törn & Zilinkas, 1989) However, global optimization area witnessed an immense increase of effort, an explosion in the number of available techniques and sublime success for many class of problems, so that even evaluating their success against each other became a challenging task. [9] "No free lunch theorem" [24] still holds today as no technique overperforms others for any problem, even not for a large class of global optimization problems.

In such a situation, we often decide which optimizer to use considering the nature of the problem and our estimation about the needed effort and performance of the implementation. We will first describe the classes and important properties of problems and define CSP problem according to this classification and then examine optimization algorithms and note some important considerations about global optimization tasks.

Target problems are classified in many different ways in the optimization related literature, but some properties are often used by authors to pose meaningful differences among problems. First frequently considered characteristic is constraints: unconstrained problems are usually tackled via different tools than constraint algorithms. (Törn & Zilinkas, 1989) CSP problem is defined by lattice vectors, whose negative or very small magnitudes yield physically impossible or meaningless structures, and positions of atoms in the lattice. This definition implies that CSP is a constraint problem. On the other hand, constraints are also very important in the selection of a promising method for a problem. In the scope of CSP, it should be indicated that those constraints do not reduce the number of candidate global optimizers to a great extent. (Oganov, 2011)

Another important property is the definition of parameters: continuous or discrete. Discrete problems are being selected from a finite set, whereas continuous parameters are defined on a bound or unbound infinite set and the number of feasible values is called infinite neglecting the numerical precision limits. (Mikosh & Jorge, 2006) Feasible and promising solution methods greatly differ among these two classes, but there are many global optimization methods able to solve both class of problems, including problems whose candidate solutions are a combination of discrete and continuous parameters. (Törn & Zilinkas, 1989) CSP is a fully continuous problem by its nature, but some implementations prefer to define crystal structure with a combination of continuous and discrete parameters, where first imply positions and lattice vectors and second is used for cell type or symmetry type. Many groups including ours prefer to repeat the optimization process for certain types of cells, but solving the problem for every possible alternative of a discrete parameter do not indicate an optimization procedure, so the problem can be considered fully continuous unless optional discrete parameters are used. (Oganov, 2011)

Computation cost of a single evaluation of the objective function is also an essential property that depends on the problem. Cheaper objective functions could be optimized through much more evaluations and the cost of candidate solution generator may become very important. The opposite case is valid for high cost evaluations. (Törn & Zilinkas, 1989) In general, the optimization procedures used to predict crystal structures uses potential energy of the candidate structures. CSP may be performed using more expensive *ab initio* energy calculation methods, cheaper semi-empirical approximations of them or much cheaper methods such as force fields. (Oganov, 2011) These methods are explained in the previous section. Whereas our group generally use the potential energy values calculated via force field methods, some of our applications such as FFCASP (Fast & Flexible Crystal Structure Prediction), which is the modernized and parallel version of CASPESA (Crystal Structure Prediction via Simualted Annealing) used different objectives to predict structures mostly bulk crystals. [25] Note that evaluation cost of a single energy calculation via force field is much cheaper, however its cost is not low enough to be deemed cheap for large systems with many molecules. Therefore, even if relatively expensive calculations are held for offering next candidates by the optimization algorithm their share in the total cost is not expected to have an important share.

Most of local and some global optimization methods make use of the derivative of the objective function. These can be implemented using analytical, numerical or approximated derivatives when possible. Derivative including methods usually converge to local minima much faster (Törn & Zilinkas, 1989) and frequently adopted for local optimization purpose in the CSP field. (Oganov, 2011) On the other hand, some functions are totally impossible to generate a reliable derivative and some derivative methods add a grand cost to the process and using derivatives become a trade-off. Therefore, appropriateness of a method for a particular problem is closely dependent on the cost of derivative, even if a derivative is possible.

Energy value of a candidate structure is the same anytime when calculated with the same method, so it is called as a deterministic function like any function optimized for CSP purpose so far (Oganov, 2011). But there are problems that exhibit different results for distinct evaluations, called stochastic functions, and their optimization is held by special methods or adaptations of optimization routines for deterministic functions (Törn & Zilinkas, 1989) which are away from our scope in this work. Note that optimization techniques are also classified as deterministic vs. stochastic procedures; stochasticity of objective functions and of optimization process refer to totally different topics and should not be confused.

Dimensionality, i.e. the number of dimension or parameters to optimize, is very crucial for solution method. The essential contribution of number of dimensions is about the size of search space and it is one of the main reasons why brute force search or similar applications are not feasible or possible for highly dimensional problems. Most objective functions that are posed or may be posed with a large number of dimensions do not have an efficient solving method and therefore they are often challenging for global optimization and no guarantee can be given for founding the global optimum. This property is usually referred as the curse of dimensionality. (Törn & Zilinkas, 1989) The curse of dimensionality is effective for CSP problem and optimization for bigger number of molecules is one of the main challenge in the field. (Oganov, 2011)

The shape of objective surface posed by the objective function, for example its convexity, (Mikosh & Jorge, 2006) as well as the size ot the basin of attraction and especially the shape of the function around the global minimum [9] are some other major considerations as well. For CSP, a rough, very wavy and most unpredictable objective surface is faced which makes the task a particularly challenging case. (Hartke, 2011)

One of the most important aspects about the the function is the number of local minima for optimization purpose. An optimum is defined as a point in the search space which gives better results than any other point surrounding it, or close to it for a given distance measure. Problems that have more than one optimum, such as multimodal problems, can be solved via a series of nested local optimizers in theory. But many levels of local minimization are only possible or much more feasible to be held in an approximate fashion by global optimizers in practice [9]. CSP problem is highly multimodal, i.e. it contains a large number of local minima, as we mentioned before in the section 1.1.2. In addition, it is not possible to take derivatives or other useful information for local optimizer at second or greater level. (Oganov, 2011)

Even reaching to the point which is better than any surrounding point in a space would need an infinite number of objective function evaluations by brute force approach for the case of continuous functions and the number of possible evaluation points often achieve unsustainable numbers for discrete problems. Thus, any solving method must make use of less point to find this point in a reliable and deterministic fashion. This goal is reached by local optimizers. Two most popular examples of local optimizers are the gradient decent method and its derivatives such as BFGS, that make use of the derivative information, and the simplex method, which is often used for functions with impossible or very expensive derivatives for its derivative free usage capacity. But these methods can only provide the best point in a limited part of search space usually called a basin of attraction in the literature.(Hartke, 2011)

Global optimization appears as the job to find the best point in the search space, defined by parameters and constraints, and brute force approach already seems impossible for smallest and easiest cases for continuous problems as it would need infinite number of objective function evaluations to cover all the space. A combination of a local optimization technique and an approach to provide starting points covering hopefully every attraction basin would yield a deterministic global search algorithm. Even if authors of these algorithms may claim to guarantee exact global optimum, most methods offer wrong predictions occasionally. Usage of deterministic global optimization techniques in the CSP field could provide successful solutions for very small systems and failed for even tasks with two dozens of atoms. (Hartke, 2011)

A degree of stochasticity appears when we generate starting points with a stochastic approach. Some examples for this are pseudo-random and quasi-random starting point in the case of random search, and Sobol' series that tries to maximize coverage with fewer point than randomly proposed point sets. Both find an area of application in the CSP field, but the problem of scaling with dimensionality could not be alleviated enough to examine systems with hundreds to thousands atoms using reasonable computational resources. (Hartke, 2011) Still, many teams in the CSP blind tests continue to use this approach. [4] [12]

Stochastic global optimization methods combine deterministic process to promote convergence with randomness that helps avoiding the deadlocks of deterministic approaches. Generally, insight about target problem or optimization process is used to come up with special heuristics in order to obtain additional bias to shift the coverage through more promising subspaces to tackle global optimization more cleverly. (Hartke, 2011) Algorithms that make use of heuristics to explore and exploit application specific heuristics on the fly, the metaheuristics, hold a dominant position among the cutting-edge optimization methods for the last decades. [26]

Metaheuristics are game plans to guide the optimization through optimal or near-optimal solutions. They are not problem-specific but mostly they are able to make use of domain-specific knowledge in the form of heuristics that are still supervised by upper level strategies. A large range of metaheuristics are present, from simple local search to complicated learning systems. [26] Obtaining information about the search space from previous attempts and taking advantage of it is the essential idea underlying this class of techniques. Incorporation of emerging swarm intelligence, which is a mechanism to avoid traps or confined closures of the search space, special tricks to increase coverage or eliminate bias are popular ways to achieve this goal. [27]

Coverage is one of the most important criteria to evaluate an optimization algorithm, but directing search into more promising areas is another achievement that conflicts with the first. The success of an optimization algorithm design seems to depend on achieving an appropriate balance in such trade-off conditions. (Törn & Zilinskas, 1989)

Two popular algorithmic concept in this field are determination of trajectory according to previous trials by algorithms usually referred to as trajectory methods and making use of information stored by a group of past trials collectively, namely by population-based algorithms. [26] We will mention a few examples from both groups in the next section which is dedicated to nature-inspired methods.

An important remark can be made on hybridization of metaheuristics. An important share of successful domain-specific applications is hybrids of metaheuristic global search algorithms together with a number of general-purpose algorithms. Hybridization can be achieved via component exchange among different methods, cooperative search methods involving information flow between algorithms running in parallel and integration of metaheuristics with some other artificial intelligence methods. [26] A detailed taxonomy of hybrid metaheuristics can be found in [28]. This group of methods is especially noteworthy for our purpose, because most of leading CSP applications involve local optimization and global optimization hybridization [12] [4] and, similarly to many others, some applications of our group show relay hybridization according to the taxonomy in [28] by combining particle swarm optimization and simulated annealing in a pipeline fashion. [25]

Optimization research field developed predominantly in practical applications and most researches focused on creation of novel tools or implementation of existing methods for new areas. Mathematical proofs and theoric understanding are considered insufficient by majority of experts, including those who contribute greatly to this repository of methods. Concordantly, convergence and efficieny analysis of metahuristics is an open research field. On the other hand, there is a noteworthy disagreement among academics about the significance of the current research trends: some assert that it provide accumulation of insight and experience which is indispensable for further development, some others claim that it already become a distracting and detaining behavioural pattern [29].

2.2 Nature-inspired Methods

Last sixty years has witnessed a particular attention towards optimization algorithms inspired from natural process. Numerous accomplishments in various disiplines obtained by nature-inspired optimization methods, many of which invented by interdisciplinary partnerships, pumped the popularity of these algorithms, which bring discovery of new ones and development of existing methods in return. Many natural processes may be interpreted as optimization procedures. Some phenomena seems optimizing certain properties of its subjects because more stable structures are the ones optimized for that property, as in the case of biological evolution and reproductive success of species, and some systems are driven to more appropriate variations by natural forces, like crystal structure in annealing process. Also, some systems evolved and are evolving dominantly for their capacity to produce favorable outcome in some area, like neural system of animals on learning or collective behavior of social insects on finding food, so their mechanism serve well as a source of inspiration or discovery of optimization process. In the historical development of numerical optimization and especially stochastic global optimization, natural phenomena played the roles of metaphor and source of discovery extensively. [30]

On the other hand, concerns and criticism made for metaheuristics is even more sharp when it comes to most nature-inspired metaheuristics. [29] In addition, some researchers state that using natural metaphors as justification of production of new methods lead the field into an unnecessary chaos by creating a surplus of similar methods and an exaggerated value given to inspirations drives the discipline away from scientific rigor. [31] Even by those who defend the current research line for providing a historically appropriate accumulation of information, the need to examine and clarify the extent of natural inspiration is generally excepted [10] [30].

The limits of inspiration from nature is well beyond the area of optimization. The popular research line of artificial neural networks, for example, stands tall among other natural inspirations out of optimization domain. Still, the global optimization enjoys the biggest contributions of natural metaphors and inspirations by far in the computer science and mathematics. [32]

Major examples of general purpose nature-inspired optimization algorithms can be listed as Ant Colony Optimization, Particle Swarm Optimization, Evolutionary Algorithms (including Evolution Strategies, Genetic Algorithms and Evolutionary Programming), Artificial Bee Colony Optimization, Firefly Algorithm, Cuckoo Search and Simulated Annealing. In the field of CSP, notable examples include but are not limited to: Firefly Algorithm [33], Particle Swarm Optimization [34], Evolutionary Algorithms [35] [36], Simulated Annealing (Oganov, 2011) on which the first computer simulation of a molecular system is performed [17], and Parallel Tempering. [12].

2.3 Other Methods Developed by Our Group

CSP strategy of our group is consistent with many other groups: generation of a reliable cheap energy calculation method (for the molecular crystals) is followed by a global optimization process that uses the generated method, then prescreening of obtained structures, DFT calculations and result analysis. Prescreening includes

elimination of repeated occurrences of structures, very similar structures and unfeasible structures. This step is performed with a multistep data mining strategy, in particular, the Structure-Matcher feature of PyMatGen, followed by selection of lowest scored structures, that are further eliminated with a density threshold. After the number of possible structures is reduced (in a study [J. Phys. Chem. C, just accepted, 2019], from more than five thousands to two hundreds), Quantum Espresso program is employed to perform DFT calculations. GGA exchange-correlation approximation of Perdew, Burke, and Ernzerhof (PBE) and the ultrasoft pseudopotentials of all atoms are selected among DFT methods, considering the problem in hand. After a second similarity search among relaxed structures, results are analyzed. Analysis of the results is made by clustering using matminer, resulting in a similarity matrix where structure similarity is determined using site and structure fingerprints. Visualization of found and clustered structures with their clusters is made via a dendrogram plot, where important results such as experimental structures and common appearance of clusters are given as structural models expressing crystal morphologies in addition. [25]
Different algorithms are developed for various CSP tasks, such as CSP of nucleic acids and Cu-TCNQ [J. Phys. Chem. C, just accepted, 2019] systems. One of the main differences among those is the objective function. Along with force fields, an objective function that serves to maximize the number of interaction that stabilizes the crystal (this kind of objective function is used primarily in covalent crystals (also known as network solids)). Bond length constraints and type of intermolecular interactions are gathered from crystal structuredatabases, available experimental data or DFT calculations. [25] Employed force fields are either the ones presented in the section 1.2.3.1, or generated in a similar fashion for the examined system.

The other main distinction among our methods is the global optimization algorithm. CrsytAl Structure PrEdiction via Simulated Annealing (CASPESA), the first version of those, uses a Simulated Annealing implementation. This implementation utilises many tricks to evade unnecessary calculations of intermolecular interactions. First, only half of the surrounding cells are token into accounts and these contribution is multiplied by two, as the surrounding 26 cells are point symmetric around the examined cell. Second, very little changes are ignored and the molecules that changed their pose significantly are labeled as “dirty” and only the interactions among those and between those and other molecules are recalculated, if the cell is not dirty. This trick works well with simulated annealing, because it needs subsequent evaluations of often slightly changed systems, moving in space with steps that change only one molecule's pose. A further preceeding of CASPESA is FFCASP, added a previous PSO step to the optimization procedure, resulted in a relay hybridization of two nature-inspired algorithms. The following figure exhibits the flowchart of whole process of FFCASP application for a molecular crystal, summarizing the process explained in the previous paragraphs.

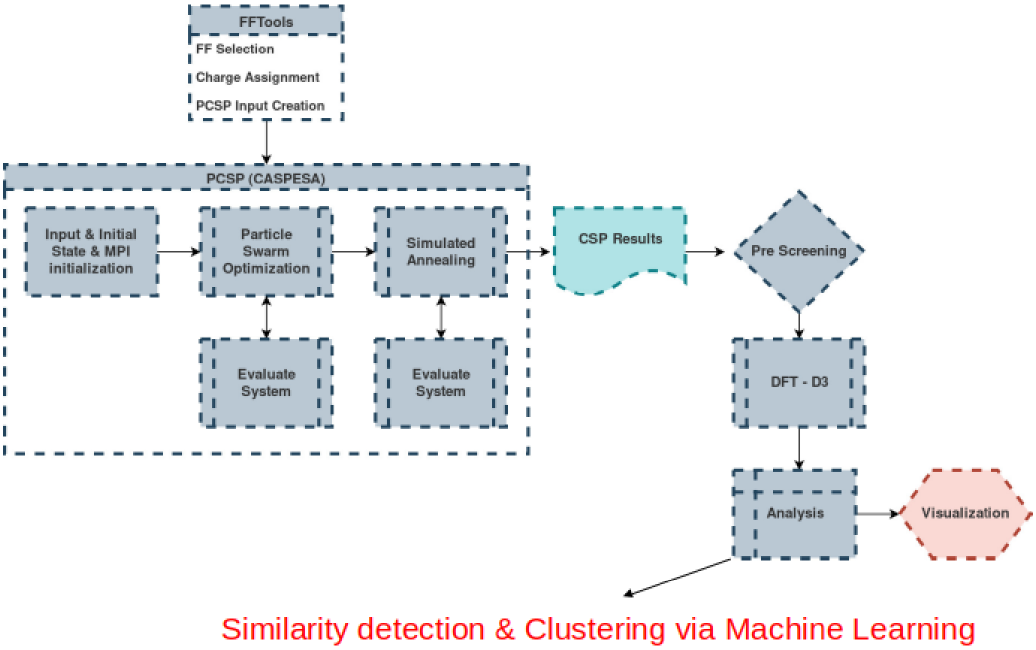

Figure 2.1: Flowchart of FFCASP procedure. CSP procedure used within studies of our group differs only in first or second step, so it can be token as a general model for our CSP procedure.

2.4 Other applications

Along with global optimization methods that combine coverage methods (such as random points or Sobol sequences) with local search, applications of a rapidly developing field in the CSP studies if noteworthy: machine learning and data mining. Deep learning strategies generally aim to retrieve useful information from the datasets or produced candidate structures on the fly, and put this information to use in order to make a guess about lower energy structures. [37] While many of them can be found in the literature (Hartke, 2011) (Oganov, 2011) (Leszczynski, 2012), a remarkable example of crystal structure classification [38] and material discovery [5] is promising for the close future of machine learning applications even in rather different CSP field.

3. EVOLUTIONARY ALGORITHMS (EA)

3.1 Introduction

Evolutionary Algorithms (EAs) is one of the main branches of nature-inspired metaheuristic stochastic global optimization algorithms [26] and find a large application areas in science and engineering since last decades. [32]

In this section, historical background of EAs and underlying inspirations will be shortly presented first. The united EA approach and a practical framework which is used as model in our study will be presented next, followed by a summary of basic concepts and considerations in EA design together with some noteworthy components addressing those concerns. Two popular example of sophisticated EA variations will be mentioned shortly, addressing their promising features for CSP problem. Then, two most successful EAs in CSP field, which are encouraging and reference for current and future work of our study, will be explained in detail.

3.1.1 History and motivation of EA

History of EAs starts in the middle of last century, with independent inventions of Genetic Algorithms (GA) (Holland, 1975) and Evolution Strategies (ES) (Rechenberg & Schwefel, 1971) in different continents. Both groups (Back,2000) remained uninformed about the other discovery for a few years, but the notification of the other branch was far than enough to join those. (Back, 2000) Evolutionary Computation (EC) field was mostly an academic sphere of interest for a considerable time and gained notable functionality on real world problems a few decades after its invention, which decelerated the rate of interaction among competing groups (Back,2000). Another reason for that was probably the differences in their philosophy and understanding of evolution, which was apparently the inspiration source of both algorithms, and further distinction in their development lines due to these. Both aimed the same achievements, exploring the evolutionary intelligence and exploit that for optimization purposes along with the examination of biological evolution itself; but GA got an advancement history with simpler operators and an emphasis of selection and mating strategy, whereas ES tend to keep those components unchanged and use more sophisticated mutation operators in general. Also, GA got its first successes in discrete optimization, but ES was mainly used for continuous problems. For a long time of their history, inspirations from the other method was not mainstream among researches. They kept adding stochasticity and maintaining convergence issues in different ways. (Back,2000) Some researchers refer to that point by calling GA more crossover oriented and ES more mutation oriented. Also note that mutation is considered more successful for randomize the search process and convergence to local optima, and crossover is considered more fruitful for combining information of different candidates.

This distinction seemed artificial and detrimental to many researchers. Further researches showed the merits and weaknesses of various components used in both techniques, which empowered the tendency to join those areas in a single frame, together with some other variations of EAs like Evolutionary Programming. While some researches called this distinction artificial and started to use the same term covering all related work, such as Evolutionary Computation in 1991 (Back, 2000), Simulated Evolutionary Optimization or Evolutionary Algorithms.

3.1.2 General frameworks suitable for main variations of EAs

One of main differences between GA and ES terminology is what they call “population”: both models use two set of individuals, offspring and parents, and GA prefer to call offspring for the population ES use this term for parents. On the other hand, they hold same operations with same order: fitness evaluation is followed by selection, then recombination and mutation, as showed in the figure 3.1. . On the other hand, this difference also represents a differentiation to the process, the emphasis on variations at distinct places in the flow, reinsertion and selection as they are commonly named. [39] Like some other implementations, we preferred to use three population in our work: the main, the offspring and the parents population, in order to be able to optimize our algorithm for computational performance in the future works by tuning stack sizes of novel candidate generation and reinsertion, and interfere to critical conditions like selective pressure easierly.

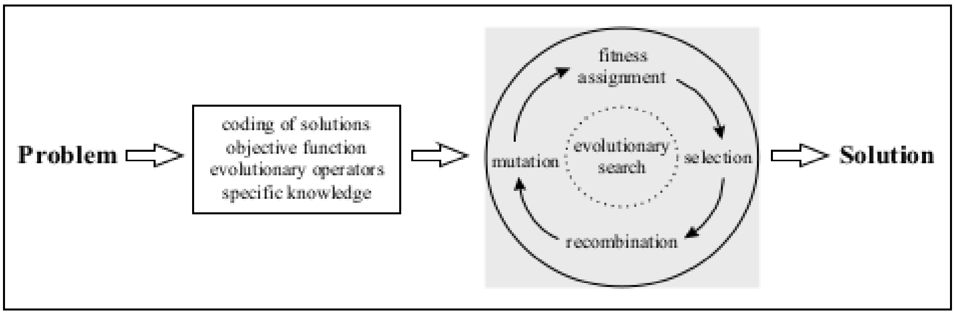

Figure 3.1: Problem solution using evolutionary algorithms [39].

3.1.3 Basic concepts of EA design

First of all, an optimization algorithm should be able to cover all the space, in other words, it should generate any point with a finite probability from another point within a finite number of steps. Even if it does not seem a problem in our case, asymmetric probability distribution of steps create a bias in the search process and controlling the bias is a major subject for also our case. It is already known that EAs could have an uncontrolled structural bias especially for bigger population which can reduce coverage [40], but controlled bias towards more promising areas is a desirable aspect. (Back, 2000) Uncontrolled bias should be limited by correct choices of superparameters and algorithmic components. Moreover, controlled bias may be added by recombination and mutation operators like we tried to do in our implementation.

Second, it should combine convergence towards local optima and stochasticity should be well. Selection pressure, which is the probability of the best individual being selected as parent divided by that of mean fitness individual, is the most important aspect to optimize in this trade-off. Sharper selection would yield faster convergence, but the diversity in the population may be lost and both coverage and stochasticity would be ruined and the search can be stuck because of emerging premature convergence. (Back, 2000) A number of various algorithmic components are designed and used to address these problems, maintaining diversity[41] and preventing premature convergence [42] which are linked closely. Notable ones are crowding in which an offspring can replace only the closest parent, fitness sharing that often introduced by a grouping and mostly clustering routine like in Gator [REF], parameter control like increasing mutation rate when diversity is lost, local selection/reinsertion, island model, novelty search, sexual selection to incest prevention and so on. We needed to control mutation and crossover parameters and restart the population when necessary according to our diversity criteria, later we implemented crowding and island model parallelization for improved diversity.

Also note that fitness function and chromosomal representation of parameters are very important, even if their effect on the performance is not always clear and tuning of those parameters may necessitate expertize (Back, 2000). In our case, chromosomes are a set of real parameters, where a set of six parameters define the pose of a molecule and last three to nine parameters define lattice vectors. Our problem definition posed different challenges as well as advantages over other CSP parametrizations in the literature. One of the important challenges aroses from the fact that the same crystal structue may usually expressed by very distict parameter sets, because of similarities, symmetries and combinations of 6-parameter sets that each represents a molecule of the same kind. Abscence of explicit symmetry information, on the other hand, is mostly facilitating the optimization process by allowing us to work with continuous parameters only.

A different trade-off is about information loss and again population diversity. Good solution can be stochastically lost in non-elitist evolutionary algorithms, and elitism can restain this loss. An extreme example of that is (μ + λ) ES, in which every individual in the main population are elite, which means that they will be in the next population without getting eliminated in a stochastic process even if it is not guarantee that they do reproduce. A small number of elites (usually 1) is used in this algorithm.

3.1.4 Examples for popular and promising variations of EA

Covariance Matrix Adaptation – Evolution Strategies (CMA-ES) is noteworthy not only for its popularity and success in a large number of studies but also for its capacity to implicitly make mutations if an explicit information is used to define more probable mutation orientations. [43] [26]

Estimation of Distribution algorithms target to make more appropriate crossovers. They rely on schema theorem which briefly states that schema destruction rate should be less than schema propagation rate where schema is defined as smallest possible parameter set of the genome that could need and be treated as if the problem is separable. This class of algorithms aim to solve linkage relations among parameters, i.e. how they contribute to the objective value collectively. [44] For the future works of this study, these may be especially promising, as solving linkage problem is naturally very close to protect promising relations among molecules by transferring frequent relative poses in the previous generation to the offspring in a much better rate than uniform or one-point crossover.

On the other hand, Estimation of Distribution Algorithms (EDAs) naturally work with discrete parameters and perform better with smaller alphabets of parameters, in other words, they may need much larger population sizes when the number of possible values a parameter may have is high. This implies that an implementation of EDAs to the CSP problem should solve a discretization problem in the first place. [45] Also note that, they usually does not scale sufficiently good with the number of dimension in the problem (remember that it is one of the main challenges in CSP), although there is special versions made for better scaling. [46] A domain specific solution may be the integration of EDAs to our refined implementation, so using EDA together with intermolecular interaction energy information.

3.2 EAs Specialized for CSP

There is a number of evolutionary algorithms implementation for the CSP problem in the literature. However, for their performance in molecular CSP field, two of them are noteworthy: USPEX [35] and GAtor [36]. Those two algorithms use different techniques to prevent diversity loss (aging for USPEX and fitness sharing for GAtor), different heuristics for crossover and mutation operators, employ different local optimization routines (Molecular Dynamics and BFGS relaxation, respectively) and energy calculation methods (DFT calculation for every candidate in USPEX, and force fields in GAtor) but achieve similar performance in blind tests in some extent. USPEX is a more general CSP tool, actually more focused on inorganic (covalent crystals) CSP, whereas GAtor is more specialized for molecular systems. For the sake of their relatively small performance difference in favor of GAtor, and similarity for the scope of application, GAtor was a more important subject of examination in our pre-studies. On the other hand, detailed examination of these methods is out of the limits and scope of this work.

4. EXPLANATION AND PERFORMANCE OF MCaSP-Evo

4.1 Aims, Motivations and General Structure of MCaSP-Evo

The ultimate aim of the project is to create an efficient optimization tool for CSP of molecular systems. Like many other real world problems that cannot be tackled with general black box optimization tools, problem specific heuristics should be considered in the design of the optimizer. EAs are selected because it is not only an efficient optimization tool that offers a large variety of general purpose operators and algorithmic variations, but also its modular structure provides a unique plasticity that allows users to implement specific and novel operators or algorithmic components.

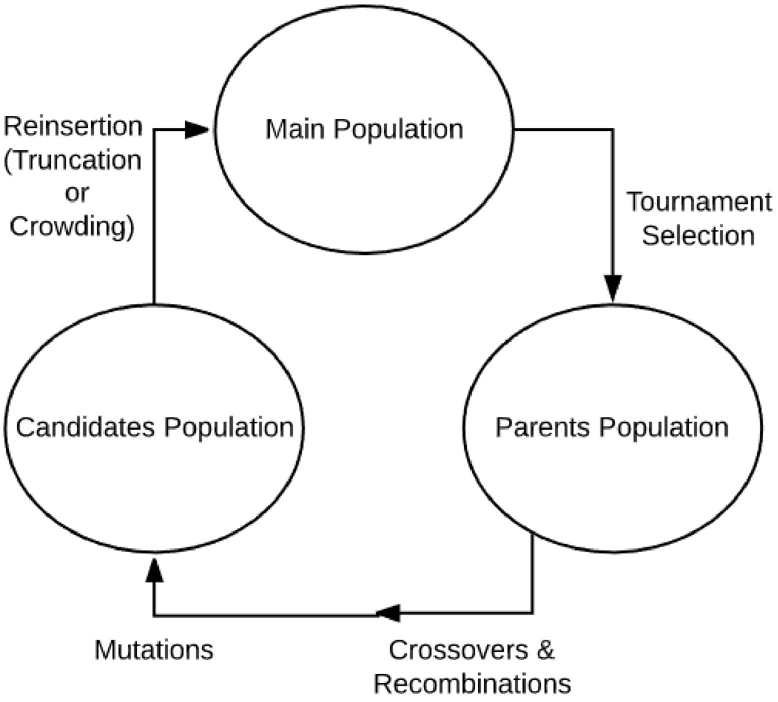

Figure 4.1: Population structure of MCaSP-Evo

Beside its merits, EAs have some drawbacks and deficiencies comparing with some other alternatives and their performance rely on appropriate selection of algorithmic components and adequate tuning of its super-parameters. A number of contradicting concerns should be paid or problems such as premature convergence, loss of diversity or lack of convergence can arise in practice. Fortunately, EA literature contains a great number of offered solutions to address those problems. Unluckily, any systematic approach to diagnose and treat these problems did not prevail yet, thus selection among a large number of possibilities relies on the implementer.

As a result, our effort was guided and shaped by two different objectives. First, we aimed to benefit from present expertise and insight on CSP by creating special operators that make use of problem specific heuristics. These were meant to reduce the unlikelihood of lucky steps to proceed towards better solutions. Second, we need to change and tune EA in order to satisfy performance criteria for an optimization algorithm and overcome structural biases added by problem specific operators for the sake of reliability of our optimizer. These were also claimed to be main goals by the authors of two previously mentioned EAs tailored for CSP, USPEX and GAtor. We set the first goal as the implementation of a reliable GA, to add special operators and promising EA extensions such as covariance matrix adaptation and estimation of distribution.

The general structure of the algorithm is constructed with three population model, namely the main population, parents and offspring. This model differs from traditional GA and ESs population structures that use two populations, parents and offspring, where both traditions do not agree on which list of individuals would be called the population. This population structure provides us an extended flexibility to adjoin various components that would be inconsistent otherwise and ease to tune the algorithm. Even if two domination models continue to dominate the field, alternative three population models are getting more popular as the area proceed to a unified approach from two independent branches of GAs and ESs [39].

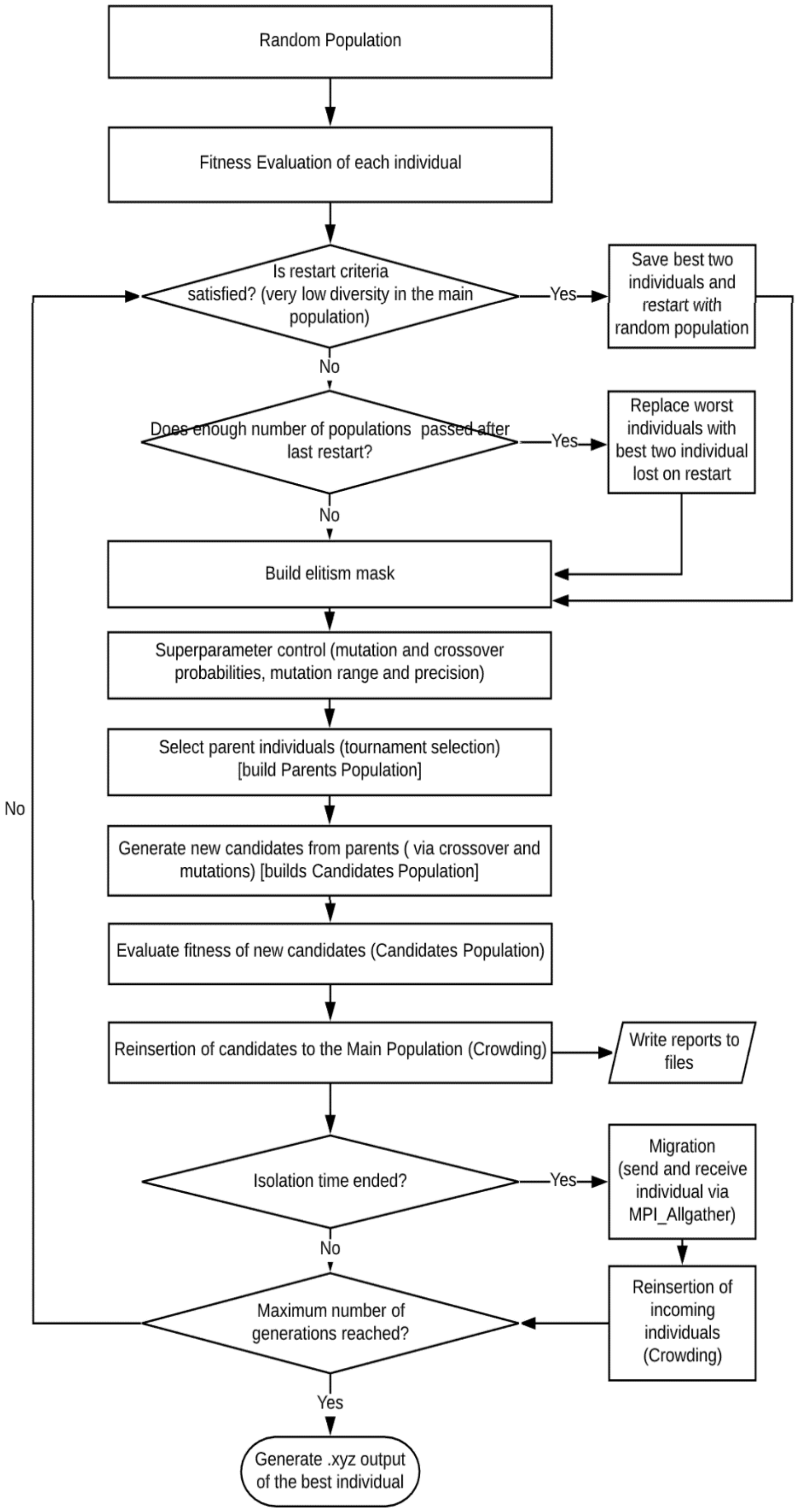
Random Population
Fitness Evaluation of each individual
Is restart criteria satisfied? (very low diversity in the main population)
Yes
Save best two individuals and restart with random population
No
Does enough number of populations passed after last restart?
Yes
Replace worst individuals with best two individual lost on restart
No
Build elitism mask
Superparameter control (mutation and crossover probabilities, mutation range and precision)
Select parent individuals (tournament selection) [build Parents Population]
No
Generate new candidates from parents (via crossover and mutations) [builds Candidates Population]
Evaluate fitness of new candidates (Candidates Population)
Reinsertion of candidates to the Main Population (Crowding)
Write reports to files
Isolation time ended?
Yes
Migration (send and receive individual via MPI_Allgather)
No
Maximum number of generations reached?
Reinsertion of incoming individuals (Crowding)
Yes
Generate .xyz output of the best individual

Figure 4.2: Flowchart of McaSP-Evo

The search space of molecular CSP is exceptionally crinkled, which means that small changes have a very little chance of producing similar or better objective value than parent structure. This complication can be alleviated via either proper local optimizations or global optimization algorithms that intrinsically contain fast convergence to first level local optima such as Monte Carlo simulated annealing or parallel tempering. EAs do not exhibit such a fast local convergence unless very sharp selective pressures are used. On the other hand, high selective pressure often yields to premature convergence around local optima found at an early stage due to the fast diversity loss. This problem is usually tackled by using high selective pressures together with additional elements, such as crowding, fitness sharing (e.g. in GAtor), aging (e.g. in USPEX), increasing mutation rate (e.g. in PIKIA general purpose GA algorithm), restarting population with random individuals except best solution, island model parallelization or pseudo-parallelization with sophisticated migration models, and many more.

In addition to that, almost all of EAs that deal with CSP problem use a local optimization procedure for every new candidate solution to lower dimensionality of the problem, in other words to reduce the number of local optimization stages that global optimizer have to deal with. The usage of local optimizers also procures a softer and less wavy fitness surface to the global optimizer. Local optimizers used in the field are either general purpose local optimizers such as BFGS implemented using the derivative of the force fields or numeric derivatives, or problem specific procedures are employed for that reason such as relaxation via molecular dynamics. On the other hand, local optimization process itself has a considerable cost and using that for every new feasible candidate increases the cost of whole process dramatically as it is repeated very high number of times. Furthermore, the derivative of the implemented force fields is not always reliable, which forces us to use more stable local optimization routines that have even higher cost. Thus, we decided to implement an EA that works without a local optimizer for making it less dependent on local optimization and see how far we can go in order to attach a cheapest possible local optimization routine in a further step of development when a further improvement without local optimization cannot be achieved anymore.

The absence of a local optimization makes the computational costs affordable for us, but also makes the design process of a reliable EA harder. We witnessed an exceptional tendency to premature convergence with the simple GA, due to higher levels of selective pressure is needed. We implemented some superparameter control techniques, i.e. lowered crossover rate with time and increased mutation rate when median of the population gets closer to the best in terms of objective value, i.e. more or less similar to PIKIA algorithm. Many different variations of superparameter control are tested and we achieved best result on practical problems with current schema.

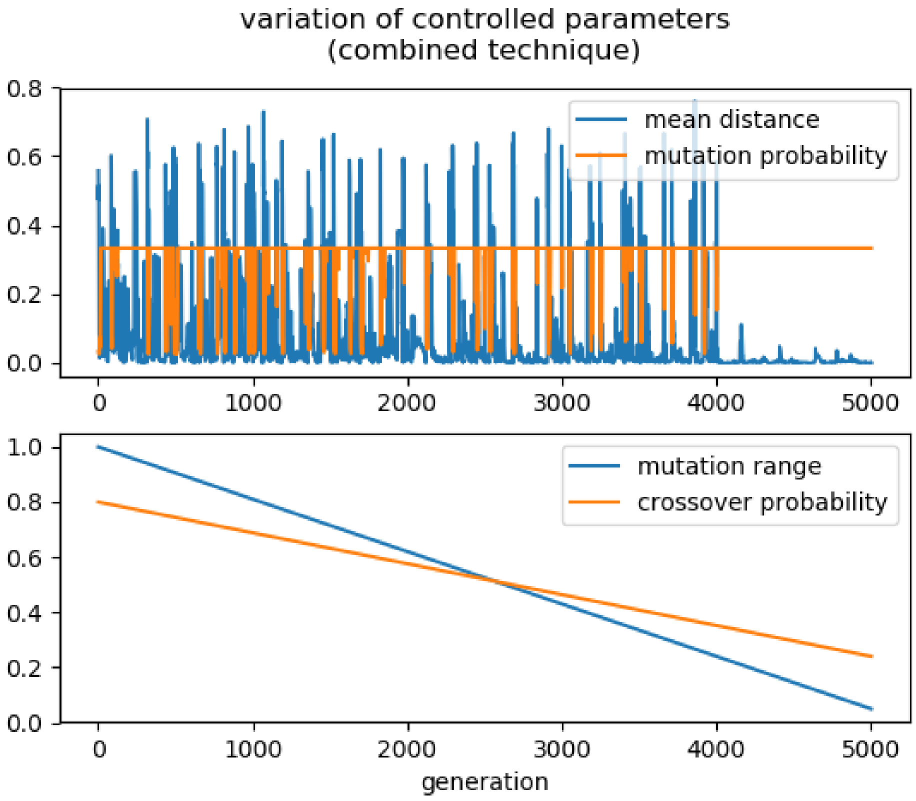

Figure 4.3: Variation of superparameters in a run of EA using superparameter control and restarts combined.

Using mutation rate change and restarts together,mutation probability fastly oscillates between its limit values as it may be seen in the figure 4.3. On the other hand, corssover probabily and mutation range vanishes in a constant and slow manner l,ra alternatif-Superparameter control improved the performance at a point which may be successful for very little systems, such as CSP with two or three molecules in a single lattice. But the success was not stable, and the performance was not sufficiently reliable to go further. Investigation over the process of optimization revealed that the fast diversity loss that appears in many runs is due to the dominance of an individual and it happens when one single good local optimum could be found in the early stages. This makes the optimization less global and more local and causes it to get stuck around a single or a group of closely located local optimum. Our solution was restarting the population with random individuals using current superparameters, like microGA technique. But out implementation differs from usual microGA restarts in the sense that we do not let the best solution in the population. Instead, we resurrect the best and second best solutions after a certain number of generations passed hopefully to find other local minima with similar fitness that reduce the risk of premature convergence. An important part of this implementation was the restart decision criteria.. We decided to restart if 90% of individuals are almost copies of the best individual. The criteria for being an almost copy is to have more than 95% of their genes less different than 0.1% of parameter interval than the best individual’s corresponding parameter. This criteria is developed after our trials with standard deviation and distance from the best did not produce satisfactory results. Reinsertion or resurrection of the best individual happens 150 generation later than restart.

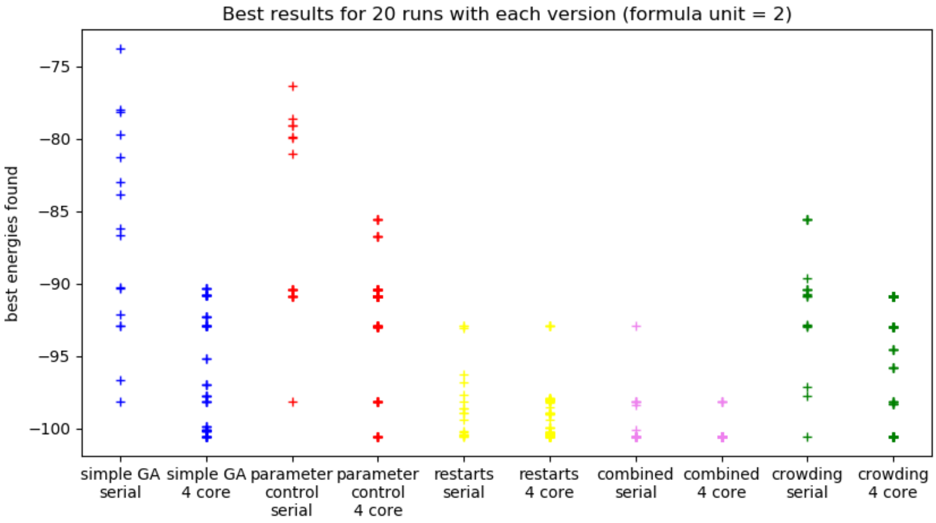

Figure 4.4: Comparison of different versions of EA by best results.

In the figure 4.5, it is clear that all the changes improved performance of our EA, except crowding. We already chosed crowding for its potential of polymorph discovety, so it is not discouraging by its own.

We further reduced the diversity loss trend by replacing global reinsertion which works in a similar fashion with truncation selection between new candidates and the main population by nitching techniques. To avoid potentially high computational cost of clustering algorithms for high population sizes, we chose to use a rather simple but flexible crowding technique. Within that reinsertion method, we calculate a modified

Euclidean distance between new candidate structure and those in the main population to replace the closest one with the new candidate if it has a lower energy. Modified Euclidean distance is calculated by weighted summation of distances between molecular poses in two compared structures and their lattice parameters. Each molecule is represented by 6 parameters, where first 3 imply the position of center of mass. Distance between two molecules in two different structures is defined as Euclidean norm of differences between parameters of them. Structural distance between two structures is calculated as sum of molecular distances after matching the molecules with the unmatched molecule in the other structure closest to itself in terms of only position. We aimed the lattice distance to be as effective as half of the structural distance, like GAtor. Similar to arbitrary heuristic used in some other algorithms such as GAtor, so we multiply lattice distance by the number of molecules and summation of the distances is taken as modified Euclidean distance.

For the parallelization, we used the island model which is known for yielding even better performance for the quality of solutions than serial implementation. A probable reason of this eligibility is that migration helps the subpopulations to break diversity loss and premature convergence that can also happen in isolation times. This effect is inspired from a discuss in evolutionary biology field that leaded to development of punctuated equilibrium hypothesis of Stephan Jay Gould. He argues that large evolutionary changes often happens in relatively short periods among small populations and large unified populations stay stagnant as it is observed in the fossil records. (Back, 2000) In this model, populations go through an isolation time in which they act as independent serial EA runs and a migration process in handled among them at the end of each isolation time. In our work, migration model is selected as an all-to-all communication among all populations sending the best 2 individuals, and reinsertion of migrated individuals are handled by crowding.

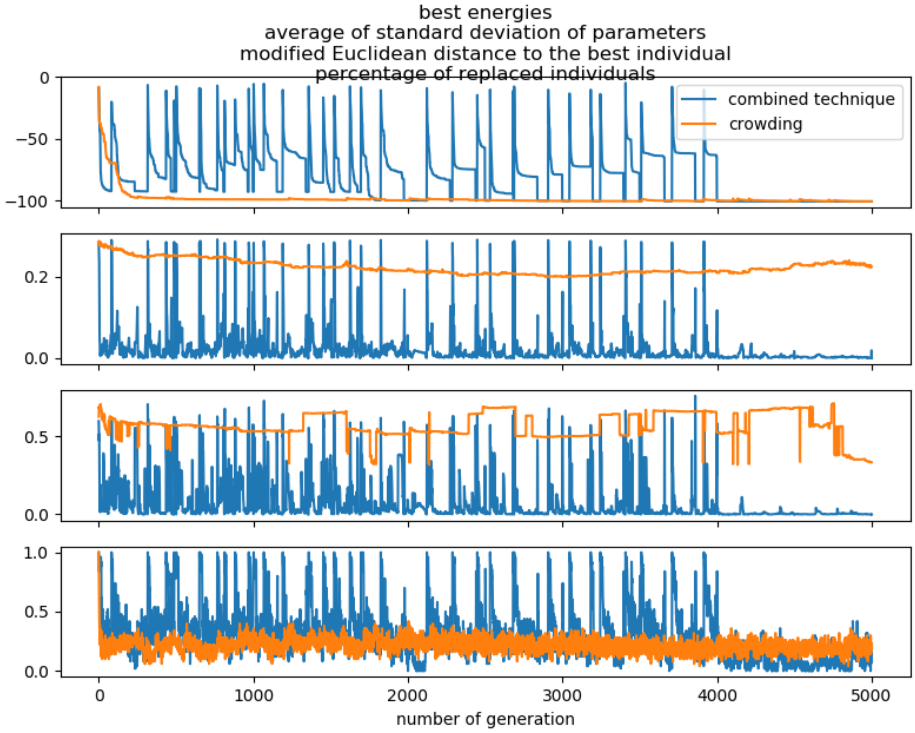

Figure 4.5: Comparison of the last two versions by best energies, average of standard deviation of parameters, modified Euclidian distance to the best individual and percentage of replaced individuals respectevly.

The figure 4.5 provides precious data for comparison of two last versions: crowding version and the other which combines parameter control and restarts. It shows, from top to bottom, best energies in the seach process, average standard deviation of parameters and average modified Euclidean distance to the current best solution, and the ratio of reinserted or replaced individuals. First two provides information about how better solutions are found: crowding shows a more steady search process and combined approach increases its chance by restarts when diversity in the population vanishes. Second and third subplots clearly indicate the vanishing diversity loss caused by truncation reinsertion, and success of crowding over truncation for preserving diversity.And the last one shows replacement rates, which is and important hint for how often the algorithm produce and select promising now solutions.

The problem specific operators are also a very important part of the EA design. Including GAtor and USPEX, many successful implementation make use of heuristics or information about the problem they tackle by problem specific operators. Our problem specific mutation operator is based on an anticipation about coordinate systems used in the mutation operators. We previously mentioned the performance gap between Cartesian coordinates and internal coordinates in the molecular configuration optimization, which is about 7 times for a rather small molecule and scales with the problem size, in the section 1.2.1. This impressive contrast originates from the difference of likelihood of lucky steps: small changes in one parameter in a system expressed in Cartesian coordinates is unlikely to be sufficiently close to a progressive direction in contrary to internal coordinates or other coordinate systems where parameters are directly related to physical aspects that govern the interaction, such as distance between atoms or bond angles. A similar performance increase would not be surprising between small steps taken in Cartesian coordinates and a more reliable coordinate system. We propose to make use of relative poses, i.e. the relative position and orientation of a molecule in the reference frame of another instead of making mutations by changing individual parameters in Cartesian coordinate system. This would also eliminate a structural bias of the mutation operators. Mutations in Cartesian coordinates would be less likely to produce a better candidate in some parts of the search space comparing with other parts, because the angle between the axis and the direction of change of a potentially related physical property like relative distance of two molecules is dependent on the orientation of the reference molecule. For that purpose, mutations are made on the relative poses of a selected molecule with respect to the selected reference frame which is another molecule in this case, so that likelihood of lucky steps become independent of the orientation of the reference molecule and hopefully increased. Figure 4.7 shows a flowchart that explains general structure of our mutation operator, and Figure 4.8 shows an example of crystal structures before and after special mutation.

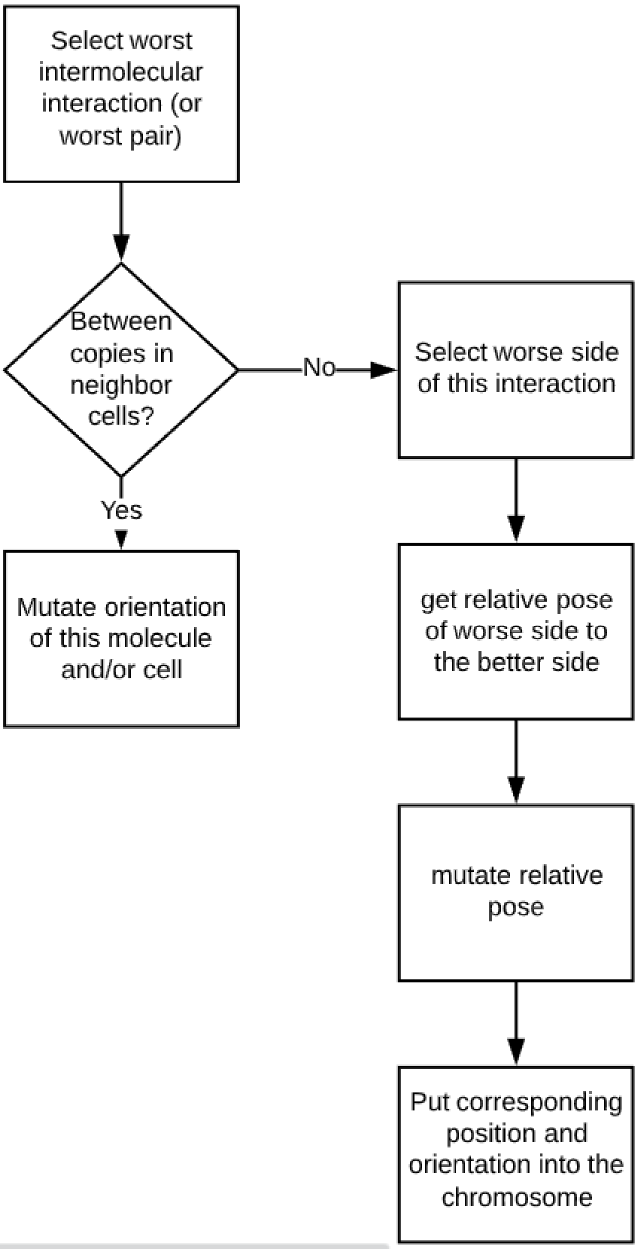
Select worst intermolecular interaction (or worst pair)
Between copies in neighbor cells?
No
Yes
Select worse side of this interaction
Mutate orientation of this molecule and/or cell
get relative pose of worse side to the better side
mutate relative pose
Put corresponding position and orientation into the chromosome

Figure 4.6: Flowchart of special mutation

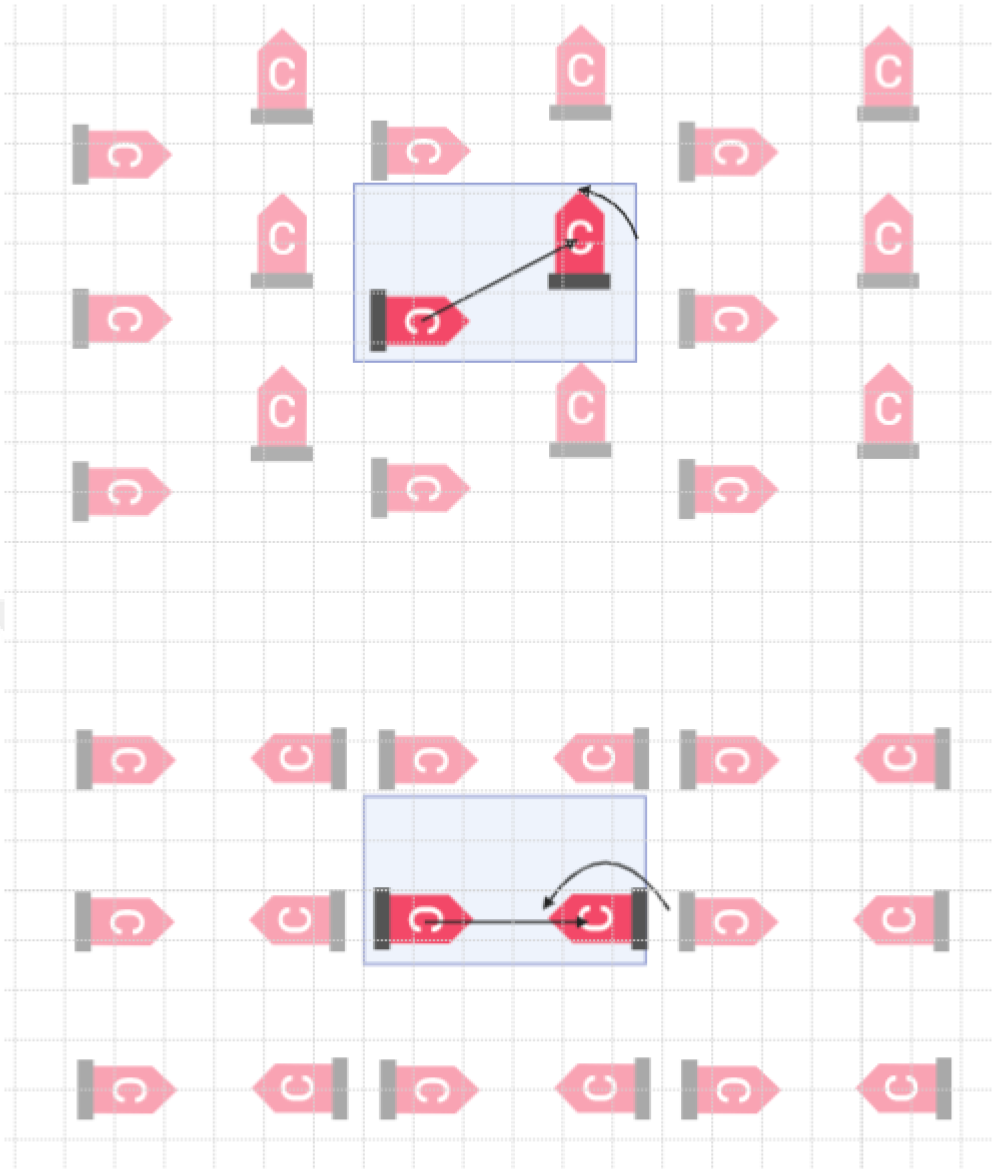

Figure 4.7: Simplified crystal structures demonstrating a system before (upper figure) and after (lower) special mutation.

In this simplified example, the interaction with highest energy contribution is between first and second molecules, and second molecule which is on upper left of the lattice is the molecule with higher total energy contribution among two sides of that worst interaction. Thus, relative position and orientation of the second molecule with reference to the first is mutated. This relative position and orientation before and after mutation are represented with black arrows in the figures.

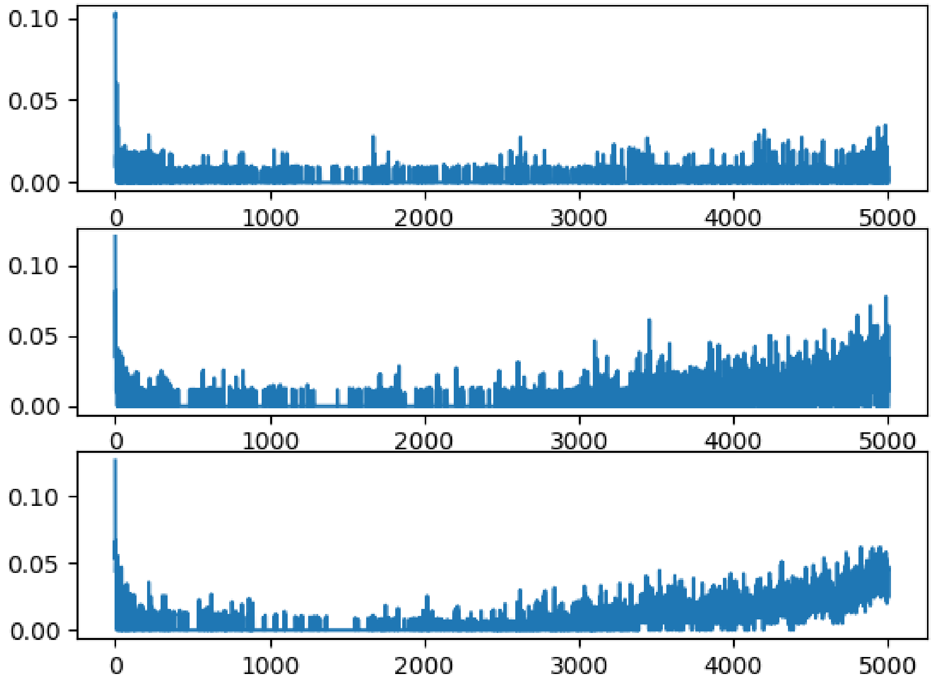

Figure 4.8: Success rate of mutation operations: uniform mutation, exponentially scaled mutation, our special mutation operator, respectively(population size is 50 for main population, 500 for candidate population).

Second, we tried to put the information about the molecular interactions in service. Our potential energy calculation method calculates total energy as a sum of individual contributions of molecule pairs. By passing these contributions from energy calculation module to EA, we made use of the simple heuristic that it is a better idea to mutate molecules that have worst interactions. For that, a molecular interaction is chosen with a probability proportional to the related energy contribution, and relative pose of the molecule that has worse total energy contribution with respect to the other side of chosen interaction. The mutation probability of these mutation operators will be controlled such that first one is more likely to happen when the median of the population gets close to the best in terms of total energy and the probability of the latter is inversely proportional to the diversity of substructures in the population.

Figure 4.8 shows succes rate of mutation operators. Success rate is calculated as the ratio of the number of mutants reinserted to the main population and the number of mutants in the candidates population. The outcome of all mutation operators is low, as well as crossover operators, because of two main reason: the fitness surface is rugged and rough and we lack locak optimization to deal with it for every individual in the offspring. Here, first two are general purpose mutation operators: first is the uniform mutation, that changes selected parameter(s) by adding it a scaled random number from uniform distribution between zero and mutation range. Second one is widely used to increase low level local optimization ability of optimizer for real parameter problems. It makes use a random number as a negative power of two, so it produces smaller steps much more probably. The third one is our special mutation operator.

The underlying concept of our special crossover and mutation operators, whose some are still in design process, is to protect better interactions by creating corresponding relative pose in the offspring. We name groups of molecules that have favorable internal energy as substructures. The substructure concept is the logical basis of our problem specific recombination and crossover operators. Crossover is claimed to be the most important part of a genetic algorithm, as it defines the information flow among candidate solutions. These were an important part of the examined examples of EAs specialized to CSP, namely GAtor and USPEX. As a reminder, GAtor make use of symmetry information to derive a problem specific crossover operator, named “symmetric crossover”. Heredity operator, used in USPEX and many other algorithms in the field, tries to preserve the essential information of parent structures, which is defined as relative position of nearby atoms by creators of the algorithm, by slicing two parent structure and combining slabs of each parent. Our method is more similar to that of USPEX in sense of preserving relative position information, but it differs from it as we do not select an arbitrary spatial region but favorable intermolecular poses that have better interaction energies than most of the others. The employed heuristic can be expressed so that it would be more probable to obtain good novel candidates when better parts in sense of internal interactions of two parents are combined than a combination of more or less random parts of parent solutions. As a part of future works, it is planned to implement EDAs (Estimation of Distribution Algorithms) to detect and evaluate substructures in order to reduce or eliminate the tendency of oversampling a group of molecules that interacts well with each other but badly with most of or any other molecule. In other words, the underlying heuristic may become the frequency or change in the frequency of occurrence of a partial solution in the population after the selection in the future. In order to add stochasticity to the process to overcome a structural bias towards combinations of small molecule groups such as dimers, crossover operators are intentionally designed and the aforementioned mutation operator that seek to break substructures is added and a special crowding scheme seeking the variety of substructures will substitute for the current simpler crowding.

We call the selected group of molecules a substructure and their employment seems promising for a number of reasons. First, it is similar to divide and conquer approach in the sense that discovery of substructures and their combinations is being preferred over direct search of the full structure. A limitation is to emphasize for divide and conquer approaches in general: we usually cannot asume that the porblem in hand is seperable, including CSP. The fact that CSP for larger systems does not always result in the repetitions of the crystal structures found with smaller molecule counts implies inseparability of the CSP problem, which makes the use of divide and conquer methods questionable. On the other hand, the noteworthy success of EDAs whose the main mathematical proof relies on separability of the solution (see schema theorem) over a considerable class of inseparable problems shows that this major concern is not enough to invalidate all the methods that seek for partial solutions at least those which do not explicitly need separability. Our method does not assume that CSP can be obtained as a combination of crystal substructure discoveries, instead it employs related operators in order to reduce complexity of problem by oversampling more promising areas of the search space. The second inspiration is already mentioned: the success of EDAs that relies on building block notion which is similar to our substructure concept. Third, determination of frequent substructures provides a well exploited guide for discovery of novel substances [47] and this functionality may be repeated in the case of crystal structures.

Moreover, the computational complexity of the problem can be further reduced in the future by employment of special operators inspired or adapted from fairly large docking and self-assembly literature for combining substructures, as these problems have many aspects similar to this challenge. Now, it seems the least studied and critical part of the algorithm, and the part that would benefit from local optimization most.

Lastly, constructive heuristics [26] may be insufficient for modern CSP challenges for which even local optimization approaches does not seem promising. But they are often faster than other methods and may be adapted into global optimization procedures in certain examples. Discovery and recombination -or exploration and exploitation- of substructures can be seen as a constructive heuristic for CSP and may be functional for speeding up the global optimization process by stochastic implementations and additional precautions are taken for overcoming potential biases.

4.2 Performance of MCaSP-Evo

4.2.1 Test case

We tested our algorithm to find the crystal structure of cytosine . Cytosine, is one of the nucleobases found in nucleic acids , forms anhydrous crystals. In 1964, Barker and Marsh determined the first crystal structure of cytosine using photographic intensity data.[48] They found that cytosine crystallizes into orthorhombic P2 1 2 1 2 1 space group with lattice parameters a=13.041, b=9.494 and c=3.815 Å and Z=4. The same crystal structure has been reproduced by McClure and Craven in 1973. [49] A new polymorph of cytosine (orthorhombic Pccn with a=15.104, b=15.1212 and c=9.2948 Å and Z=16) recently have been reported by Sridhar et al. [50] This findings, for the first time, shows that polymorphism is possible in nucleobases. Along with polymorphism, nucleobases are in our interest area for their potential semiconductor properties [51] and interactions with metals.

4.2.2 Results

We could not achieve to predict neither the polymorph discovered 4 years ago nor the other, but we are able to successfully predict theoretical crystal structure for 2 formula units, i.e. 2 molecules in a single cell.

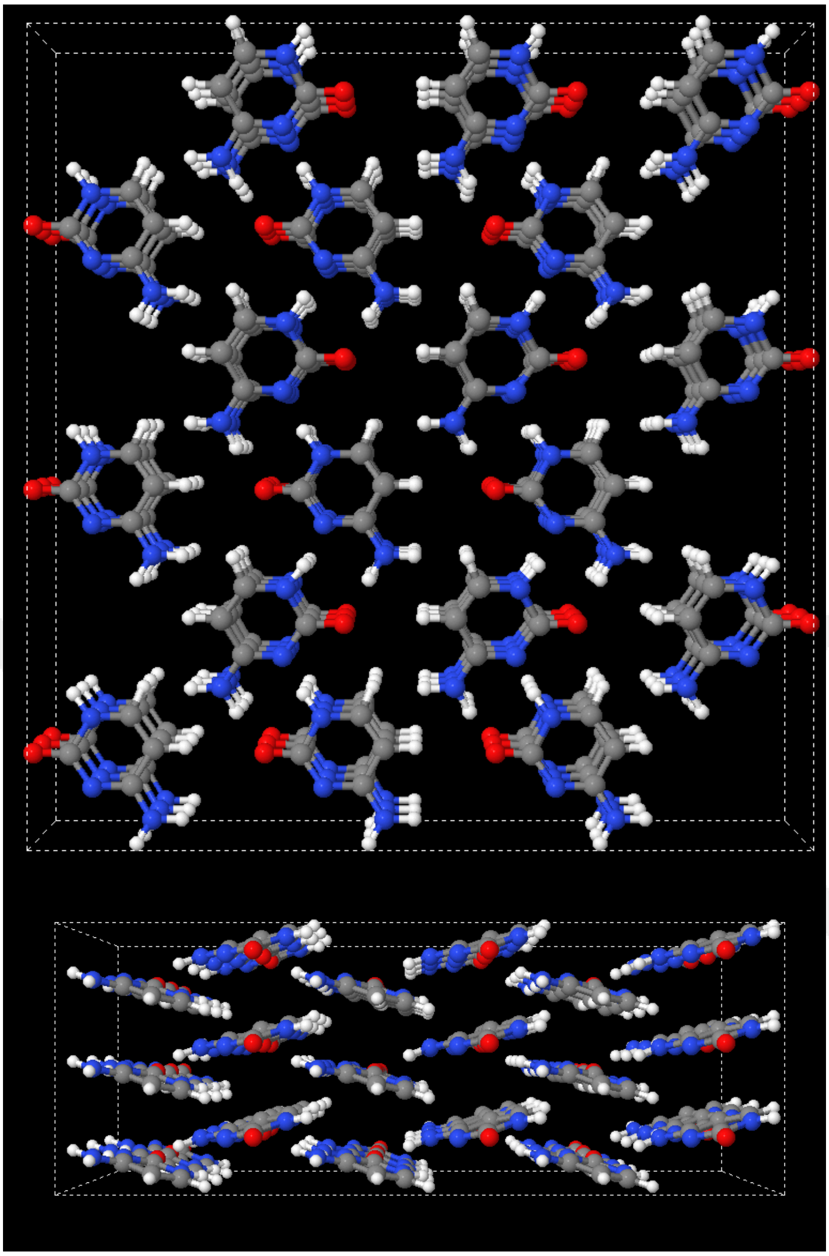

Figure 4.9 : Two explanatory views of the computational predicted crystal structure of cytosine with Z=2 showing structural conformation.

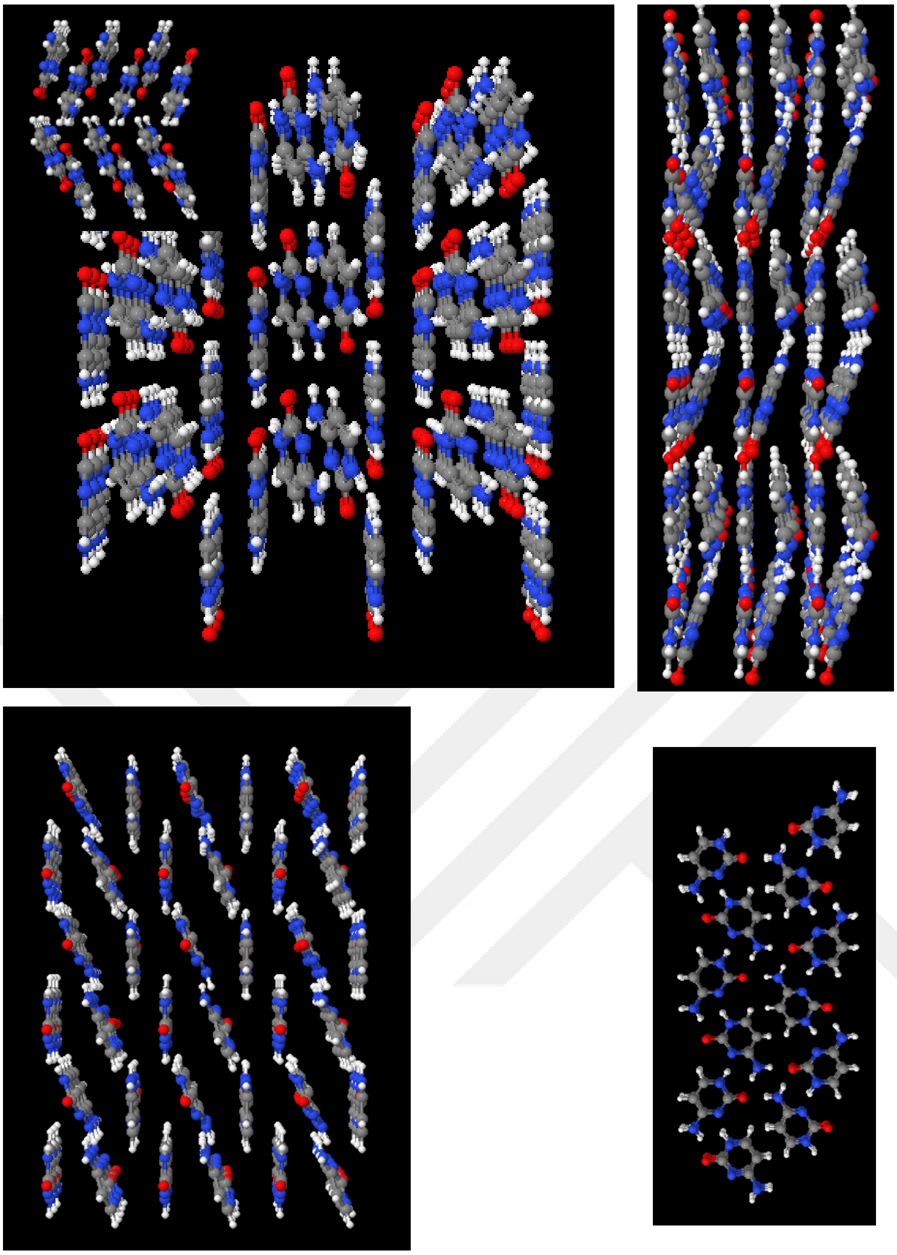

Figure 4.10: Various crystal structures found at the end of runs.

4.3 Conclusion

With crowding, our EA searches many diverse points of the search space simultaneously and contains many different structures at the same time. This diversity in the population is promising for both our aim of discovering polymorphs and metastable phases, and tyring our novel crossover operators in proper conditions in the future.

As conclusion, we can say that we reached one of our goals by implementing a population based optimization algorithm that preserves diversity in the population. At this point, we can try many insightful heuristics for generating good candidate solutions based on information implicitly stored in a group of candidate solutions on a proper infrastructure.

On the other hand, our problem specific mutation operator showed the potential of this research direction. The rate of replacement for the offspring created via special mutation operator was increasing or stable for the test runs. The similar performance of exponentially scaled mutation operator seems to be making this remark suspicious, but we should consider the meaning of exponentially decreasing step size and success rate which is inversely proportional to the step size. A potential explanation for the performance of exponentially scaled mutation may be higher replacement rates by contributing to the fine tuning of the found crystal structures instead of helping to find different crystal structures. In that case, this operator would not be considered as a competitor for our special mutation operator for their dissimilar contributions to the population. The increase of the success rate of the whole optimization process after adding special mutation can be a clue for this kind of different functions of this two mutation operator, nevertheless further investigation and analysis is required to be sure of the functionality of the special mutation operator among other mutation operators.

The performance of MCaSP-Evo is not satisfying comparing with other CSP optimization tools of our research group at the present moment. Nevertheless, the advancements in the development process show the potential of the underlying ideas and promise a more modifiable optimization engine that achieves similar or better performance in the close future.

REFERENCES

Back, T., Fogel, D. B., & Michalewicz, Z. (Eds.). (2000). *Evolutionary Computation 1: basic algorithms and operators*. Bristol and Philadelphia: Institute of Physics Publishing.

Back, T., Fogel, D. B., & Michalewicz, Z. (Eds.). (2000). *Evolutionary Computation 2: Advanced algorithms and operators*. Bristol and Philadelphia: Institute of Physics Publishing.

Chipot, C., & Pohorille, A. (2007). *Free Energy Calculations Theory and Applications in Chemistry and Biology*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Hartke, B. (2011). Global optimization. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *1*(6), 879–887. doi: 10.1002/wcms.70

Leszczynski, J. (2012). *Handbook of computational chemistry*. Dordrecht: Springer.

Mikosch, T. V., J., W. S., & Jorge, N. (2006). *Numerical Optimization*. New York, NY: Springer New York.

Oganov, A. R. (2011). *Modern methods of crystal structure prediction*. Weinheim, Germany: Wiley-VCH.

Törn Aimo, & Zilinskas, A. G. (1989). *Global optimization*. Berlin: Springer-Verlag.

Young, D. C. (2001). *Computational chemistry: a practical guide for applying techniques to real world problems*. New York: Wiley.

[1] **Introduction: Crystal Structure Prediction, a Formidable Problem. (2010)**. *Modern Methods of Crystal Structure Prediction*, 11–21. doi: 10.1002/9783527632831.ch

[2] **Maddox, J. (1988)** *Crystals from first principles*. Nature, 335, 201.

[3] **Groom, C. R., & Reilly, A. M. (2014).** Sixth blind test of organic crystal-structure prediction methods. Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials, 70(4), 776–777. doi: 10.1107/s2052520614015923

[4] **Appendix: First Blind Test of Inorganic Crystal Structure Prediction Methods. (2010)**. *Modern Methods of Crystal Structure Prediction*, 223–231. doi: 10.1002/9783527632831.app1

[5] **Meredig, B., Agrawal, A., Kirklin, S., Saal, J. E., Doak, J. W., Thompson, A., Wolverton, C. (2014)**. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*, *89*(9). doi: 10.1103/physrevb.89.094104

[6] **Ryan, K., Lengyel, J., & Shatruk, M. (2018)**. Crystal Structure Prediction via Deep Learning. *Journal of the American Chemical Society*, *140*(32), 10158–10168. doi: 10.1021/jacs.8b03913

[7] **Podryabinkin, E. V., Tikhonov, E. V., Shapeev, A. V., & Oganov, A. R. (2019)**. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Physical Review B*, *99*(6). doi: 10.1103/physrevb.99.064114

[8] **Bhadeshia, H. (1987)**. *Worked examples in the geometry of crystals*. London: the Institute of Metals.

[9] **Addis, B., & Locatelli, M. (2006)**. A new class of test functions for global optimization. *Journal of Global Optimization*, *38*(3), 479–501. doi: 10.1007/s10898-006-9099-8

[10] **Chiong, R. (Ed.). (2009)**. *Nature-inspired algorithms for optimisation*. Berlin: Springer Verlag.

[11] **Piela, Lucjan. (2015)**. Computational Chemistry: From the Hydrogen Molecule to Nanostructures. 10.1007/978-94-007-6169-8_1-2.

[12] **Görbitz, CH. (ed.), Reilly, AM., Cooper, R.I., Adjiman, C.S., Groom, C. R. (2016)**. Report on the sixth blind test of organic crystal structure prediction methods, *Acta Cryst B72, 439-459,* https://doi.org/10.1107/S2052520616007447

[13] **Christ, C. D., Mark, A. E., & Gunsteren, W. F. V. (2009)**. Basic ingredients of free energy calculations: A review. *Journal of Computational Chemistry*. doi: 10.1002/jcc.21450

[14] **Zipse, H. (2019)**, *Computational Chemistry 1, comparative examples web notes* [website notes as html]. Retrieved from https://www.cup.lmu.de/oc/zipse/teaching/computational-chemistry-1/topics/comparative-examples/

[15] Ponder, J. (2019), *chem430, Lecture 20: Density Functional Theory, class notes* [pdf]. Retrieved from https://dasher.wustl.edu/chem430/lectures/lecture-20.pdf

[16] Burke, K. (2012). Perspective on density functional theory. *The Journal of Chemical Physics, 136(15), 150901*. doi: 10.1063/1.4704546

[17] Leach, A. R. (2009). *Molecular modelling: principles and applications*. Harlow: Pearson Prentice Hall.

[18] Manukyan, A., & Tekin, A. (2015). First principles potential for the cytosine dimer. *Physical Chemistry Chemical Physics*, *17*(22), 14685–14701. doi: 10.1039/c5cp00553a

[19] Manukyan, A., & Tekin, A. (2017). The intermolecular dimer potential for guanine. *The Journal of Chemical Physics*, *147*(15), 154311. doi: 10.1063/1.4998792

[20] C. Leforestier, A. Tekin, G. Jansen, and M. Herman, (2011) "First principles potential for the acetylene dimer and refinement by fitting to experiments," J. Chem. Phys. 135, 234306.

[21] Hartke B. (1998) Global geometry optimization of small silicon clusters at the level of density functional theory. Theor Chem Acc, 99:241–247.

[22] Podryabinkin, E. V., & Shapeev, A. V. (2017). Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, *140*, 171–180. doi: 10.1016/j.commatsci.2017.08.031

[23] Meredig, B., Agrawal, A., Kirklin, S., … Wolverton, C. (2014). Combinatorial screening for new materials in unconstrained composition space with machine learning. *PhysRevB, 89, 9*. doi: 10.1103/PhysRevB.89.094104

[24] Wolpert, D.H., Macready, W.G. (1997), *No Free Lunch Theorems for Optimization*, IEEE Transactions on Evolutionary Computation 1, 67.

[25] Demir, S. (2020), *Machine Learning Assisted Massively Parallel Crystal Structure Prediction*, (Unpublished doctoral dissertation), Istanbul Technical University, Istanbul, Turkey

[26] Alba, E. (2005). *Parallel Metaheuristics: A New Class of Algorithms*. John Wiley & Sons.

[27] Alba, E. (2009). *Optimization techniques for solving complex problems*. Hoboken, NJ: Wiley.

[28] E.G., Talbi. (2002) *A Taxonomy of Hybrid Metaheuristics*. Journal of Heuristics, 8(5): 54 1-564

[29] Yang, X.-S. (2011). Metaheuristic Optimization: Algorithm Analysis and Open Problems. *Experimental Algorithms Lecture Notes in Computer Science*, 21–32. doi: 10.1007/978-3-642-20662-7_2

[30] Steer, K.C.B., Wirth, A. Halgamuge, S.K. (2009) The Rationale Behind Seeking Inspiration from Nature. In Chiong, R. (Ed.). *Nature-inspired algorithms for optimisation*. Berlin: Springer Verlag.

[31] Sörensen, K. (2013). Metaheuristics-the metaphor exposed. *International Transactions in Operational Research*, *22*(1), 3–18. doi: 10.1111/itor.12001

[32] Alanis, A. Y., Arana-Daniel, N., & López-Franco, C. (2018). Bio-inspired Algorithms. *Bio-Inspired Algorithms for Engineering*, 1–14. doi: 10.1016/b978-0-12-813788-8.00001-9

[33] Avendaño-Franco, G., & Romero, A. H. (2016). Firefly Algorithm for Structural Search. *Journal of Chemical Theory and Computation*, *12*(7), 3416–3428. doi: 10.1021/acs.jctc.5b01157

[34] Ma, Y., L, J., & Wang, Y. (2015). CALYPSO structure prediction method. *Chinese Science Bulletin*, *60*(27), 2580–2587. doi: 10.1360/n972015-00575

[35] Glass, C. W., Oganov, A. R., & Hansen, N. (2006). USPEX—Evolutionary crystal structure prediction. *Computer Physics Communications*, *175*(11-12), 713–720. doi: 10.1016/j.cpc.2006.07.020

[36] Curtis, F., Li, X., Rose, T., Vázquez-Mayagoitia, Á., Bhattacharya, S., Ghiringhelli, L. M., & Marom, N. (2018). GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *Journal of Chemical Theory and Computation*, *14*(4), 2246–2264. doi: 10.1021/acs.jctc.7b01152

[37] Ryan, K., Lengyel, J., & Shatruk, M. (2018). Crystal Structure Prediction via Deep Learning. *Journal of the American Chemical Society*, *140*(32), 10158–10168. doi: 10.1021/jacs.8b03913

[38] Ziletti, A., Kumar, D., Scheffler, M., & Ghiringhelli, L. M. (2018). Insightful classification of crystal structures using deep learning. *Nature Communications*, *9*(1). doi: 10.1038/s41467-018-05169-6

[39] Pohlheim, H. (2006, December). Retrieved from http://www.geatbx.com/docu/

[40] Kononova, A. V., Corne, D. W., Wilde, P. D., Shneer, V., & Caraffini, F. (2015). Structural bias in population-based algorithms. *Information Sciences*, *298*, 468–490. doi: 10.1016/j.ins.2014.11.035

[41] Bhattacharya, M. (2016). Evolutionary Landscape and Management of Population Diversity. *Combinations of Intelligent Methods and Applications Smart Innovation, Systems and Technologies*, 1–18. doi: 10.1007/978-3-319-26860-6_1

[42] Pandey, H. M., Chaudhary, A., & Mehrotra, D. (2014). A comparative review of approaches to prevent premature convergence in GA. *Applied Soft Computing*, *24*, 1047–1077. doi: 10.1016/j.asoc.2014.08.025

[43] Hansen, N. (2016). The CMA Evolution Strategy: A Comparing Review. *Towards a New Evolutionary Computation Studies in Fuzziness and Soft Computing*, 75–102. doi: 10.1007/3-540-32494-1_4

[44] Hauschild, M., & Pelikan, M. (2011). An introduction and survey of estimation of distribution algorithms. *Swarm and Evolutionary Computation*, *1*(3), 111–128. doi: 10.1016/j.swevo.2011.08.003

[45] Chen, C.-H., & Chen, Y.-P. (2014). Quality Analysis of Discretization Methods for Estimation of Distribution Algorithms. *IEICE Transactions on Information and Systems*, *E97.D*(5), 1312–1323. doi: 10.1587/transinf.e97.d.1312

[46] Wang, Y., Li, B., (2009). A Self-adaptive Mixed Distribution Based Uni-variate Estimation of Distribution Algorithm for Large Scale Global Optimization. In Chiong, R. (Ed.). *Nature-inspired algorithms for optimisation*. Berlin: Springer Verlag.

[47] Yan, X., & Han, J. (2006). Discovery of Frequent Substructures. *Mining Graph Data*, 97–115. doi: 10.1002/9780470073049.ch5

[48] Barker, D. L.; Marsh, R. E. (1964) Acta Crystallographica, 17, 1581–1587.

[49] McClure, R. J.; Craven, B. M. **(1973)** Acta Crystallographica Section B, 29, 1234–1238.

[50] Sridhar, B.; Nanubolu, J. B.; Ravikumar, K. (2015) Acta Crystallographica Section C, 71,128–135.

[51] Maia, F. F., Freire, V. N., Caetano, E. W. S., Azevedo, D. L., Sales, F. A. M., & Albuquerque, E. L. (2011). Anhydrous crystals of DNA bases are wide gap semiconductors. *The Journal of Chemical Physics*, *134*(17), 175101. doi: 10.1063/1.3584680

CURRICULUM VITAE

Name Surname **:** Denizhan Tutar

Place and Date of Birth: Üsküdar 18/10/1992

E-Mail **:** tutard@itu.edu.tr

EDUCATION:

- **B.Sc. :** 2016, Istanbul Technical University, Faculty of Mechanical Engineering, Mechanical Engineering.
- **M.Sc.:** Present, Istanbul Technical University, Informatics Institute, Computational Science and Engineering.