Multi-modal memetic framework for locating saddle points, with application to biomolecular systems : water clusters and ring-deficient covalently-bonded small molecules

Mostafa Ellabaan

2013

Mostafa Ellabaan. (2013). Multi-modal memetic framework for locating saddle points, with application to biomolecular systems : water clusters and ring-deficient covalently-bonded small molecules. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/54820>

<https://doi.org/10.32657/10356/54820>

**Multi-Modal Memetic Framework for Locating Saddle Points,
with Application to Biomolecular Systems:
Water Clusters and Ring-deficient Covalently-bonded Small Molecules**

A thesis submitted to
School of Computer Engineering
Nanyang Technological University

by

Mostafa Ellabaan (G0702650D)

in Partial Fulfillment of the Requirement for the Degree of
Doctor of Philosophy (Ph.D)

2013

Abstract

Saddle points have important applications in different fields of science and engineering, including transition states identification, path-planning and robotics navigation. Identifying such points using memetic approaches is a field, remained yet to be well-investigated. In this dissertation, we propose a multi-modal memetic framework, composing of 1) a fitness function that transforms the problem landscape to favor saddle points instead of local or global optima, 2) evolutionary operators of crossover and mutation for efficient exploration of the problem landscape, 3) niching operator based on the valley-adaptive clearing scheme for maintaining multiple saddle points during the evolutionary process, and 4) memetic operator consisting of saddle point local searcher for attaining precise saddle points. Experimental study on synthetic benchmark functions of varying complexity demonstrates the efficacy of the proposed framework for locating multiple first-order saddle points of higher precision, when compared to its evolutionary compeers.

From synthetic benchmark problems to real-world problems, the multi-modal memetic framework is applied to systems representing the two main streams of the molecular systems, namely, water clusters and small ring-deficient molecules. In particular, the framework has been extended to efficiently explore the landscape of biomolecular systems for transition states, especially when using first-principle calculations which describe mathematically the interaction among atoms and their subatomic structures in biomolecular systems. For water cluster, the proposed framework is extended with water-specially-designed reproduction operators. The application of the proposed framework on water cluster has shown not only to uncover those previously reported but also to establish newly discovered transition states of size 2-4 water molecule, with the minimal computational cost among to its compeers.

For ring-deficient covalently-bonded small molecules, the multi-modal memetic framework is extended with tree-based evolutionary operators, to induce configurational changes on the molecules while maintaining both the tree-structure and the molecular properties unchanged. Comparisons to recent advancements in related state-of-the-art algorithms on several small, yet very important, molecules showed that the multi-modal memetic framework not only uncovered the largest set of transition states, but also attained the achievements at significantly lower computational costs. With further analysis on the transition states uncovered by the framework, it has been found out that the number of oxygen atoms and rotatable bonds play a key role in the flexibility of the molecules, allowing larger number of transition states to be uncovered while carbon atoms have been observed to limit the flexibility of the molecule, decreasing the number of transition states to be uncovered. The number of hydrogen atoms as well as total number of atoms, in contrast, has not shown any significant impact.

The framework, extended for small ring-deficient molecules, has been, equally important, considered for locating glutamic acid stereoisomers, which later uncovered to have a modulation effect on the glutamatergic systems. Such findings open new door for neuroscientist to study the impact of the glutamic acid stereoisomers on the function and dysfunction of the glutamatergic systems.

From this study, we conclude that the memetic computing paradigm is the most efficient evolutionary paradigm ever applied for saddle point identification, and incorporation of domain knowledge into the evolutionary operators improves search performance significantly, constraining search process to explore only the feasible region of the landscape.

Acknowledgements

I would like to express my deepest gratitude to my academic supervisor **Associate Professor Ong Yew Soon** of Nanyang Technological University.

His sincere encouragements during my indecisive moment and inspirational insight into the many aspects of my research direction have faithfully guided me throughout my research work. Without Professor Ong's guidance, the fruitful achievement of this project will never be possible.

I would also like to thank my friends for their encouragements and the staffs at Emerging research Lab (ER Lab) and Center of computational intelligence (C2I) for their technical supports.

Last but not the least, special thanks to my parents, wishing God blessing to cover them, for their constant support, understanding and love, until their last breaths.

Mostafa Mostafa Hashim Ellabaan

Table of Contents

Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
List of Symbols	xi
Chapter 1: Introduction	1
1.1 Research Objectives	3
1.2 Thesis Contribution	6
1.3 Organization of the Thesis	8
Chapter 2: Locating First-order Saddle Point - A Literature Review	11
2.1 Conventional Optimization Methods	14
2.1.1 Single-ended Methods	14
2.1.2 Doubly-ended Methods	22
2.2 Stochastic First-order Saddle Points Finding Methods	26
2.2.1 Simulated Annealing for First-order Saddle Point Optimization	27
2.2.2 Evolutionary Optimization Methods	28
2.3 Discussions	31
2.3.1 Discussion on Conventional Methods	32
2.3.2 Discussion on Stochastic Methods	38
2.3.3 Comparison between Stochastic and Conventional Methods	40
PART I: Methodology	43
Chapter 3: Stochastic Multi-start Local Search for Locating Multiple Saddle Points	46
3.1 Stochastic Multi-Start Local Search (SMLS)	47
3.2 Experimental Study	47
3.2.1 Benchmark Problems	48
3.2.2 Performance Analysis	54
3.3 Conclusion	57
Chapter 4: A Multi-Modal Memetic Framework for Locating Multiple Saddle Points	59
4.1 Evolutionary Search Operators	62
4.2 Niching Operator: The Valley-adaptive Clearing Scheme	66
4.2.1 Hill-valley Detection	68
4.2.2 Valley-adaptive Clearing Scheme	69
4.2.3 Complexity of Valley-adaptive Clearing Scheme	73
4.2.4 Experimental Study on Clearing Schemes	74

4.3	Memetic Operator: Life-time Learning in Memetic Algorithms and Valley Elitism	77
4.4	The Multi-modal Memetic Framework	81
4.5	Experimental Study on Saddle Point Optimization	83
4.5.1	Precision Accuracy of Uncovered First-order Saddle Points	85
4.5.2	Convergence Speed	86
4.6	Conclusion	89
	PART II: Biomolecular Applications: Water Cluster and Ring-deficient Covalently-bonded Small Molecules	91
	Chapter 5: Biomolecular Systems	94
5.1	Experimental Approaches for Studying Biomolecular Systems	96
5.2	Modeling Approaches for Biomolecular Systems	98
5.2.1	Quantum Mechanics Modeling	98
5.2.2	Molecular Mechanics Modeling	99
5.2.3	Hybrid Quantum Mechanics and Molecular Mechanic Modeling	100
5.3	Types of Biomolecular Systems	100
5.3.1	Water Cluster	101
5.3.2	Covalently-bonded Molecular Systems	102
5.4	Transition States Identification in Biomolecular Systems	103
5.4.1	Structural Configuration/Representation Space and Potential Energy Function	103
5.4.2	Transition State Sampling and Optimization	104
5.5	Molecular Extensions for Multi-Modal Memetic Framework	106
5.5.1	Life-Time Learning Procedure	106
5.5.2	Identification of Duplicate Molecular Structure Configurations and Archival Procedure	108
5.6	Experimental Study	108
5.6.1	Benchmark Example Biomolecular Systems	109
5.6.2	Results	112
5.7	Conclusion	118
	Chapter 6: Evolutionary Discovery of Transition States in Water Clusters	119
6.1.	Water-Specially-Designed Operators	121
6.1.1	Molecular Water Clusters Initialization	122
6.1.2	Molecular-Based Water Evolutionary Operators	123
6.2.	Experimental Study	128
6.2.1	Water Cluster Test Problems	129
6.2.2	Performance Measures	129
6.2.3	Numerical Results	131
6.3.	Visualization of Discovered Water Cluster Transition States	136
6.4.	Conclusion	142

Chapter 7: Evolutionary Discovery of Transition States in Ring-deficient Covalently-bonded Small Molecules	144
7.1. Covalent-bond-driven operators	145
7.1.1 Chromosome Representation	146
7.1.2 Knowledge Extraction and Embedding	148
7.1.3 Tree-Based Molecular Evolutionary Operators	149
7.2. Experimental Study	160
7.3. Transition State Web Server	165
7.4. Toward Better Understanding of Glutamatergic Systems: Glutamate Stereoisomers Discovery and Receptor Binding Affinity	167
7.5. Conclusion	171
Chapter 8: Concluding Remarks and Future Research Investigations	174
8.1. Conclusion	174
8.2. Potential Future Research Investigations	177
8.2.1. Algorithmic Issues	177
8.2.2. Domains of Application	182
Author's Publications	184
References	185

List of Figures

Figure 2-1: Visualization of a saddle point, redrawn from [35].	12
Figure 2-2: Dimer structure as defined by Henkelman, redrawn from [18].	21
Figure 3-1: 2D Sines function landscape with first-order saddle points tabulated and highlighted by white dots in the range of [-2, 2].	50
Figure 3-2: 2D Multi-function landscape with first-order saddle points tabulated and highlighted by white dots in the range of [-0.5, 05].	51
Figure 3-3: 2D Rastrigin landscape with first-order saddle points tabulated and highlighted by white dots in the range of [-0.75, 0.75].	52
Figure 3-4: 2D Schwefel function landscape with first-order saddle points tabulated and highlighted by white spots in the range of [-75, -75].	53
Figure 3-5: Percentages of uncovered first-order saddle points using SMLS with varying gradient precision accuracy, on all benchmark test problems.	55
Figure 4-1: Stochastic universal sampling.	63
Figure 4-2: Scatter crossover.	64
Figure 4-3: Arithmetic scatter crossover.	65
Figure 4-4: Hill-valley detection.	69
Figure 4-5: Percentages of uncovered optima by clearing schemes, on the three 2D benchmark problems.	75
Figure 4-6: Percentages of optima uncovered by archiving clearing schemes and fitness sharing on the benchmark problems, for different niche radii (Q_{clear}).	77
Figure 4-7: Valley-elitism-local refinement strategy.	79
Figure 4-8: Whole-valley-local refinement strategy.	80
Figure 4-9: A Multi-modal Memetic Framework for locating first-order saddle points. [Note that the components that will be customized later for the problem at hand are highlighted by <<Customized>>. The evolutionary operators will be customized for the water clusters and ring-deficient covalently bonded small molecules. The memetic operator will be customized for biochemical systems in Chapter 5].	83
Figure 4-10: Percentages of uncovered first-order saddle points using different algorithms with varying gradient precision accuracy, on all benchmark test problems.	88
Figure 5-1: An illustrative example of the A) Dimer water cluster, B) Trimmer water cluster, and C) Tetramer water cluster.	110
Figure 5-2: Ring-deficient covalently-bonded small molecules.	112
Figure 5-3: Performance of different algorithms.	114
Figure 5-4: Nonsensical molecular structures and valid GABA structures.	115
Figure 5-5: Percentage of valid and nonsensical molecular structure generated after initialization.	116
Figure 5-6: Percentage of valid and nonsensical molecular structure generated after both crossover and mutation.	117
Figure 6-1: Basic water monomer.	123
Figure 6-2: An example initial population consisting of 12 water chromosomes (W_c), on the four water cluster problem or $(H_2O)_4$.	124
Figure 6-3: An illustration of the molecular crossover operator on two water clusters or $(H_2O)_2$.	125

Figure 6-4: An illustration of rotation-based mutation operators for water cluster A) a two water cluster molecular structure, B) upon rotation of a molecule, around the center of mass of the water cluster, and C) upon rotation of a water molecule in the water cluster, around the center of mass of the water molecule.....	127
Figure 6-5: An illustration of translation-based mutation operators for water cluster A) a two water cluster molecular structure, B) upon translation of a random molecule in the water cluster, and C) upon translation of a random atom in the water cluster.	127
Figure 6-6: An illustrative example of the A) Dimer water cluster, B) Trimmer water cluster, and C) Tetramer water cluster.	130
Figure 6-7: Mean wall clock time of 10 independent runs incurred by MOL-MA, SNMA, SMLS, M-S-CMA-ES, BGA, and CHGA on the Dimer, Trimmer and Tetramer water clusters.	131
Figure 6-8: Box plot on the percentage of transition states uncovered (%Ts) by each algorithm on the Dimer, Trimmer, and Tetramer water cluster problems.	134
Figure 6-9: Box plot on the number of gradient evaluations count incurred by the algorithms considered on the Dimer, Trimmer, and Tetramer water cluster problems.	135
Figure 6-10: Dendrogram for the hierarchical clustering of Dimer transition states.....	138
Figure 6-11: Dendrogram for the hierarchical clustering of Trimmer transition states.....	139
Figure 6-12: Dendrogram for the hierarchical clustering of Tetramer transition states.	140
Figure 7-1: Tree node in UML class notation.....	148
Figure 7-2: Chromosome tree-structure exemplified on GABA, taking C ₆ or alpha carbon as a pivot atom or roo/t node.....	150
Figure 7-3: Knowledge extraction and embedding initialization procedure.	151
Figure 7-4: An example initial population of GABA structures.....	152
Figure 7-5: Tree-based crossover operator, exemplified on GABA.....	156
Figure 7-6: Rotation of randomly chosen substructure, exemplified on GABA.	157
Figure 7-7: Rotation of randomly chosen leave nodes, exemplified on GABA.....	158
Figure 7-8: Translation of a random set of atoms, exemplified on GABA.	159
Figure 7-9: Distribution of uncovered transition states across relative energy and structure dissimilarity with reference to the lowest-energy transition state	165
Figure 7-10: TranSGen webserver main page.	167
Figure 7-11: Examples of glutamate stereoisomers identified. AMPA receptor binding domain (right) is enlarged (right bottom) to show the structural details of the receptor and the glutamate molecule (red). Distinct properties in AMPAR interaction with varied glutamate stereoisomers is the focus of our study.	169
Figure 7-12: (A) the binding energy distribution of glutamate stereoisomers against IGluR2. (B) statistical summary on the binding energy of glutamate stereoisomers.	170
Figure 8-1: Algorithmic-issues depicted in the framework: A) shows where the machine learning stage should be incorporated; B) shows where the symbioses life-time learning procedure should be considered; C and D) show where graph mining and indexing should be incorporated.....	179
Figure 8-2: The multi-scale framework for the current algorithmic framework proposed.....	181

List of Tables

Table 2-1: Quasi-Newton iterative scheme.....	16
Table 2-2: Comparison among single-ended methods.....	34
Table 2-3: Comparison among doubly-ended methods.....	37
Table 2-4: Stochastic methods.....	41
Table 3-1: A stochastic multi-start local search.....	47
Table 3-2: Number of gradient calls incurred by the algorithms considered on different landscapes.....	57
Table 4-1: Valley-adaptive clearing scheme.....	70
Table 4-2: Pseudo code of valley identification phase.....	71
Table 4-3: Pseudo code of valley clearing phase.....	72
Table 4-4: Pseudo code of valley replacement phase.....	73
Table 4-5: Success rate of different local search strategies.....	80
Table 4-6: Average number of gradient calls incurred by different local search strategies.....	81
Table 4-7: Descriptions of the state-of-the-art algorithms considered.....	84
Table 4-8: Experimental settings.....	85
Table 4-9: Standard deviation of uncovered first-order saddle points using different algorithms with varying gradient precision accuracy, on all benchmark test problems.....	87
Table 4-10: Number of gradient calls incurred by the algorithms considered on different landscapes.....	89
Table 5-1: Molecular parametric settings of the algorithms.....	109
Table 6-1: Molecular water clusters crossover.....	124
Table 6-2: Descriptions of the state-of-the-art algorithms considered.....	129
Table 6-3: Descriptions of the performance measures considered.....	131
Table 6-4: Parametric settings of the algorithms. These settings are obtained from the domain knowledge of scientists and physicists on water clusters.....	132
Table 6-5: Properties of the algorithms considered (Y denotes the existence of a property while N refers to its absence).....	132
Table 6-6: Cost reduction in (%) by MOL-MA over the other counterpart algorithms (N/A indicates not applicable as the algorithm fails to locate any transition states).....	135
Table 6-7: Dimer transition states.....	138
Table 6-8: Trimmer transition states.....	139
Table 6-9: Tetramer transition states.....	141
Table 7-1: Bond lengths constraints [118].....	149
Table 7-2: Properties of the algorithms considered (Y denotes the existence of a property while n refers to its absence).....	161
Table 7-3: Number of transition states uncovered (improvement rate in percentage) [Note that the best is highlighted in bold].....	161
Table 7-4: Number of gradient evaluations (the best is highlighted in bold).....	162
Table 7-5: Characteristics of uncovered transition states in terms of number of atoms, number of rotatable bonds, and fitness-distance correlation.....	163

List of Symbols

$f(\mathbf{x})$	Function landscape
$\mathbf{X}_T$	A set includes all possible first-order saddle points
$\mathbf{H}_k$	Hessian Matrix at step k
$\gamma_{i,k}$	The i^{th} eigenvalue of $\mathbf{H}_k$
$\mathbf{g}_k$	Gradient at step k
$g_{i,k}$	i^{th} component of the gradient $\mathbf{g}_k$
$\Delta \mathbf{x}_k$	Step vector at step k
$\Delta \mathbf{x}_{k,i}$	i^{th} component of the Step vector $\Delta \mathbf{x}_k$
$\mathbf{t}$	Tangent vector
$\mathbf{x}_k$	Intermediate solution vector at step k
$\mathbf{v}_{i,k}$	i^{th} eigenvector of $\mathbf{H}_k$
x_{max}	Trust region radius
$Q(\Delta \mathbf{x}_k)$	Quadratic region approximation
$R(\Delta \mathbf{x}_k)$	Rational region approximation
$\mathbf{T}$	A tridiagonal matrix
C	Curvature at $\mathbf{x}_k$
$\Delta \mathbf{v}$	The velocity
M	Mass
Δt	Time interval
D	Number of design vector variables
$f^*(\mathbf{x})$	Fitness function
$\mathbf{X}$	Vector of design variables
$\mathbf{P}$	Population individuals
$\mathbf{X}_T$	Set of the archived solutions
N	Population size
Q_{clear}	Niche radius
$\mathbf{V}_{\text{id}}$	A set of all individuals sharing a common valley
$\mathbf{V}_{\text{groups}}$	Set of all $\mathbf{V}_{\text{id}}$ in the current population $\mathbf{P}$
ξ	Fitness penalty
w	Number of elites in $\mathbf{V}_{\text{id}}$
k	The maximum number of elites $\mathbf{V}_{\text{id}}$ can have
x_i^g	The i^{th} component of $\mathbf{x}_i$ at g^{th} generation
e_n	Number of negative eigenvalues
ε	A small real number
$\ \mathbf{g}\ $	Gradient norm
δ	Gradient precision accuracy threshold
MEP	The Minimum Energy Path
QM	Quantum Mechanics
MM	Molecular Mechanics
QM/MM	Hybrid molecular and quantum mechanics
NMR	Nuclear magnetic resonance
IR	Infrared
$\mathbf{L}$	Landscape
$\mathbf{X}$	The structural configuration space

f	The fitness function
φ	The distance measure between two structural configurations in $\mathbf{X}$.
f_{max}	A user-defined energy threshold.
σ_1, σ_2	Random angles
$\mathbf{T}$	A random translation vector
$O-O_{min}$	A minimum (O-O) distance between two neighboring oxygen atoms
$O-H_{min}$	A minimum (O-H) distance between oxygen and hydrogen atoms of two neighboring water molecules
W_{cl}	Water Cluster
$\alpha_{b_T(y-x)_{min}}$	The minimum bond length between atom y and atom x
$\alpha_{b_T(y-x)_{max}}$	The maximum bond length between atom y and atom x
$b_T(y-x)$	Bond type between atom y and atom x
Σ	Povit atom or root node
$\Omega_{A,\Sigma}$	Atom sets of a given molecule
$\Omega_{B,\Sigma}$	Bond sets of a given molecule
$\mathbf{P}$	Coordinate system of a given molecule
$\mathbf{p}_{cm}$	The center of mass $\mathbf{P}$
$\mathbf{P}_t$	Coordinate system of a given molecule ($\mathbf{P}$) after translation to the center of mass
$\mathbf{Q}$	Coordinate system of a given molecule
$\mathbf{Q}_{cm}$	The center of mass $\mathbf{Q}$
$\mathbf{Q}_t$	Coordinate system of a given molecule ($\mathbf{Q}$) after translation to the center of mass.
$\mathbf{C}$	The covariance matrix $\mathbf{C}$ between ($\mathbf{P}_t$ and $\mathbf{Q}_t$) is then calculated as
$\mathbf{R}_{opt}$	The optimal rotation matrix
Ω_R	Reference set
$\mathbf{a}(\mathbf{s})$	A set of atoms falls on nodes precedents to $\mathbf{s}$
ϑ_1, ϑ_2	Random angles
$\mathbf{R}_{(\vartheta_1, \langle p(\mathbf{s}) - \mathbf{s} \rangle)}$	The rotation matrix where rotation is performed around $\langle p(\mathbf{s}) - \mathbf{s} \rangle$ axis with a random angle ϑ_1 .
$\mathbf{R}_{(\vartheta_2, \langle p(p(\mathbf{x})) - p(\mathbf{x}) \rangle)}$	The rotation matrix where rotation is performed around $\langle p(p(\mathbf{x})) - p(\mathbf{x}) \rangle$ axis with a random angle ϑ_2 .
Ω_T	A set of randomly chosen atoms/nodes to be arbitrarily translated.

Chapter 1: Introduction

Saddle points are important in many areas of science and engineering, which include fields like economy [1], chemistry [2], and biology [3]. In engineering, saddle points have wide spectrum of applications. In robotic navigations for example, first-order saddle points have been used as a highly accurate guide for identifying x-junction in the images captured by robotic visual sensors. Building roadmaps in visual models also represents another important general-purpose application of saddle points in computer science and robotics [4]. The roadmaps could be used, for example, in estimating human pose captured from monocular images and identifying the minimum energy path which is important for path-planning and robotic navigation [5].

In science, the geometry optimization of chemical structures, especially transition structures, represents another important application of saddle points. Transition state structures play a key role in the chemical reaction rate theory [3, 6], especially for estimating the reaction rate of chemical reactions and processes. Examples of some chemically interesting processes include the transportation of ions across membranes [7, 8], diffusion of clusters of absorbed atoms on solid surfaces [9], nucleation of first-order phase transitions [10], denaturation of biological molecules [11], membrane fusion, exchange of O₂ or CO from myoglobin and, last but not the least, protein folding. Transition structures or first-order saddle points, unlike other stable structures, are energy maxima along the minimum energy path connecting reactant and product for a chemical reaction, posing fundamental difficulties in finding such structures. Due to their fleeting nature, transition structures are almost impossible to be isolated experimentally.

Therefore, search for such structures computationally is unavoidable and remains a great challenge [12].

To identify first-order saddle points, a plethora of optimization methods have been proposed in the last decades. They can be classified as conventional and stochastic methods [12]. Conventional methods make use of domain specification information such as gradient and/or Hessian in the search process. These methods can be further segregated into doubly-ended and single-ended approaches [13]. In doubly-ended methods, two minima (reactant and product) are assumed to be available to find transition states that lie within [14-17]. In contrast to doubly-ended methods, single-ended methods require no prior information on the destination of the transition [13, 18]. However, like most conventional numerical methods, it requires good initial guesses to locate the transition states. Due to the use of prior information, conventional methods are well established to converge efficiently to a transition state, given an appropriate initial guess.

In stochastic search methods, instead of using only one or two solution(s) to start the search, a population of individuals is usually used to explore the search space [19]. To date, few stochastic methods have been proposed for locating of transition states, making it a fertile area for further research investigations. Chaudhury et al., for example, suggested a simulated annealing method for locating saddle points [20]. Bungay et al. [19] also presented a Genetic Algorithm (GA) method that uses information on the order of eigenvalues to bias the evolutionary search towards first-order saddle points. As both methods require second derivative information (Hessian matrix), Chaudhury et al. [21] extended the effort with a GA that estimates the eigenvalues from gradient information,

with the assumption that these points lie on the minimum energy path. For such an assumption to hold, the GA must start its search from the global minimum configuration [21]. Besides, the proposed method impedes the intrinsic parallelism benefits of GA. In addition, most stochastic methods proposed for finding transition states are plagued with many problems, including the lack of sufficient precision, slow convergence speed, as well as, the inability to maintain multiple saddle points in the search.

1.1 Research Objectives

As highlighted previously, stochastic search approaches designed to identify saddle points are yet to be well investigated, especially those algorithms stemming from an evolutionary or memetic methodology. Evolutionary algorithms (EAs) represent a new class of stochastic optimization methods, which imitate the biological evolution and/or the social behavior of species. In light of their excellent adaptabilities, evolutionary algorithms are well-established as strong candidates for handling multimodal optimization problems. One of their key strengths lies in their ability to maintain a diverse set of solutions [22-26] in a single execution of the algorithm, instead of a series of separate executions as in the conventional methods. Additionally, evolutionary algorithms are described by less susceptibility to the shape or continuity of the landscape, whereas these issues always challenge the search progress in conventional methods.

In addition to these advantages, EAs have several other merits as detailed below:

- **Parallelism:** In EAs, population individuals can evolve independently of the evolution of other competing solutions. Thus, the evolution of each individual can be achieved in a parallel manner. It is only the selection which requires some serial

degree of processing. Therefore, the computational time required may be reciprocally proportional to the computing machines, providing sufficient speed to produce solutions for difficult problems in reasonable periods of time.

- **Robust to Dynamic Changes:** Conventional methods of optimization are not robust to changes in the environment and often necessitate a new complete run to identify a solution. In contrast, EAs can identify solutions for varying circumstances. The available evolved population provides a basis for further enhancements and in most cases it is not necessary to reinitialize the population at random.
- **Self-Optimization Capability:** Appropriate settings of exogenous variables are a necessity for proper operation of most conventional optimization techniques as well as evolutionary algorithms. However, using evolutionary process, EAs can, itself, optimize their parameters as part of the search for optimal solutions.
- **Knowledge Incorporation:** Incorporating domain knowledge into the algorithm when dealing with a particular problem is always recommended [27]. Specialized algorithms generally surpass general-purpose algorithms on a particular domain of interest. Unlike other optimization framework, EAs present a framework that can easily include such knowledge. For example, knowledge can be integrated straightforwardly in mutation or recombination operations. Embedding such information centered the evolutionary search on the feasible region of the search space, leading to a more efficient exploration of possible solutions.
- **Hybridization:** Evolutionary algorithms can also be integrated with more conventional optimization techniques. This can be as straightforward as the use of

a Newtonian optimization method after major exploration with an evolutionary algorithm, or it may perhaps involve concurrent application of specialized algorithms.

Extending evolutionary computation, memetic computation represents an emerging subject that has engrossed growing research interest in the recent few decades, with a mounting number of accomplishment reported [28]. Memetic computation [29] in its earliest form was introduced as a memetic algorithm (MA), which is a coupling between global, population-based search and individual learning, where the second is often referred to as a meme, capable of individual fine-tuning in converging to the accurate local optima swiftly. They are stimulated by Darwin's theory of natural evolution and Dawkins' notion of a meme. MA has manifested as a special type of hybrid global-local approach that enhances both exploration and exploitation in the search. Many MAs have, up to date, been developed for solving real-world problems, attaining better quality solutions with minimal computational cost. Such hybrids have been fruitfully employed to solve various medical, scientific and engineering design problems [30,31], where higher quality solutions are reached more efficiently than conventional evolutionary algorithms [32, 33].

Thus, the main objectives of this research include the followings:

- Reviewing and categorizing existing methods proposed for locating saddle points,

- Development of novel approaches and methodologies based on the evolutionary and memetic theory [34] for the purpose of locating saddle point, and
- Incorporating domain knowledge and applying the proposed methodologies to identify transition states of biomolecular systems including water cluster and ring-deficient covalently-bonded small molecules.

1.2 Thesis Contribution

This thesis has contributed to the scientific communities through different folds:

1. It provides a comprehensive survey on the state-of-the-art methodologies of finding saddle points. In this survey, we have categorized saddle points finding methods into conventional and stochastic search methodology. The conventional methodologies have been further categorized according to their prerequisite into single-ended which necessitate an initial guess of the transition states to start with the other one, doubly ended methods necessitate two minima, representing the reactant and product for their execution. The stochastic search methods explore the problem search space globally guided by a fitness function. A few methods have been proposed to locate saddle points. Most of the existing state-of-the-art methods lack efficiency in term of computational cost and effectiveness in terms of number of saddle points uncovered and the quality of saddle points uncovered.
2. To explore search space efficiently and locate multiple high-quality saddle points, a multi-modal memetic framework has been presented. The framework composes of evolutionary operators such as crossover and mutation, multi-modal operator,

called valley-adaptive clearing, and memetic operator, which enable efficient exploration and exploitation of the search space. This framework can also be considered for any other problems. However, incorporation of domain knowledge significantly improves its search performance.

3. For effective exploration of biomolecular systems, the proposed framework has been extended, considering an efficient molecular local searcher and structure similarity metric.
4. Two other extensions of the framework have been proposed in this thesis. The first extension considers water-specially-designed operators that have enabled the framework to locate not only previously uncovered water cluster transition states but also establish newly-uncovered ones.
5. The second extension of the framework has enabled the framework to deal efficiently with ring-deficient covalently-bonded molecules. This extension involves tree-representation of the molecular structure and tree-based evolutionary operators which enables the structure or the bonding-signatures of the molecule to be maintained across the whole execution. This extension has enabled the framework to locate largest set of uncovered transition states with the minimal computational cost when compared to the state-of-the-art.
6. Based on these extended forms of the framework, a web services have been developed to allow scientists to find transition states of biomolecular systems.
7. Also we had employed the algorithm for finding stereoisomers of glutamic acid, finding more than 524 stereoisomers which have been examined against glutamate receptor 2, showing significant modulation of the glutamate receptors. This not

only will help scientists to better understand the glutamatergic system, the major communication systems in the central nervous systems, but also may help scientists in developing therapeutical management tools for many neurodegenerative disorders including Alzheimer's disease.

1.3 Organization of the Thesis

The remaining parts of the thesis are arranged as follows:

- An intensive review of the existing methods capable of locating saddle points is presented in Chapter 2.
- Chapters 3-7 contain the core contributions of this dissertation, which are divided into Part 1 Methodologies, Part 2 Biomolecular Applications: Water clusters and Ring-deficient Covalently-bonded Molecules.
- Part 1 Methodologies: This part compose of Chapters 3 and 4:
 - Chapter 3 presents a preliminary investigation on saddle point stochastic search where a simple stochastic multi-start local search is investigated for the discovery of multiple first-order saddle points.
 - Chapter 4 presents a novel multi-modal memetic framework to attain multiple solutions precisely and efficiently. This framework composes of crossover and mutation operators, valley-adaptive clearing scheme, a saddle point local searcher and archival procedure. The framework manages to locate multiple first-order saddle points of high precision and with minimal computational cost, rectifying the issues raised in Chapter 3.

- Part 2 Biomolecular Applications: Water Clusters and Ring-deficient Covalently-bonded Small Molecules:
 - Chapter 5 presents the biomolecular systems applications considered in the present dissertation namely, water clusters, ring-deficient covalently-bonded molecules. In this chapter, we present different types of biomolecular systems which are classified into two main streams: molecular clusters and covalently-back-boned systems. The general framework has been extended for biomolecular systems, considering molecular distance metric called USR and an efficient molecular transition states local searcher, called Berny algorithm.
 - Chapter 6 presents another extension that would allow the multimodal memetic framework to explore the water cluster solution space efficiently. The multi-modal memetic framework is extended with a specially-designed water cluster crossover operators and mutation strategies that are inspired by the dynamics of water molecules.
 - Chapter 7 presents another extension that would allow the multimodal memetic framework to efficiently explore the solution space of ring-deficient covalently-bonded small molecules. The multimodal memetic framework is extended with tree-structured covalently-bonded molecular operators that enable bonding-signatures maintenance across the whole execution.

- Last but not least, Chapter 8 concludes this thesis and outlines some potential future research directions of transition states identification, both algorithmically and application-wise.

Chapter 2: Locating First-order Saddle Point - A Literature Review

Saddle points are a subset of stationary points, where the Hessian matrix possesses at least one positive eigenvalue. The problem of locating saddle points is defined mathematically as follows:

$$\min_{\mathbf{x} \in \mathbf{X}} \max_{\mathbf{z} \in \mathbf{M}} f(\mathbf{x}, \mathbf{z}) \quad (\text{Eq. 2.1})$$

where $\mathbf{X}$ and $\mathbf{M}$ are convex sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively while f is a convex-concave function defined over $\mathbf{X} \times \mathbf{M}$. In particular, $f(\cdot, \mathbf{z}): \mathbf{x} \rightarrow \mathbf{R}$ is a convex for every $\mathbf{z} \in \mathbf{M}$, and $f(\mathbf{x}, \cdot): \mathbf{z} \rightarrow \mathbf{R}$ is a concave for every $\mathbf{x} \in \mathbf{X}$ [35].

A solution to problem (Eq. 2.1) is a vector pair $(\mathbf{x}^*, \mathbf{z}^*) \in \mathbf{X} \times \mathbf{M}$ such that

$$f(\mathbf{x}^*, \mathbf{z}) \leq f(\mathbf{x}^*, \mathbf{z}^*) \leq f(\mathbf{x}, \mathbf{z}^*), \quad \text{for all } \mathbf{x} \in \mathbf{X} \text{ and } \mathbf{z} \in \mathbf{M} \quad (\text{Eq. 2.2})$$

Such a vector $(\mathbf{x}^*, \mathbf{z}^*)$ represents a saddle point on the function $f(\mathbf{x}, \mathbf{z})$ over the set $\mathbf{X} \times \mathbf{M}$.

An example of such points is presented in Figure 2-1, where $f(\mathbf{x}, \hat{\mathbf{z}}(\mathbf{x}))$ is the curve of maxima $\hat{\mathbf{z}}(\mathbf{x}) = \arg \max_{\mathbf{z}} f(\mathbf{x}, \mathbf{z})$; and $f(\hat{\mathbf{x}}(\mathbf{z}), \mathbf{z})$ is the curve of minima, $\hat{\mathbf{x}}(\mathbf{z}) = \arg \min_{\mathbf{x}} f(\mathbf{x}, \mathbf{z})$. Saddle point corresponds to the point where these two curves intersect.

As higher order saddle points are not of any chemical interest, we consider in this study only the case where z is a scalar and $\mathbf{M}$ is a vector, which is associated with first-order saddle points or transition states. For simplicity, $f(\mathbf{x}, z)$ will be represented as $f(\mathbf{x}')$, where $\mathbf{x}'$ is the concatenation of $\mathbf{x}$ and z or $\mathbf{x}' \in \mathbf{X} \times \mathbf{M}$.

Alternatively, first-order saddle points are defined as stationary points with a vanished gradient and only one negative eigenvalue in the Hessian matrix [14]. These points can be mathematically expressed as:

$$\mathbf{X}_T = \left\{ \mathbf{x}_k \mid \left(\frac{\Delta f(\mathbf{x}_k)}{\Delta \mathbf{x}_k} = 0 \right) \& (\gamma_{k,i} < 0) \right\} \quad (\text{Eq. 2.3})$$

where $\mathbf{x}_i \in \mathbf{R}^d$, d is the dimensional size, $f(\mathbf{x}_i) \in \mathbf{R}$, $\mathbf{X}_T$ is the set that includes all possible first-order saddle points, and $\gamma_{k,i}$ is the only negative eigenvalue of the Hessian matrix $\mathbf{H}_k$, respectively. ‘&’ denotes a logical AND operator.

Figure 2-1: Visualization of a saddle point, redrawn from [35].

As highlighted in the first chapter, saddle points are very important for a large spectrum of applications in numerous subjects of science and engineering. However, saddle points are in general difficult to locate. The difficulties originate, first and foremost, from the lack of intuitive knowledge about the first-order saddle points, so that the search can rarely be launched from a point close enough to the first-order saddle point being sought [5]. Large number of local minima in a complex landscape can also hinder the progress of the search algorithm. Further, locating saddle points become even more tedious when the area around saddle points is a plateau [6].

In the last decades, many methods have been proposed to tackle the problem of finding first-order saddle points or transition states. They can be classified into two categories: conventional and stochastic methods. Conventional methods (see Section 2.1) typically use gradient and/or Hessian information in the process of finding first-order saddle points. These methods can be further classified into doubly-ended and single-ended approaches [14]. In doubly-ended methods, the availability of two known minima is required in order to find saddle points in between. In contrast to doubly-ended methods, single-ended methods require no information about the location of the transition or saddle points. However, single-ended search methods assume that a good initial guess is available.

In stochastic search methods (see Section 2.2), instead of utilizing only one or two solution(s) to find saddle points, a population of individuals is used to explore the search space for saddle points [1]. To date, few stochastic methods have been proposed for finding first-order saddle points, making it a fertile area for further research investigation.

A detailed presentation and discussion of different transition state finding methods are presented in the following subsections.

2.1 *Conventional Optimization Methods*

Conventional optimization methods are mathematics based techniques in which information required to solve optimization problems are assumed to be available and, therefore, the effect of any variable can be computed with certainty (i.e. without any room for random variation). The common theme among the conventional methods is that their search includes proceeding with successful candidate solutions sequentially until the convergence criteria are fulfilled. With a given starting solution, the search space is explored locally. Conventional methods proposed for finding transition states can be classified into single-ended (see section 2.1.1) and doubly-ended (see section 2.1.2) methods.

2.1.1 *Single-ended Methods*

Single-ended methods are a class of transition state finding approaches in which a good initial guess of the first-order saddle points is assumed. Such methods can be further classified according to the number of eigenvectors that a search method follows, namely, all-mode and minimum-mode following methods.

A. *All-Mode Following Methods*

All-mode following method refers to the group of first-order saddle point finding approaches that move the system configuration along all directions simultaneously, in each iteration during the search. Quasi-Newton provides the theoretical basis for many

exciting all-mode following methods. In this section, we present some of the Quasi-Newton methods reported in the literature for finding transition state, which involves an initial estimation $\mathbf{x}_k$ of the first-order saddle-point. The second-order expansion of Taylor series in the step direction $\Delta\mathbf{x}_k$ can be represented as follows:

$$f(\mathbf{x}_k + \Delta\mathbf{x}_k) \approx Q(\Delta\mathbf{x}_k) = f(\mathbf{x}_k) + \mathbf{g}_k^T \Delta\mathbf{x}_k + \Delta\mathbf{x}_k^T \mathbf{H}_k \Delta\mathbf{x}_k \quad (\text{Eq. 2.4})$$

where $f(\mathbf{x}_k)$ is the energy or objective function at a point $\mathbf{x}_k$, $\mathbf{g}_k$ is the gradient vector and an approximate Hessian matrix $\mathbf{H}_k$, given the stationary condition $\nabla Q(\Delta\mathbf{x}_k) = 0$, Newton step ($\Delta\mathbf{x}_k$) is calculated as $\Delta\mathbf{x}_k = -\mathbf{H}_{k-1}^{-1} \mathbf{g}_k^T$, or $\Delta\mathbf{x}_k = -\sum_{i=1}^d \frac{g_{k,i}}{\gamma_{k,i}} \mathbf{v}_{k,i}$ where $g_{k,i}$ is the component of $\mathbf{g}_{k,i}$ along the eigenvector $\mathbf{v}_{k,i}$ and the $\gamma_{k,i}$ is the i^{th} eigenvalue of $\mathbf{H}_k$. However, it is not essential to calculate the exact Hessian matrix during each iteration. Instead, an approximated Hessian matrix can be used to avoid the expense of recalculating Hessian matrix. The Hessian matrix can be approximated as $\mathbf{H}_{i+1} = \mathbf{H}_i + \Delta\mathbf{H}_i$. To calculate $\Delta\mathbf{H}_i$, there are many ways. Among the most common are Powell updates and Bofill updates [36, 37].

Thus, at the beginning of the search process, an iterative calculation of the Hessian should be carried out using an appropriate approximation of the second derivative matrix. The basic outline of general Quasi-Newton algorithm is presented in Table 2-1. The differences among the algorithms lie on how each algorithm implements each step of Table 2-1. For example, to update the Hessian matrix, Baker [38] and Simons et al. [39] consider the formula of Powell [40], while Schlegel [41] proposed otherwise. The choice

of the direction $\Delta \mathbf{x}_k$ in step (1) and the scaling procedure in step (2) varies across all three methods. The details of each method are discussed in what follows.

Table 2-1: Quasi-Newton iterative scheme.

1. Determine a search direction $\Delta \mathbf{x}_k$ 2. Scale the search direction step such that $\Delta \mathbf{x}_k$ to be below an upper bound x_{max} 3. Calculate the new point $\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta \mathbf{x}_k$ 4. Update of the approximate Hessian matrix $\mathbf{H}_{k+1}$ from $\mathbf{H}_k$, $\Delta \mathbf{x}_k$, $\mathbf{g}_{k+1}$, and $\mathbf{g}_k$

A.I.Schlegel Algorithm

The basic idea of the Schlegel algorithm is to adjust the signs of inadequate eigenvalues. Particularly, all negative eigenvalues are changed to their absolute value, except the smallest one [41]. In case that all the eigenvalues are positive, the sign of the smallest eigenvalue is then negated. The quasi-Newton step is subsequently modified as $\Delta \mathbf{x}_k = -\sum_i^d \frac{g_{k,i}}{\gamma'_{k,i}} \mathbf{v}_{k,i}$, $|\gamma'_{k,i}| = |\gamma_{k,i}|$, such that $\gamma'_{k,1} < 0 < \gamma'_{k,i} < \gamma'_{k,i} < \dots < \gamma'_{k,d}$. This, subsequently, leads to reverse the search direction. The resulting direction, however, is not necessarily the converse of the initial, especially if the explored area is distant from an extremum and thus may be erroneous. This issue can be amended by a scaling factor in [42].

In case that $\Delta \mathbf{x}_k$ crosses x_{max} —the maximum allowed step, its length is set to x_{max} , necessitating the addition of a shift parameter λ attained by seeking the extremum of $Q(\Delta \mathbf{x}_k)$ on the sphere with radius x_{max} . In practice, the shift parameter λ is attained by minimizing the Eq. 2.5, below.

$$(||\Delta \mathbf{x}_k(\lambda)|| - x_{max})^2 \quad (\text{Eq. 2.5})$$

The radius $x_{\max}$ is then altered by means of a trust region method [12]. $\Delta \mathbf{x}_k(\lambda)$ is recalculated as $-\sum_{i=1}^d \frac{g_{k,i}}{(\gamma'_{k,i}-\lambda)} \mathbf{v}_{k,i}$. In the trust region method considered in [41], the minimization of (Eq. 2.5) is achieved by finding out the zero of its first derivative.

A.II. Simons Algorithm

Simons et al. [39] proposed a method that uses the augmented quasi-Newton step: $\Delta \mathbf{x}_k = -\sum_i^d \frac{\hat{g}_{k,i}}{(\gamma'_{k,i}-\lambda)} \mathbf{v}_{k,i}$ to walk uphill from a minimum towards a first-order saddle point. Such a walk can be realized by maximizing $Q(\Delta \mathbf{x}_k)$ the along the one eigenvector and minimizing along the others. To achieve that, the parameter λ is chosen so that $\gamma'_{k,1} < \lambda < \frac{1}{2}\gamma'_{k,2}$. This selection ensures that the step has the same direction as the gradient along $\mathbf{v}_{k,1}$ and the converse along all others. Moreover, $Q(\Delta \mathbf{x}_k)$ increases in the direction $\mathbf{v}_{k,1}$ and decreases along the $\mathbf{v}_{k,j}$ directions ($j = 2 \dots d$). Sometimes, such a situation, where $\gamma'_{k,1} > \frac{1}{2}\gamma'_{k,2}$, may result. Hence, a scaling of the coordinate during the walk is completed. According to Simons et al. [39], λ is set to 0 if $\gamma'_{k,1}$ is negative and $\gamma'_{k,2}$ is positive. In case of surpassing the trust radius, the step is scaled down. Otherwise, the parameter is calculated such that $\Delta \mathbf{x}_k$ [39] lies on the border of the trust region. However, there may not exist a suitable parameter λ value such that $\Delta \mathbf{x}_k$ falls on the edges of the trust region. Therefore, Simons et al. suggested the value of λ to be $\frac{1}{2}(\gamma'_{k,1} + \frac{1}{2}\gamma'_{k,2})$. This scheme tries to tune the scaling procedure on the basis of a trust region method. But random selection of λ may result in a truncation of the $\Delta \mathbf{x}_k$ step. Eq. 2.6 given below may thus resolve this issue, providing a framework for calculating, λ .

$$\text{minimize } (||\Delta \mathbf{x}_k(\lambda)|| - x_{max})^2 \quad (\text{Eq. 2.6})$$

$$\text{Subject to} \quad \gamma'_{k,1} < \lambda < \frac{1}{2}\gamma'_{k,2}$$

A.III. Baker Algorithm

Instead of a quadratic model, Baker [38] used the rational model defined as follow:

$$f(\mathbf{x}_k + \Delta \mathbf{x}_k) \approx R(\Delta \mathbf{x}_k) = f(\mathbf{x}_k) + \frac{\mathbf{g}_k^T \Delta \mathbf{x}_k + \frac{1}{2} \Delta \mathbf{x}_k^T \mathbf{H} \Delta \mathbf{x}_k}{1 + \Delta \mathbf{x}_k^T \Delta \mathbf{x}_k} \quad (\text{Eq. 2.7})$$

The augmented quasi-Newton step is then computed as follows:

$$\Delta \mathbf{x}_k = - \sum_i^d \frac{(\hat{g}_{k,i})^2}{(\gamma'_{k,i} - \lambda)} \mathbf{v}_{k,i}, \quad (\text{Eq. 2.8})$$

where $\lambda = \mathbf{g}_k^T \Delta \mathbf{x}_k$, such that $\lambda = - \sum_i^d \frac{(\hat{g}_{k,i})^2}{(\lambda - \gamma'_{k,i})} \mathbf{v}_{k,i}$.

The problem is, thus, broken into two different optimization tasks: 1) a maximization of $R(\Delta \mathbf{x}_k)$ relative to an eigenvector $\mathbf{v}_{k,1}$, with a shift parameter λ_p , and 2) a minimization of $R(\Delta \mathbf{x}_k)$ relative to the other eigenvectors, with a shift parameter λ_n .

B. Minimum-mode Following Methods

Instead of moving the system along all-directions in every iteration, minimum-mode following methods are a class of first-order saddle points finding methods that involves two distinctive steps: 1) finding the minimum-eigen-mode or direction along which the system is moved for the purpose of maximizing $f(\mathbf{x})$ and 2) translating along the minimum mode towards the saddle point. In this category, two methods were proposed so far, namely, Lanczos and Dimer approaches.

B.I. Lanczos Approach

Since the real and symmetric matrices can be converted into tri-diagonal ones, Lanczos method can then be used to find the minimum eigenmode[43]. Since $\mathbf{T} = \mathbf{Q}^T \mathbf{H} \mathbf{Q}$, where $\mathbf{T}$ is a tridiagonal matrix, $\mathbf{H}$ is a Hessian matrix and $\mathbf{Q}$ is an orthogonal matrix [26], it is assured that the eigenpairs $\{\lambda_i, \mathbf{v}_i\}$ of $\mathbf{H}$ are given by $\{\lambda_i, \mathbf{v}_i = \mathbf{Q} \mathbf{v}_i^*\}_{i=1, \dots, n}$, where the superscript * indicate eigenvector obtained from $\mathbf{T}$. Lanczos scheme can be, thus, used to get the lowest eigenpair, $\{\lambda_i, \mathbf{v}_i\}$, through the construction of a set of tridiagonal matrices, $\{\mathbf{T}_1, \dots, \mathbf{T}_j\}$, whose lowest eigenvalue will eventually converge to λ_i as j increases.

The tridiagonal matrix is constructed according to [44-46]

$$\mathbf{T} = \begin{pmatrix} \alpha_1 & \beta_1 & & & & \\ \beta_1 & \alpha_2 & \beta_2 & & & \\ & \beta_2 & \alpha_3 & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \beta_{j-1} \\ & & & & \beta_{j-1} & \alpha_j \end{pmatrix} \quad (\text{Eq. 2.9})$$

where $\alpha_{k=1, \dots, j}$ and $\beta_{k=1, \dots, k-1}$ are calculated iteratively, starting with vector $\mathbf{r}_0$ (for the first iteration, a random vector is chosen, otherwise the eigenvector calculated in the previous iteration is considered), while setting $\beta_0 = \|\mathbf{r}_0\|$ and $\mathbf{q}_0 = 0$. The following steps are recurred ($k = 1 \dots j$)

$$\mathbf{q}_k = \frac{\mathbf{r}_{k-1}}{\beta_{k-1}} \quad (\text{Eq. 2.10})$$

$$\mathbf{u}_k = \mathbf{H} \mathbf{q}_k \quad (\text{Eq. 2.11})$$

$$\mathbf{r}_k = \mathbf{u}_k - \beta_{k-1} \mathbf{q}_{k-1} \quad (\text{Eq. 2.12})$$

$$\alpha_k = \mathbf{q}_k^T \mathbf{r}_k \quad (\text{Eq. 2.13})$$

$$\mathbf{r}_k = \mathbf{r}_k - \alpha_k \mathbf{q}_k \quad (\text{Eq. 2.14})$$

$$\beta_k = \|\mathbf{r}_k\| \quad (\text{Eq. 2.15})$$

For each new set $\{\alpha_k, \beta_{k-1}\}$, the lowest eigenvalue λ_1^{*k} of T_k is calculated once $\left| \frac{\lambda_1^{*k} - \lambda_1^{*k-1}}{\lambda_1^{*k}} \right| < \lambda_{min}$. The eigenvalue is considered converged to $\lambda_1 = \lambda_1^{*k}$, otherwise the same process repeats for n_l^{max} iterations. It is worth noting that the full Hessian is sometimes not available or difficult to calculate. In such case, $\mathbf{u}_k$ in (Eq. 2.11) can be estimated as $\mathbf{u}_k = \frac{\mathbf{g}(\mathbf{x}_{l+1}) - \mathbf{g}(\mathbf{x}_l)}{\Delta x_l}$ with $\mathbf{x}_{l+1} = \mathbf{x}_l + \Delta \mathbf{x}_l \mathbf{q}_k$, where $\mathbf{g}$ is the gradient. $\Delta \mathbf{x}_l$ is the step size vector. Afterwards, the eigenvalues of the matrices can be computed efficiently by an eigenvalue computing algorithm [47]. For complex systems, it is recommended to use an advanced eigenvalue computing algorithm such as the divide-and-conquer method and MRRR (Multiple Relatively Robust Representations) algorithm [47]. Once the lowest eigenvalue converged to $\lambda_1 = \lambda_1^{*j}$, the corresponding eigenvector $\mathbf{v}_1^{*j}$ can be calculated by inverse iteration algorithm.

B.II. Dimer Approach

Dimer approach was introduced by Henkelman and Jonsson [18] for finding saddle points. It is a methodology for following the lowest eigenmode using only the gradient. Structurally, Dimer (see Figure 2-2) has two images ($\mathbf{r}_1, \mathbf{r}_2$). $\mathbf{r}_1 = \mathbf{r} - \Delta r \mathbf{n}$, and $\mathbf{r}_2 = \mathbf{r} + \Delta r \mathbf{n}$. $\mathbf{n}$ is a unit vector and refers to the Dimer direction. $\mathbf{r}$ is the center of Dimer, Δr is the distance between Dimer center $\mathbf{r}$ and $\mathbf{r}_1$. $\overline{\mathbf{f}}_1$, and $\overline{\mathbf{f}}_2$ are the forces or the negative of the gradients at $\mathbf{r}_1$ and $\mathbf{r}_2$, respectively. $\mathbf{f}_1^{\parallel}$ and $\mathbf{f}_2^{\parallel}$ are the components of the forces: $\overline{\mathbf{f}}_1$ and $\overline{\mathbf{f}}_2$ parallel to $\mathbf{n}$, while $\mathbf{f}_1^{\perp}$ and $\mathbf{f}_2^{\perp}$ are the components of the forces perpendicular to

$\mathbf{n} \cdot \mathbf{f}_r = \overline{\mathbf{f}}_1 + \overline{\mathbf{f}}_2$, where $\mathbf{f}_r$ is the total translational force, while $\mathbf{f}^\perp = \mathbf{f}_2^\perp + \mathbf{f}_1^\perp$, where $\mathbf{f}^\perp$ is the total rotation force.

Figure 2-2: Dimer structure as defined by Henkelman, redrawn from [18].

The minimum eigenmode in the Dimer approach is determined by rotating Dimer around its center $\mathbf{r}$ for minimizing rotational forces as the rotational force aligned the Dimer to the direction associated with lowest eigenvalue.

In [18], Henkelman and Jonsson suggested a method for minimizing rotational forces. The method works as follow: 1) Dimer is rotated by small angle $d\theta$ in the direction of Θ (the unit vector in the direction of rotational forces); 2) the curvature (C) is calculated as in (Eq. 2.16) as well the magnitude of the rotational forces as in (Eq. 2.17); 3) $\Delta\theta$ by which the rotational force become zero is determine as in (Eq. 2.17)

$$C = \frac{(\mathbf{f}_1 - \mathbf{f}_2) \cdot \mathbf{n}}{2\Delta r} \quad (\text{Eq. 2.16})$$

$$f' = (\mathbf{f}^\perp \cdot \Theta) / \Delta r \quad (\text{Eq. 2.17})$$

where $\mathbf{f}^\perp$ is the total perpendicular force at $\mathbf{r}_1$; it is also called the rotational force; f' represents the scalar rotation force.

$$\Delta\theta = \left(\frac{1}{2}\right) \arctan(2f_0/f'_0) \quad (\text{Eq. 2.18})$$

where f_0 is the magnitude of the rotational forces and f'_0 is the curvature and both are computed at the current configuration.

Once the minimum eigenmode is determined either by Lanczos iterative or Dimer methods, the system should move along this direction toward the saddle points. The force ($\mathbf{f}^+$) that enforces the system to move can be calculated as follow:

$$\mathbf{f}^+ = \mathbf{f}_R - 2\mathbf{f}^{\parallel} \quad (\text{Eq. 2.19})$$

Such a force is called the effective force. Unlike the total translational forces to pull the system back towards the minimum, the effective force lessen the parallel component force of the total translation force, allowing the system to move towards the saddle point, instead of the minimum point.

2.1.2 Doubly-ended Methods

Doubly-ended methods refer to first-order saddle points finding methods that require two predetermined local minima (reactant and product). There are many doubly-ended methods that will be detailed next.

A. *Drag Method*

The drag method starts by choosing one dimension—called the drag coordinate—to be held fixed during the whole execution of the algorithm. The energy of the system is thus minimized in all other directions. Along the drag coordinate, the system is iteratively

dragged from the reactant to the product in stepwise increments. Once approaching the product, the maximum energy configuration during the execution is taken to be the saddle point. This drag direction can either be intuitively chosen or to be a straight line interpolation between the given two minima [48].

B. Dewar, Healy, and Stewart Method (DHS)

DHS method introduced by Dewar et al. in [49] is a simple method by which two endpoints minima $\mathbf{r}$ and $\mathbf{p}$ are pulled together in order to locate the saddle points. The search procedure of DHS starts by joining $\mathbf{r}$ and $\mathbf{p}$ using a line segment. Then, the lowest energy image or point is chosen to climb the valley by moving the image, with lowest energy, $x\%$ (typically 5%) of the way. The image then relaxes within a fixed distance constraint. By repeating this procedure both images climb the valley until the distance between the both images falls below a given threshold. The main advantage of DHS methods is its capacity to climb the valley and locate the saddle region very quickly. However, it converges very slowly and sometimes both images may end on the same side of the valley leading the search to fail.

C. Conjugate Peak Refinement Method (CPR)

CPR or Conjugate Peak Refinement method is proposed in [50] as an approach to determine multi-dimensional reaction coordinates between two known minima using only the energy function and gradient information. In CPR, a set of images is sampled, one at a time, between the initial and end points: $\mathbf{r}$ and $\mathbf{p}$, after images constitutes a path close enough to the MEP. At the maxima along the MEP lies the saddle point. Each point along the path is sampled in two cycles of line maximization and conjugate gradient

minimization. In first cycle (i.e. line maximization cycle) the maximum along the vector configurations $\mathbf{r}$ and $\mathbf{p}$ is computed and denoted by y_1 . Then, along the direction of each conjugate vector a minimization is carried out, leading a new point $\mathbf{x}_1$. In the second cycle, the maximum along the $\mathbf{r}-\mathbf{x}_1-\mathbf{p}$ path is calculated and denoted by y_2 . Along each of the conjugate vectors, the energy is minimized in order to give a new point that could potentially get incorporated in the path. Both cycles are repeated until a minimum along the path has a smaller gradient than the pre-specified tolerance for saddle points.

D. Nudged Elastic Bands (NEB)

NEB method was introduced by Jonsson et al in [51, 52]. In the NEB method, a group of images $[\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3, \dots, \mathbf{i}_{n-1}, \mathbf{i}_n]$ of the system are created, connecting (in spring manner) between two minima. The NEB procedure consists of two steps: the first is the creation of images for example by linear interpolation between the both minima; the second is to apply an optimization procedure to relax images toward the minimum energy path. This can be achieved by minimization of the spring force which is represented as follows:

$$s(\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3, \dots, \mathbf{i}_{n-1}, \mathbf{i}_n) = \sum_{i=1}^{n-1} f(\mathbf{i}_i) + \sum_{i=1}^n \frac{k}{2} (\mathbf{i}_i - \mathbf{i}_{i-1}) \quad (\text{Eq. 2.20})$$

$f(\mathbf{i}_i)$ is the energy at the image $\mathbf{i}_i$ and k is the spring constant.

Minimization of (Eq. 2.20) will eventually lead the spring to pass through. However, the direct usage of spring forces in the NEB method will eventually cause two problems: 1) corner cutting which occurs as a result of the perpendicular component of the spring forces to the MEP; 2) sliding down to the end points as a result of the parallel component

of the actual forces which leads the resolution of the saddle point region to be very low. In order to overcome those problems, instead of using both the true force and spring force, Jonsson et al. used only the parallel component of the spring force and the perpendicular component of the true force. The total force of those components is represented in [48].

$$\mathbf{g}(\mathbf{i}_i)^{\perp} = \mathbf{g}(\mathbf{i}_i) - [\mathbf{g}(\mathbf{i}_i) \cdot \mathbf{t}] \mathbf{t} \quad (\text{Eq. 2.21})$$

$$\mathbf{f}_i^s = k(|\mathbf{i}_{i+1} - \mathbf{i}_i| - |\mathbf{i}_i - \mathbf{i}_{i-1}|) \quad (\text{Eq. 2.22})$$

$$\mathbf{f}_i = -\mathbf{g}(\mathbf{i}_i)^{\perp} + (\mathbf{f}_i^s \cdot \mathbf{t}) \mathbf{t} \quad (\text{Eq. 2.23})$$

where $\mathbf{g}(\mathbf{i}_i)$ is the gradient at the image i and $\mathbf{t}$ is tangent along MEP which will explain next. $\mathbf{f}_i^s$ is the spring forces which used here to assure that the spaces between two neighboring images are equal-sized. k is the spring constant and $\mathbf{i}_i$ is the i^{th} image of the system.

Climbing Image Nudged Elastic Bands method (CI-NEB)

CI-NEB is a modified version of NEB method suggested in [51]. In this method, after a few relaxation steps, the highest energy image is selected to climb the energy landscape along the elastic band. This can be achieved by setting the string forces on the image to be zero and inverting the parallel component of the true forces according to the following equation:

$$\mathbf{f}_{i,max}^{climb} = \mathbf{g}(\mathbf{i}_{max,i}) \cdot \mathbf{t} \|\mathbf{t}\| \quad (\text{Eq. 2.24})$$

Unlike the drag method, in CI-NEB method, the climbing image is pulled uphill in the direction calculated using the adjacent images in the elastic band. In addition, the tangent to the path is biased by the energy of the neighboring images.

The procedure of the CI-NEB method using three movable images is summarized as follow: 1) three movable images are created between the end points: reactant $\mathbf{r}$ and product $\mathbf{p}$; 2) the central image (the one associated highest energy) climbs the hill; 3) the other two images move according to the nudging algorithm. Once converged, a rough depiction of the MEP has been achieved and one of the images lies within a prescribed threshold of the saddle point. The main advantage of this algorithm, in addition to its capability to be efficiently parallelized as well as NEB, is that the movable images can be easily tuned concurrently [51].

2.2. Stochastic First-order Saddle Points Finding Methods

Unlike the conventional optimization methods where the objective function is assumed to be precise and the computation is entirely established on values attained so far, stochastic optimization approaches refer to algorithms that consider probabilistic random elements. These random elements can be either in the problem data, including the objective function or constraints, or in the algorithm itself, through random parameter values, random choices or both. There exist many stochastic optimization methods in the literature such as simulated annealing and evolutionary algorithms. Here, we present some stochastic optimization algorithms that have been proposed in the literature for tackling the problem of locating transition states.

2.2.1 Simulated Annealing for First-order Saddle Point Optimization

Chaudhury and Bhattacharyya proposed a simulated annealing based method for locating first-order saddle points that can proceed sequentially, locating all the saddle points, and also adjusted to trace the lowest first-order saddle point lying between local minima [20].

To locate first-order saddle points, Chaudhury proposed the following objective function:

$$f^*(\mathbf{x}) = e^{-((f(\mathbf{x}) - e_g)^2 + d\beta\|\mathbf{g}\| + \sum_{i=1}^{n-6} \xi_i e^{-p_i \gamma_i})} \quad (\text{Eq. 2.25})$$

where $\mathbf{g}$ is the gradient, γ_i is the i^{th} eigenvalue of Hessian matrix $\mathbf{H}$. p_i 's and ξ_i 's are the penalty factors for the slope and curvature constraints to be forced.

For the search to start, the following configurations are considered: $e_1 = e_0 + \varsigma$, where ς is recommended to be proportional to $[|f(\mathbf{p}) - f(\mathbf{r})|]$, e_0 is the highest among the energies of the reactant ($\mathbf{r}$) and the product ($\mathbf{p}$); and (ξ_i) s are assigned with small positive values. One coordinate (say, x_i) is then randomly chosen to be mutated as $x_i^{g+1} = x_i^g + (2r - 1)s_i$, where r is a random number $r \in [0, 1]$ and s_i is small step length parameter ($s_i > 0$). Thereafter, $f(\mathbf{x}^{g+1})$, $\mathbf{g}_i$, and λ_i s are computed. If $f^*(\mathbf{x}^{g+1}) < f^*(\mathbf{x}^g)$, the new configuration is then accepted. Otherwise, the new configuration undergo another test: a Metropolis sampling test – if $e^{\frac{[f(\mathbf{x}^{g+1}) - f(\mathbf{x}^g)]}{Kt_i}} > r$, where t_i and K are the temperature and the Boltzmann constant, respectively. Otherwise, the new configuration is repudiated.

During the search, a configuration with one negative eigenvalue may be uncovered. In such case, the contribution from the curvature significantly reduces as ξ_k alters from positive to negative sign. Now, the minimization process of $f^*(\mathbf{x})$ is controlled by the k^{th} curvature and therefore enforced to continue in the direction of k^{th} eigenvector of $\mathbf{H}$. The search carries on along this orientation as long as λ_k continues to be negative and the number of samples does not exceed a maximum pre-specified number. Once one of these criteria is satisfied and $f^*(\mathbf{x}) > Kt_i$, the parameter temperature is then lower to $t_{i+1} = t_i * w$ ($w = 0.9$) and e_g is updated as $e_{g+1} = e_g + \eta [f^*(\mathbf{x}^g)]^{0.5}$, where $\eta \in [0, 1]$, and if e_g is not already close to the current value of $f^*(\mathbf{x}^g)$. Otherwise, e_g is reset $e_{g+1} = e_g + \varsigma$. If an adequate minimum is not attained even after several temperature changes, or two or more negative eigenvalues are came across, the process repeats until an acceptable minimum is reached.

2.2.2 Evolutionary Optimization Methods

Evolutionary algorithm or EA is population-based stochastic search method, where a population of individuals probabilistically roams the problem search space. By analogy of natural evolution of species, individuals evolve through various genetic operators to better solutions. These operators combine and mix the genetic information of the parents to generate offspring. The goal is to create offspring which are fitter than the parents. The fitness of an individual is often defined numerically or otherwise. The process continues for a number of generations until the specified convergence criteria are satisfied. Evolutionary algorithms have been used extensively for optimizing complex systems. For instance, a motivating example for us is the optimization of optimal chemical structures

[6-13]. In this section, we present the evolutionary algorithms proposed so far in the literature to tackle the problem of finding first-order saddle points or transition state structures. To date, two different evolutionary methods have been published. These algorithms includes the Bungay genetic algorithm [19] (Subsection A.) and the Chaudhury genetic algorithm [21] (Subsection B.).

A. *Bungay Genetic Algorithm*

Bungay method starts by encoding the initial population individuals into binary strings. In this process, each variable is encoded into a binary string and concatenated to other to form a chromosome. The initial population individuals are generated by adding small random perturbations to an initial approximation of the solution. These perturbations are restricted to a user defined region. In this method, genetic operators are applied to the binary-encoded chromosomes, whereas the fitness function is applied on real-coded numbers encoded in the binary chromosomes. Thus, decoding and encoding are necessary [19].

To bias search to the first-order saddle point, Bungay et al. suggested the following fitness function:

$$f^*(\mathbf{x}) = \frac{1}{\|\mathbf{g}\| + \varepsilon} \cdot \frac{1}{(n-1) + \varepsilon} \quad (\text{Eq. 2. 26})$$

where $\|\mathbf{g}\|$ is the norm of the gradient vector, n is the number of negative eigenvalues of the Hessian matrix, and ε was chosen small to prevent division by zero error.

Parents are selected using the roulette wheel selection where the probability of an individual to be selected is proportional to the individual fitness. Selected parents undergo mutation or crossover. In crossover, two selected parent are aligned together. For the case of one-point crossover, at a random point, half of the chromosomes are exchanged, creating new offspring. In case of mutation, a random bit is chosen to be flipped by a predefined probability. Best solutions in the current generation are kept for the next generation, which is called elitism. This cycle is repeated until the given stopping criteria are satisfied. These criteria may include a maximum number of generations or a certain number of solutions attained.

B. Chaudhury GA Method

Chaudhury, Bhattacharyya, and Quapp also proposed a genetic algorithm based technique that only uses the energy and gradient information to steer the search in the direction of a first-order saddle point [21]. Following the basic genetic algorithm life cycle of selection-crossover-mutation, Chaudhury genetic algorithm starts by adding a small random perturbation to the global optimum. Each population individual is evaluated using (Eq. 2.25). Once the population of individuals is already evaluated, they undergo the selection process, where fitter individuals have higher chances to undergo mutation and crossover, in forming the offspring. Selected individuals then undergo stochastic crossover and mutation.

Chaudhury et al. uses the single-point crossover operator with some modifications. To explain, let us assume that $\mathbf{x}_1 = (x_1^1, x_1^2; x_1^3; \dots; x_1^z; x_1^{z+1}; x_1^{z+2}; \dots; x_1^d)$ and $\mathbf{x}_2 = (x_2^1, x_2^2; x_2^3; \dots; x_2^z; x_2^{z+1}; x_2^{z+2}; \dots; x_2^d)$ are the pair of individuals of interest in the current

generation. The selected pair is aligned together. At a random crossover site (t), the two solutions exchange their genetic material as follows:

$$\begin{aligned}\mathbf{x}_1^{c'} &= r\mathbf{x}_1^c + (1-r)\mathbf{x}_2^c \\ \mathbf{x}_2^{c'} &= r\mathbf{x}_2^c + (1-r)\mathbf{x}_1^c\end{aligned}\tag{Eq. 2.27}$$

where $c = z + 1, \dots, n$, $0 < r < 1$, and r is a randomly chosen number. The solution set then undergoes the crossover and mutation operation, which introduces only subtle changes to the chromosomes. Thereafter, individuals are randomly chosen to undergo mutation operator. A chosen individual $\mathbf{x}_k$ is then mutated as $\mathbf{x}'_k = \mathbf{x}_k + (-1)^l \mathbf{r} \vartheta$, where l is a random integer, $\mathbf{r}$ is a random vector, and ϑ is a mutation amplitude which is a randomly chosen number. These operations accomplish one generation in Chaudhury GA. The new population undergoes the same sequence of GA operations again, and this process is recurred until a maximum fitness is achieved or a maximum number of generations is exceeded.

2.3 Discussions

This next section involves a detailed discussion of different methods proposed for locating first-order saddle points categorized by their strategies. Comparisons between different strategies are also considered. Some recommendations for future research investigation are also introduced.

This is composed of two major subsections. Subsection 2.3.1 involves a detailed discussion of different conventional methods proposed for finding first-order saddle points, whereas Subsection 2.3.2 involves detailed discussion of different stochastic

methods. Comparison between stochastic and conventional methods is also considered in subsection 2.3.3.

2.3.1 Discussion on Conventional Methods

In this section, we compare and discuss different conventional methods proposed in literature thus far for locating first-order saddle points. Firstly, we discuss and compare and contrast the different single-ended methods over different problems of varying complexity in terms of search efficiency and number of uncovered saddle points. The doubly-ended methods are next discussed. A critical comparison among the two major strategies (doubly-ended and single-ended) is also considered. Some recommendations are then reported for further research investigation.

A. Single-ended Methods

Olsen et al. compared the all-mode and minimum mode following methods, considering both exact and approximate Hessians, on a test problem involving transitions in a seven-atom platinum (Pt) island on a Pt(111) surface, where Morse pair-wise potential is used for energy calculation. They consider three systems of varying sizes 1, 7, and 175 atoms of 3, 21 and 525 degrees of freedom, respectively. Their study indicated that if one attempts to make a map of all saddle points for a complex system, the minimum mode-following methods are recommended in particular if the gradient can be cheaply calculated. The all-mode following methods employing an approximate Hessian represent, otherwise, the preferred choice more particularly if a good estimation for the Hessian at the initial search position can be attained at a low cost. All-mode following methods with an approximate Hessian are also favored, for small and medium sized

systems, in case that the gradient is costly to compute. Otherwise, the all-mode following methods using the exact Hessian represent the ideal choice [13].

Besides Olsen's comparison study, we have also summarized the pros and cons of single-ended methods in Table 2-2. According to the table, all single ended methods have the drawback which necessitates the existence of a good initial guess to work. At the level of strategy, the minimum mode following methods are computationally-demanding and converge to low-quality solutions. Besides, they have the drawbacks of converging not only to high-energy saddle points but also to the higher-order ones. Having these drawbacks may give them the advantage of locating more saddle points than other compeers as highlighted by Olsen.

Adapting the other strategy in single-ended methods, all-mode-following methods have the drawback of converging to low stream-bed saddle points. Besides this disadvantage, Simons and Baker's algorithms are complex-to-implement and computationally-demanding. Moreover, they converge slowly when compared to Schlegel's algorithm. Schlegel's algorithm, on the contrary, is easy to implement and converge efficiently to high-quality saddle points.

Table 2-2: Comparison among single-ended methods.

		Properties	Advantages	Disadvantages
All-mode following methods	Schlegel Algorithm	Mathematical model based on Quadratic model (Eq. 2.4:5)	<ul style="list-style-type: none"> – simple – easy to implement – the fastest among all-mode following methods – very precise solutions 	<ul style="list-style-type: none"> – Lowest stream bed – Good initial guess
	Simons Algorithm	Mathematical model based on Quadratic model (Eq. 2.4 and Eq. 2.6)	<ul style="list-style-type: none"> – Fast – Precise solutions 	<ul style="list-style-type: none"> – Complex procedure – Computationally demanding – Lowest stream bed – Good initial guess
	Baker Algorithm	Mathematical model based on rational model (Eq. 2.7:8)	<ul style="list-style-type: none"> – Fast – Precise solutions 	<ul style="list-style-type: none"> – Slow convergence – Complex procedure – Computationally demanding – Lowest stream bed – Good initial guess
Minimum-mode following methods	Lanczas	Based on the mathematical model reported in Eq. 2.9:15	<ul style="list-style-type: none"> – Converge to more saddle points, but of high energy 	<ul style="list-style-type: none"> – Computationally demanding – Converge to higher order saddle points – Low precise solutions
	Dimer methods	Based on the Dimer model reported in Figure 2-2 and Eq. 2.16:19	<ul style="list-style-type: none"> – Converge to more saddle points, but of high energy 	<ul style="list-style-type: none"> – Computationally demanding – It may experience the minimum curvature or c not turning negative, leading the Dimer to continue climbing the surface or landscape. – Converge to higher order saddle points – Low precise solutions

B. Doubly-ended Methods

In this subsection, we discuss different doubly-ended methods considering the pros and cons of each method, summarized in Table 2-3 and explained below.

Drag method adapts a strategy that uses interpolations and step-wise increments. It has the advantage of being simple and easy-to-implement. However, it has several drawbacks, including convergence to low-quality saddle points very slowly.

DHS, in brief, consists of two stages where the system climb the valley using two images, moving the image with the lowest energy image and let it to relax within a fixed distance. It has the advantage of climbing the valley efficiently and locates saddle regions quickly. Despite the noticeable advantage of locating the saddle regions quickly, it hardly converges to saddle points. Moreover, the identified points are of low-quality. Besides these drawbacks, DHS may experience the case in which the two images used to locate saddle points may end up at the same side of the valley leading the search to fail.

Conjugate peak refinement (CPR) method based on two cycles of line maximization and conjugate gradient minimization. The main advantage of CPR is that it converges quickly to the saddle point especially when it is sufficiently close. However, as doubly ended method, it requires an initial approximation of the minimum energy path and converges to low quality saddle points. According to Henkelman [48], CPR is the most un-reliable doubly-ended method. It has failed seven out of ten times on Morse Potential. It is also the most difficult method to implement, because of the complex rules needed to accept or reject new points on the path.

Nudged Elastic Band (NEB) build based on string model reported in Eq. 2-20:23. It converges fast to the attraction region of the saddle points and can be efficiently

parallelize. It still however converges slowly and requires computationally demanding as well as a good tangent estimation.

Climbing Image Nudged Elastic Band method (CI-NEB) uses the string model reported in Eq. 2-20:23 considering setting the string forces on the image to be zero by inverting the parallel component of the true forces according to Eq 2.24. The main advantage of this method, among doubly ended methods, it uses only three images to explore the minimum energy path. It can be parallelized as efficiently as NEB and converge very fast to the saddle region. However, it still has the same issue of NEB: the need of good estimation of the tangent and the low quality of saddle points. Among all doubly-ended methods, CI-NEB method was reported to be the most reliable, attaining all the saddle points between **p** and **r**, with reasonable computational cost.

All the doubly-ended methods have an issue that necessitates two minima and sometimes a good initial guess of the minimum energy path to be available at the beginning of the execution.

Table 2-3: Comparison among doubly-ended methods.

Doubly Ended Method	Properties	Advantages	Disadvantages
Drag method	Interpolation and step-wise increment.	<ul style="list-style-type: none"> – Easy to develop. 	<ul style="list-style-type: none"> – Requires two minima: reactant and product – Slow strategy of step-wise increments – Saddle points identified after the whole execution – Initial guess of the minimum energy path is needed. – Converge to low quality saddle points
Dewar, Healy, and Stewart (DHS)	Climbing the valley by moving the image, with lowest energy, x% (typically 5%) of the way and relaxing the image within a fixed distance constraint.	<ul style="list-style-type: none"> – Its capacity to climb the valley and locate saddle region very quickly 	<ul style="list-style-type: none"> – Requires two minima: reactant and product – Initial guess of the minimum energy path is needed. – Identifying Optima X% for climbing image – Converges very slowly – The two images used to locate saddle points may end at the same side of the valley leading the search to fail. – Converge to low quality saddle points
Conjugate Peak Refinement Method	Based on two cycles of line maximization and conjugate gradient minimization.	<ul style="list-style-type: none"> – Converges quickly to the saddle point, especially when it is sufficiently close. 	<ul style="list-style-type: none"> – Requires two minima: reactant and product – Initial guess of the minimum energy path is needed. – Sampling a number of images that resemble the minimum energy path – Converge to low quality saddle points – It is also the most difficult method to implement, because of the complex rules needed to accept or rejecting new points on the path.
Nudged Elastic Bands	Use string model, reported in Eq. 2.20:23	<ul style="list-style-type: none"> – Converge fast to the attraction region of saddle points. – Can be efficiently parallelized 	<ul style="list-style-type: none"> – Requires two minima: reactant and product – Initial guess of the minimum energy path is needed. – Sampling a number of images that resemble the minimum energy path – Converge to low quality saddle points – Computationally demanding – Required good tangent estimation.
Climbing Image Nudged Elastic Band Method	Use string model, reported in Eq. 2.20:23, considering setting the string forces on the image to be zero and this can be achieved by inverting the parallel component of the true forces according to Eqn.2.24	<ul style="list-style-type: none"> – Use only three images to explore the landscape. – Can be efficiently parallelized – Converge fast to the attraction region of the saddle points. 	<ul style="list-style-type: none"> – Requires two minima: reactant and product – Initial guess of the minimum energy path is needed. – Sampling a number of images that resemble the minimum energy path – Converge to low quality saddle points – Computationally demanding – Required good tangent estimation.

2.3.2 Discussion on Stochastic Methods

Stochastic search methods proposed so far include simulated annealing methods and evolutionary algorithms. Simulated annealing method proposed by Chaudhury et al. [20] has been applied to different small systems varying from 2 to 18 dimensions, namely, the Cerjan-Miller potential [53], Culot-Dive-Nguyer-Chuysen potential, Halgren-Lipscomb potential, Hoffman-Nord-Redenberg potential, and 6-atom Lennard-Jones Potential. Without explicit numerical comparison, the proposers alleged that the performance of the simulated annealing method would outperform some of doubly-ended method such as drag method since there is no any constraint to select any specific coordinate, and the simulated annealing method computational resources needed are similar to that of the gradient norm minimization method. However, the simulated annealing method, as highlighted in Table 2-4, is plagued with several pitfalls, for example the need for calculating the second derivative and eigenvalues. In addition, this method requires the information on reactant or product. Last but not the least, it may converge to higher order saddle points.

On the other hand, evolutionary algorithms proposed for tackling the first-order saddle point problem includes: Chaudhury and Bungay GAs. Bungay GA has been applied on mathematical problems such as Chong-Zak function and chemical reactions including HNC isomerization. The main advantages of these methods, as highlighted in Table 2-4, are the ability to globally search the landscapes, and convergence to multiple saddle points. Proposers alleged that the methods do not require any information about the landscapes; however it is required to have an initial approximation of the saddle points

sought around which the population would be initialized. This method seems to be much more expensive than conventional methods. In addition, it requires calculation of the number of negative eigenvalues which in turn requires calculation and diagonalization of the Hessian matrix. This method may also converge to higher order saddle points. It may maintain multiple solutions but it is not assured that these solutions would survive in the whole execution.

Chaudhury et al. proposed another method which employs the genetic algorithm to find first-order saddle points using only the energy and gradient information [21]. It has been applied to Lennard Jones potential of complex systems ranging between 7 and 30 atoms. Overcoming the need for the second derivative was a great step towards targeting complex systems. However, Chaudhury method requires the search to start from the global optimum configuration which might be difficult or impossible to know or hard to optimize in advance especially on complex systems. In addition, the trajectory nature of Chaudhury GA limits the parallelism of genetic algorithms.

Chaudhury et al. [21] compared their methods (SA and GA). Their study showed that GA methods generally outperformed the SA method. As the complexity of system increases, SA, in particular, hardly converges to saddle points. This result totally agrees with many studies that compare the performance of the evolutionary algorithms with simulated annealing methods [54-57].

All the stochastic methods proposed so far requires either the global minimum configuration or an initial approximation of the first-order saddle points, or reactant and

products. Moreover, some of the methods requires second derivatives and may converge to higher order saddle points. Most of the stochastic methods are also quite computationally expensive compared to conventional methods.

2.3.3 Comparison between Stochastic and Conventional Methods

In this chapter, both conventional and stochastic first-order saddle point optimization methods have been briefly discussed. A simple comparison between the two classes is considered here to highlight the major issues of both methods to be considered in the study.

On one hand, conventional saddle points finding methods, which have taken the main focus of saddle point optimization search, have shown to converge quickly and precisely. Some of the methods are also capable of identifying the minimum energy path. Stochastic saddle points finding methods have also shown great abilities to locate saddle region efficiently, even if the search starts far from the saddle points. Stochastic saddle points finding methods are also capable of dealing with rugged and noise-imbued complex landscapes.

On the other hand, conventional methods fail to converge to saddle points if the search starts far apart from regions where the saddle points actually lie. The stochastic saddle point finding methods also converge to low precision solutions. Both methods lack the ability to locate and maintain multiple saddle points. In addition, both assume the availability of landscape information such as an initial guess of the first-order saddle point and global optima, and may however converge to higher order saddle points.

Table 2-4: Comparison among stochastic methods.

		Properties	Advantages	Disadvantages
Evolutionary Algorithms	Chaudhury GA	Evolutionary model, using fitness function in Eq. 2.25 and Eq. 2.27, linear crossover and mutation	<ul style="list-style-type: none"> – May locate multiple saddle points. – Explore the search space globally. – Ability to be parallelized 	<ul style="list-style-type: none"> – Converge slowly – Low-quality saddle point – Require global optima – Exploring search space in a step-wise manner, limiting the flexibility in exploring search space. – May converge to higher order saddle points
	Bungay GA	Evolutionary search model, using fitness function in Eq. 2.26.	<ul style="list-style-type: none"> – May locate multiple saddle points. – Explore the search space globally. – Ability to be parallelized 	<ul style="list-style-type: none"> – Converge slowly – Low-quality saddle point – Requires an initial guess of the saddle point – May converge to higher order saddle points – May converge to high energy saddle points – Requires calculation of the number of negative eigenvalues which in turn requires calculation and diagonalization of the Hessian matrix
Simulated Annealing methods	Chaudhury SA	<ul style="list-style-type: none"> – Simulated annealing search model 	<ul style="list-style-type: none"> – May locate multiple saddle points. – Explore the search space globally. 	<ul style="list-style-type: none"> – Requires two minima: reactant and product – Converge slowly – Low-quality saddle point – May converge to higher order saddle points – May converge to high energy saddle points – Calculating the second derivative and eigenvalues

To sum up, both methods have some advantages over the others, but the ability to converge to saddle region fast and to deal with complex landscape give the stochastic saddle points finding methods an edge over the existing conventional compeers. However, many issues should be considered in stochastic methods such as the ability of discovering and maintaining multiple saddle points as well as the low precision of the uncovered saddle points. Further investigation on stochastic methods will be considered next in the methodology part.

PART I: Methodology

As discussed previously in Chapter 1 and 2, first-order saddle points are challenging to attain. They lie neither at the maxima nor at the minima. They have the characteristic of both, maximum on only one dimension and minima on all other dimensions. Many conventional and stochastic methods have been proposed in the last decades. Existing methods proposed thus for locating saddle points, especially the stochastic search ones, have, however, difficulty locating multiple saddle points both efficiently and precisely. To verify these challenges, this part includes preliminary investigation on locating multiple saddle points, using a simple form of stochastic search methods, entitled “multi-start local search” and a number of multi-modal benchmark problems of varying complexity in Chapter 3. Experimental studies on the multi-start local search and several methods proposed in the literature have highlighted the limited capabilities of existing conventional and stochastic methods in maintaining multiple saddle points. Less than 30% of the saddle points have been identified across different problem landscapes.

To identify more than saddle points both efficiently and precisely, we propose, in Chapter 4, a multi-modal memetic framework, especially designed to locate multiple high-quality saddle points. The proposed framework includes a fitness function that transforms the problem landscape such first-order saddle points lie at the optima, ranked according to their original fitness values, on the transformed landscape. The framework encompasses evolutionary operators for globally exploring the transformed landscape. The framework also includes memetic operators to search the original landscape locally and locate high quality saddle points with the minimal computational cost. To locate and maintain multiple saddle points, the framework has a niching operator called valley-

adaptive clearing scheme. The framework has successfully attained more than 90% of saddle points with accuracy up to 10^{-10} gradient norm precision.

Chapter 3: Stochastic Multi-start Local Search for Locating Multiple Saddle Points

Optimization problems have attracted a lot of attention in the last few decades. Particularly, locating global and local optima have attracted the major attentions whereas little attention has been given to the saddle point search. In the context of saddle point optimization, on the other hand, there remains still limited attention and progress made in stochastic optimization, relative to the conventional approaches. In this chapter, an initial investigation on a simple stochastic optimization of saddle points serving as a baseline algorithm is introduced.

The random multi-start local search represents one of the simplest stochastic approaches that have been widely used as an optimization method for both global and local optimization. The basic idea of this method is that a set of random points sampled from the function domain are allowed to undergo local search to arrive at a set of local optima. The approach has shown to perform reasonably well. More importantly, it is very easy to implement and hence considered in this study as an initial investigation for locating multiple first-order saddle points.

In this chapter, a stochastic multi-start local search (SMLS) designed for converging to multiple saddle points, as formalized in Chapter 2, is introduced in Section 3.1. To verify the efficacy of the proposed SMLS, experimental studies have been conducted and the results are reported in Section 3.2. A brief conclusion is then stated in Section 3.3.

3.1 *Stochastic Multi-Start Local Search (SMLS)*

In this section, the stochastic multi-start local search for locating first-order saddle points (see Table 3-1) is introduced. In the algorithm, a population of individuals is first randomly generated. After the initialization process, individuals then undergo the local search. Due to its efficacy in achieving quick and precise convergence, a single-ended saddle point searcher is considered. Individuals undergo the single-ended saddle point searcher until first-order saddle point convergence is achieved or the maximum number of iterations allowable is exceeded. All discovered first-order saddle points that satisfy a given precision criterion are subsequently archived. The process repeats until the maximum number of generation is reached.

Table 3-1: A stochastic multi-start local search.

```

1   $X_T = \{\}$ 
2  while Stopping conditions are not satisfied do
2.1 Initialize: generate a new population  $P$  randomly ;
2.2 For  $x \in P$ 
2.2.1  $x' = \text{Apply Single-ended Saddle Point Searcher on } x$ 
2.2.2 if  $\frac{\delta f(x')}{\delta x'} < \delta$ 
2.2.2.1  $X_T = X_T \cup \{x'\}$ 
2.2.3 End if
2.3 End For
3  End while

```

3.2 *Experimental Study*

In this section, the experimental settings and results of the stochastic multi-start local search algorithm (SMLS) are reported. The stochastic multi-start local search algorithm is implemented on MATLAB development environment, executed on a PC with 2.66 GHz Intel Duo Core CPU and 3 GB RAM. A population of 10 individuals is considered. Individuals then undergo local refinements until the convergence criterion has been fulfilled. The search terminate if a maximum computational budget of 10^6 gradient calls

has been incurred. Average results taken across 50 independent runs are reported. The benchmark test problems considered in the study are detailed in subsection 3.2.1. Performance analysis of the stochastic multi-start local search is then presented in subsection 3.2.2 .

3.2.1 Benchmark Problems

To visualize the first-order saddle points and test the performance of SMLS, a set of four commonly used benchmark problems in the context of global optimization are considered here. These problems are scalable, giving the flexibility to increase or decrease the domain ranges and the number of decision variables of interest. The landscape of all problems considered in the present study contains a significant number of critical points including maxima, minima and saddle points. These test problems are detailed in what follows:

Problem 1: The Sines Function

Sines function defined in (Eq. 3.1) has a multi-modal landscape with approximately 84 first-order saddle points¹ within the range of $[-10, 10]$. Some first-order saddle points uncovered by the stochastic multi-start local search are depicted in Figure 3-1.

$$f(x, y) = 1 + \sin^2(x) + \sin^2(y) - 0.1e^{-x^2-y^2} \quad (\text{Eq. 3.1})$$

where $x, y \in [-10, 10]$

Problem 2: The Multi-function

¹ The number of 1st order saddle points is approximated by means of visualization.

Multi-function defined in (Eq. 3.2) represents a challenging 2D test function for saddle point algorithms because its fitness landscape involves 144 saddle points within the small range of $[-2, 2]$. Some first-order saddle points uncovered by the stochastic multi-start local search are visualized in Figure 3-2.

$$f(x, y) = -1 - x \sin(4\pi x) + y \sin(4\pi y + \pi) \quad (\text{Eq. 3.2})$$

where $x, y \in [-2, 2]$

Problem 3: Rastrigin Function

Rastrigin function defined in (Eq. 3.3) represents another problem with a highly multimodal landscape. For the range specified (Eq. 3.3), the landscape of the 2D Rastrigin function contains one global optima lying at $(0, 0)$ and 223 first-order saddle points. Some first-order saddle points uncovered by using SMLS are visualized in Figure 3-3.

$$f(\mathbf{x}) = 10n + \sum_{i=1}^n x_i^2 - 10 \cos(2\pi x_i) \quad (\text{Eq. 3.3})$$

where $x_i \in [-5, 5]$

Problem 4: Schwefel Function

Schwefel function defined in (Eq. 3.4) denotes one of the other widely used test function in global optimization. It is composed of a large number of valleys and peaks. The distances between the global and local optima are also significant large, causing many search algorithms to be trapped at local optima. The 2D landscape in the range specified in (Eq. 3.4) contains about 66 first-order saddle point reported in Figure 3-4.

$$f(\mathbf{x}) = (\sum_{i=1}^n -x_i \sin(\sqrt{|x_i|})) \quad (\text{Eq. 3.4})$$

where $x_i \in [-350, 350]$

First-order Saddle points (x,y) visualized in the figure above	$\ g\ $	$f(x, y)$
(-0.000000000000, 1.583692845456)	7.632786E-017	1.991691271273
(0.000000000000, -1.583692845456)	7.727025E-017	1.991691271273
(1.583692845456, -0.000000000000)	7.632785E-017	1.991691271273
(-1.583692845456, 0.000000000000)	7.632786E-017	1.991691271273

Figure 3-1: 2D Sines function landscape with first-order saddle points tabulated and highlighted by white dots in the range of $[-2, 2]$.

First-order Saddle points (x,y) visualized in the figure above	$\ g\ $	$f(x,y)$
(0,-0.161443419136)	2.355139E-016	-1.144807581839
(0,0.161443419136)	2.355139E-016	-1.144807581839
(0.161443419136,0)	2.355139E-016	-1.144807581839
(-0.161443419136,0)	2.355139E-016	-1.144807581839
(0.161443419136,0.390978476619)	3.942464E-015	-0.761684241182
(0.161443419136,-0.390978476619)	2.899717E-015	-0.761684241182
(-0.161443419136,0.390978476619)	2.355139E-016	-0.761684241182
(-0.161443419136,-0.390978476619)	3.942464E-015	-0.761684241182
(0.390978476619,-0.161443419136)	3.689717E-015	-0.761684241182
(0.390978476619,0.161443419136)	4.342661E-015	-0.761684241182
(-0.390978476619,-0.161443419136)	3.689717E-015	-0.761684241182
(-0.390978476619,0.161443419136)	4.553268E-015	-0.761684241182

Figure 3-2: 2D Multi-function landscape with first-order saddle points tabulated and highlighted by white dots in the range of [-0.5, 0.5].

First-order Saddle points (x,y) visualized in the figure above	$\ g\ $	$f(x, y)$
(0.000000000000, 0.502546036555)	8.882201E-015	20.251272990990
(-0.000000000000, -0.502546036555)	8.882299E-015	20.251272990990
(-0.502546036554, -0.000000000000)	8.882126E-015	20.251272990990
(0.502546036554, -0.000000000000)	8.882126E-015	20.251272990990

Figure 3-3: 2D Rastrigin landscape with first-order saddle points tabulated and highlighted by white dots in the range of $[-0.75, 0.75]$.

First-order Saddle points (x,y) visualized in the figure above	$\ g\ $	$f(x, y)$
(5.23919930019552, -5.23919930019552)	3.14E-16	0
(-5.23919930019552, -25.8774173476187)	5.98E-16	-20.1377
(-25.8774173476187, 25.8774173476187)	7.11E-16	0
(25.8774173476187, 5.23919930019552)	1.14E-15	20.13766
(5.23919930019552, 25.8774173476187)	3.05E-15	20.13766
(-25.8774173476187, -5.23919930019553)	3.13E-15	-20.1377
(65.5478650901515, 25.8774173476187)	3.70E-15	-39.552
(-25.8774173476187, -65.5478650901515)	4.24E-15	39.55202
(-65.5478650901515, 5.23919930019553)	4.26E-15	-59.68968
(65.5478650901515, -65.5478650901515)	4.40E-15	0
(-65.5478650901515, 65.5478650901515)	4.40E-15	0
(25.8774173476187, 65.5478650901515)	4.46E-15	-39.552
(5.23919930019553, -65.5478650901515)	4.61E-15	-59.68968
(-65.5478650901515, -25.8774173476187)	5.11E-15	39.55202
(65.5478650901516, -5.23919930019553)	1.13E-14	-59.6897
(-5.23919930019553, -65.5478650901516)	1.13E-14	-59.6897

Figure 3-4: 2D Schwefel function landscape with first-order saddle points tabulated and highlighted by white spots in the range of $[-75, -75]$.

3.2.2 Performance Analysis

In this subsection, the performance of SMLS in terms of the percentage of uncovered first-order saddle points, gradient precision accuracy, and convergence speed are reported and analyzed.

A. *Precision Accuracy of Uncovered First-Order Saddle Points*

Finding multiple first-order saddle points is a very important task in theoretical chemistry. Multiple first-order saddle points are, for example, required to identify the minimum energy path between a reactant and product and characterize complex multi-modal landscapes. Moreover, multiple first-order saddle points are also useful for studying the thermo-dynamical properties of a given cluster or molecular system. Thus the percentage of uncovered first-order saddle points is considered as a crucial performance measure for studying the efficacy of any optimization algorithms.

Here, we consider SMLS configured with different saddle point local searchers (Schlegel or SMLS-Sh, Simons or SMLS-Si, Baker or SMLS-Ba, Dimer or SMLS-Di) as well as evolutionary algorithms designed to identify first-order saddle points such as CGA and BGA discussed in Chapter 2. These methods are investigated here in the experimental study for the discovery of multiple first-order saddle points where a maximum computational budget of 10^6 gradient calls is considered. During the search, all discovered first-order saddle points are archived. The experimental results are summarized in Figure 3-5 where the percentage of uncovered first-order saddle points against the gradient precision accuracy are presented for the benchmark problems.

Figure 3-5: Percentages of uncovered first-order saddle points using SMLS with varying gradient precision accuracy, on all benchmark test problems.

With a convergence criterion at a gradient precision accuracy of 10^{-1} , SMLS is observed to uncover approximately 16 to 27% of first-order saddle points across the four benchmark problems of varying landscape properties. For the same accuracy criterion, evolutionary algorithms (i.e. BGA and CGA) are observed to locate a relatively smaller percentage of first-order saddle points, varying in the range of [2%, 10%], due to the selection pressure in evolutionary algorithms.

In addition, SMLS is also observed to attain the percentage of first-order saddle points at high gradient precision accuracy of up to 10^{-10} . That is due to the well-established high convergence precision ability of local saddle point solvers used in SMLS. BGA and CGA, on the other hand, were observed to arrive at gradient precision accuracies that are under 10^{-5} for the first-order saddle points uncovered. This is expected since the GA is well-established to converge quickly at a solution region of the landscape but lacks the ability to converge to a precise solution efficiently.

B. Convergence Speed

Convergence speed is another important performance measure that shows how fast a given method can converge to a solution. The number of gradient calls is considered as the convergence speed indicator. The fewer the number of the gradient calls, the more efficient is the algorithm.

Here, the number of gradient call is reported as the computational budget incurred by the method upon uncovering a single first-order saddle point with the gradient precision accuracy threshold of 10^{-1} . Experimental results tabulated in Table 3-2 show that stochastic multi-start local search configured with the Schlegel saddle point local solver incurs the lowest number of gradient calls and thus achieve the highest level of convergence speed. Among all-mode following methods, quadratic-approximation-based methods tends to be faster than the rational-approximation-based one. This highlights the efficacy of quadratic-approximation in accurately modeling the landscape and speeding up convergence. On the contrary, minimum-mode following methods, represented by Dimer method, tend to incur more gradient calls than all-mode following methods. The

extra gradient calculations are needed to identify the minimum eigen-mode in the sake of attaining the first-order saddle points.

Comparing stochastic multi-start local search methods with evolutionary saddle points finding methods (i.e. CGA and BGA), we noted that the latter incurred 3- to 4-fold more gradient calls than the former. This highlights another pitfall of the evolutionary saddle points finding methods.

Table 3-2: Number of gradient calls incurred by the algorithms considered on different landscapes.

				Sines Function	Multi-Function	Rastrigin Function	Schwefel Function
Stochastic Multi- Start Local Search	All-mode following methods	Quadratic Model	Schlegel	1,187	1,313	1,207	1,086
			Simons	1,373	1,501	1,445	1,345
		Rational Model	Baker	1,431	1,761	1,575	1,505
	Minimum-mode following methods		Dimer	2,183	2,813	2,207	2,086
	Canonical EA			BGA	3,758	2,750	3,294
CGA				4,246	3,147	3,972	5, 241

3.3 Conclusion

In this chapter, the stochastic multi-start search method has been investigated as a baseline approach for finding multiple first-order saddle points. It has been compared against evolutionary saddle points finding methods, namely, Bungay genetic algorithm and Chuadhury genetic algorithm. Due to the selection pressure and lack of gradient precision accuracy in genetic algorithms, SMLS was observed to outperform these genetic algorithms in terms of the number of first-order saddle point discovered and the gradient precision accuracy as well as the convergence speed.

However, the percentage of uncovered saddle points is still significantly very low, below 30%, across the state-of-the-art methodologies. To rectify this issue, the next

chapter, Chapter 4 introduces the multi-modal memetic framework which considers valley-adaptive clearing scheme proposed for saddle points, coupled with a saddle point local searcher to improve the quality of the saddle points and speed up search convergence.

Chapter 4: A Multi-Modal Memetic Framework for Locating Multiple Saddle Points

As we have discussed in chapter 3, the existing saddle points optimization methodologies have several issues, including the slow convergence and the difficulty in maintaining multiple saddle points, and the low-quality of uncovered saddle points. To elaborate, single ended methods, especially the Schlegel method, have the drawback of attaining few saddle points. However, the attained saddle points are of high quality. On the other hand, evolutionary algorithms are capable of attaining a larger percentage of saddle points compared to conventional methods. However, the uncovered solutions are still of low quality and the percentage of uncovered saddle points requires a significant increase. Here, we are interested in identifying the set of all acceptable high-quality solutions (or saddle points in this context). Such paradigms are commonly known as multi-modal optimization problems (MMOP) [58]. In this work, our research interest is to deal with MMOPs containing a finite set of acceptable solutions.

Such problems are of great abundance in science and engineering [30, 59]. Aerodynamic design, chemical isomers, scheduling and assignment problems represent some of the areas that possess fitness landscapes that are multi-modal in nature. Among these problems, most effort to date has concentrated on revealing the location and properties of the global optimum(s). Significantly, lesser effort has been placed on identifying the set of all acceptable solutions. A motivating example for us in chemistry and physics is the discovery of low energy stable and meta-stable molecular structures which has remained an important and unsolved problem. Global optimization is often

conducted at the expense of locating other low-lying isomers or otherwise known as local minima in the context of optimization. It is worth noting that isomers not only provide key insights into resultant properties but a statistical comparison between isomers also represents a more robust methodology for comparing models and determining fit to the quantum mechanical calculations [60].

In light of their excellent adaptabilities, evolutionary algorithms are well-established as strong candidates for handling multimodal optimization problems. One of their key strengths lies in their ability to maintain a diverse set of solutions in a single population [22-26] when appropriately designed. Hence, several new challenges to evolutionary computation have arisen when dealing with problems plagued with multiple optima, among which, the number of solutions attained and the quality of these solutions as well as the convergence speed.

To tackle the problem of attaining multiple solutions, several efforts by researchers have been made. The most common methods such as fitness sharing and clearing methods are summarized and compared in [26]. The comparison has shown that clearing methods efficiently reduce genetic drift and maintain multiple solutions. However, clearing methods generally assume the landscape to contain equal-sized basins, with a basin size larger than or equal to the clearing radius. Such information requires a prior knowledge about the landscape. An inappropriate value assumed would seriously affect eventually the search performance. Further, some individuals in the clearing methods have no chance to participate in the reproduction operators, limiting the exploration capabilities of the clearing methods.

To tackle the issue of quality in evolutionary algorithm, we consider a recently-born evolutionary paradigm called memetic computation. Memetic computation represents an emerging subject that has engrossed growing research interest in the recent few decades, with a mounting number of accomplishment reported [28]. Memetic computation [29] in its earliest form was introduced as a memetic algorithm (MA), which is a coupling between global, population-based search and individual learning, where the second is often referred to as a meme, capable of individual fine-tuning in converging to the accurate local optima swiftly. MAs are often regarded as population based meta-heuristic search approaches, inspired by Darwin's theory of natural evolution and Dawkin's notion of meme defined as unit of information that reproduce itself while people exchange ideas [33, 61-68]. MA has manifested as a type of hybrid global-local approach that improves both exploration and exploitation in the search. Many MAs have, up to date, been developed for solving real-world problems more efficiently. Such hybrids have been fruitfully employed to solve various medical, scientific and engineering design problems [30-31], where higher quality solutions are reached more efficiently than conventional evolutionary algorithms [32, 33].

In this chapter, we propose a multi-modal memetic framework for locating first-order saddle points, which takes the explorative merits of evolutionary search presented in Section 4.1 and the multi-modal feature of the valley-adaptive clearing scheme, which is briefly introduced in Section 4.2, and the exploitation merit of memetic operator presented in Section 4.3. To tide up the exploration operators with exploitation operators, in specific valley-adaptive clearing scheme with local search strategy considered in the memetic framework, we discuss different strategies of applying local search on valley groups in

order to ensure the effective use of computational resources in Section 4.3. The proposed memetic framework is then detailed in Section 4.4. To verify the efficacy of the proposed MA, experimental studies against different saddle points optimization approaches and recent advances in niching and memetic algorithms are considered in Section 4.5. Section 4.6 finally concludes and summarizes the chapter and put forward future research investigations.

4.1 Evolutionary Search Operators

Evolutionary theory has been developed based on two basic concepts random changes and natural selection. In evolutionary algorithms, three operators implement these concepts: Selection operators implement the natural selection principle which states that, in a given population, the fittest individuals have better chance to survive than others, and crossover and mutation provides the random changes. The operators used in this study are detailed below.

4.1.1 Selection: Stochastic Universal Sampling

Providing zero bias and minimum spread, stochastic universal sampling has been recommended as a selection tool for multi-modal optimization especially clearing methods [15]. Population individuals are mapped contiguously on a line such that each individual occupying a space proportional to its fitness. A number of pointers equivalent to the number of individuals to be selected are placed over the line such that they are equally distant. An example for the stochastic universal sampling is indicated in Figure 4-1.

Figure 4-1: Stochastic universal sampling.

4.1.2 Crossover

Crossover is an important operator in biological evolution as well as in evolutionary algorithms. Crossover is used in genetic algorithms to vary the programming of a chromosome from one generation to another. It simply works by interchanging the information and genes between chromosomes. Different crossover operators have been proposed in literature. Among these operators are single-point, double-point and scattered crossover operators. Here, individuals selected to undergo crossover are modified by the scattered crossover. Unlike single or double point crossover in which close genomes tend to survive together and far apart genomes tend to separate [69], scatter crossover give each gene an equal chance for coming from either parent. Scattered crossover begins by generating a random binary vector. Genes are, then, selected from the first parent where the vector is a 0 and the second where the vector is 1 [44]. The first child is formed from the combined genes, and the vice-versa to form the other child. An example for scattered crossover is present in Figure 4-2. Such a crossover operator, as reported in many studies [70-72], converges faster and produces better solutions than others.

However, scatter crossover is proposed to tackle the combinatorial optimization problem. To customize scatter crossover operator to be capable of handling continuous optimization problems, we propose an arithmetic scatter crossover operator. In this

operator, genomes, to be exchanges in the canonical scatter crossover, are modified as follow:

$$x_{i_{new}}^1 = r_1 x_i^1 + (-1)^l (1 - r_1) x_i^2 \quad (\text{Eq. 4.1})$$

$$x_{i_{new}}^2 = r_2 x_i^2 + (-1)^l (1 - r_2) x_i^1 \quad (\text{Eq. 4.2})$$

where r_1 and r_2 are two random numbers, x_i^1 and x_i^2 are the i^{th} gene of the first and second parent at which the mask vector is one, respectively, l is a random integer, and $x_{i_{new}}^1$ and $x_{i_{new}}^2$ are the i^{th} gene of the offspring. An example for arithmetic scattered crossover is present in Figure 4-3.

5	6	7	8	9	10	11	Parent 1
1	2	3	4	5	6		Parent 2
1	0	0	0	1	0	1	Mask
1	6	7	8	5	10	7	Child 1
5	2	3	4	5	6	11	Child 2

Figure 4-2: Scatter crossover.

Figure 4-3: Arithmetic scatter crossover.

4.1.3 Mutation

Another important operator in evolutionary algorithm is the mutation operator. Mutation or variation operator, unlike crossover operator which requires, at least, two parents, changes the information contained in the genome or (genes) of single parent according to a given probability distribution. In the last decades, dozens of mutation operators have been proposed, these operators can be classified according the chromosome representation in EA into binary and real coded operators. Binary coded mutation operators refer to the group of mutation operators which is capable of dealing with binary coded genetic algorithms. Among these operators are uniform, greedy, translocation, transposition and frame-shift mutation operators. Whereas real-coded mutation operators are the mutation operators proposed to handle real-coded genetic algorithms. The most widely used real-coded mutation operators involve Gaussian, Cauchy, and non-uniform mutation operators. Unlike Gaussian mutation operator slightly changes all the genes of the individual using a small, normally distributed disturbance, non-uniform mutation operator randomly selects a gene from the parent individual and replaces it with a random number within the range of the solution space. At early stages,

non-uniform mutation operator achieves the long jumps of the Cauchy mutation operator. At later stages, non-uniform mutation operator attains the fine-tuning of the Gaussian mutation operator.

Non-uniform operator is proposed by Michalewicz [73] for the purpose of overcoming the disadvantages of random mutation in the real-coded GA. In this operator, each individual x^g selected to undergo mutation in g^{th} generation are modified as follow,

$$x_i^{g+1} = \begin{cases} x_i^g + \Delta(g, u - x_i^g) & \text{if } \varepsilon > 0.5 \\ x_i^g - \Delta(g, x_i^g - w) & \text{if } \varepsilon \leq 0.5 \end{cases} \quad (\text{Eq. 4.3})$$

where u and w are the upper and lower limits of the variable x_i^g , respectively. The function $\Delta(g, q)$ returns a number between 0 and y such that $\Delta(g, q)$ move toward at zero as g raises. This property forces the mutation operator to explore the search space uniformly at early phases, and very locally at final phases of the search. Working in such a way raises the chance of producing a new number near to its predecessor. We use the following function:

$$\Delta(g, y) = y \left(1 - r^{\left(1 - \frac{g}{g_{max}}\right)^b} \right) \quad (\text{Eq. 4.4})$$

where r and g_{max} denote a randomly-chosen, uniformly-distributed number from $[0, 1]$ and the maximum number of generations, respectively, while b denotes a system parameter, specifying the degree of dependency on the iteration number.

4.2 Niching Operator: The Valley-adaptive Clearing Scheme

Seeking for more than one solution is a common theme in numerous problems of science and engineering. Using evolutionary algorithms to address such problems is not generally recommended, since the conventional evolutionary algorithm is designed to converge at single solution. In recent decades, many niching methods have been proposed to make genetic algorithms capable of tackling multi-modal optimization problems. The more popular approaches are fitness sharing [22], deterministic crowding [74], probabilistic crowding [75], clustering [76], restricted tournament selection [77], and the clearing methods [26, 78]. A comparison of these multi-modal optimization methods in [26] has reported that clearing methods are shown to be generally more efficient for exploration of the problem search spaces. Unlike other niching methods where resources are shared among similar individuals of the population, clearing methods only consider the elite individuals [26], while other less fit individuals are treated with a killer penalty. Petrowski [78] also reported that clearing methods effectively succeed in reducing the effect of genetic drift. Nevertheless, it is worth noting that clearing methods suffer from several limitations [26]. Particularly, inferior individuals have no chance of participating in the mating operations, limiting the capability of the algorithm in exploring the search space sufficiently well. There is also a poor usage of resources since those individuals that generate little effectiveness hog up the limited finite population slots. It is also worth noting as well that the efficacy of both canonical clearing [78] and modified clearing methods [26] remains to be sensitive to the configuration of Q_{clear} or clearing niche radius. An inappropriate value assumed would seriously affect the eventual search performance. Lastly, since clearing methods do not adapt to the nature of the problem landscape, there

is a high chance that some local optimum solutions are left uncovered or missed in the search.

To mitigate these challenges, we propose a valley-adaptive scheme. In this scheme, the population of individuals are segregated together shall share a common valley using, using a *hill-valley* detection *scheme* (see Section 4.2.1)². Valley-adaptive clearing scheme which involves the *identification* (see Section A.), *clearing* (see Section B.) and *valley replacement* phases (see Section C.) is discussed in Section 4.2.2. The complexity of the proposed valley-adaptive clearing scheme is then presented in Section 4.2.3. The proposed valley-adaptive clearing scheme is then compared to its clearing compeer in Section 4.2.4.

4.2.1 Hill-valley Detection

The hill-valley detection procedure [79] begins by generating a line connecting two points (**s** and **e**) in the Euclidean space. Subsequently, a number of intermediate points are sampled within the line connecting these points. The fitness values of these points are then calculated. A valley existence, on one hand, is identified, if the fitness of any sampled points represents an improvement over that of the given points. Otherwise, a hill is found (see Figure 4-4 for an illustration).

$$\mathbf{p} = \mathbf{s} + h (\mathbf{e} - \mathbf{s}) \quad (\text{Eq. 4.5})$$

where h is a scaling factor.

² A minimization problem is assumed

Figure 4-4: Hill-valley detection.

The procedure begins with the sampling of a point that lies at h distance from the starting point. If the fitness of both end points fares better than that of the sampled point, a hill is detected and the scheme terminates. Otherwise, the process is repeated with another point at $(2 * h)$ of the distance between s and e according to (Eq. 4.5). If no hill has been detected, new points will be sampled, until a maximum number of points (m) is sampled.

4.2.2 Valley-adaptive Clearing Scheme

Valley-adaptive clearing scheme, summarized in Table 4-1, is proposed to adapt to non-uniform width of the valleys in the problem landscape. The main idea of the valley-adaptive clearing scheme is to group the population individuals into niches based on whether they share a common valley or not, and then clear the lowest fit individuals from the niches, by relocating them on unexplored area of the landscape.

The valley-adaptive clearing scheme is composed of three core phases. The valley identification phase categorizes the population of individuals into groups of individuals sharing the same valley, denoted as V_{groups} . Subsequently, the dominant individual (i.e., in terms of fitness value) of a valley group or V_g is archived if it represents a unique local optimum solution, while all other members of the same group undergo the valley replacement phase where relocation of these individuals to new basin of attractions or valleys are made so that unique local optimum solution elsewhere may be uncovered. In

the event that no local optimum solution exists in a valley group, all individuals of the group will undergo the valley clearing stage where elite individuals are ensured to survive across the search generation while all others are relocated to new basin of the attractions.

A. Valley Identification Phase

Valley identification phase (see Table 4-2) details the process of clustering individuals into valley groups. It simply works by ranking the population individuals according to their fitness value. The ranked population individuals are then traversed from the best individual ($\mathbf{x}$) to the worst ($\mathbf{y}$). Individuals within Q_{clear} of $\mathbf{x}$ are segregated forming niche, $\mathbf{V}_g$. The whole population is traversed once forming several niches, $\mathbf{V}_{\text{groups}}$. Therefore, $\mathbf{V}_g \in \mathbf{V}_{\text{groups}}$ refers to a set of ranked individuals, sharing a common niche. The k best individuals of $\mathbf{V}_g$ are taken to be the elites and will be considered for further evolutionary process, while the rest are considered inferior and undergo valley detection procedure if $\mathbf{x}$ and $\mathbf{i}$ share the same valley. Otherwise, a new valley group is created around $\mathbf{i}$, say $\mathbf{V}_{g'}$, and the rest of valley inferiors join $\mathbf{V}_{g'}$ if they are closer to $\mathbf{i}$ than $\mathbf{x}$. The same process repeats for the next inferior until all inferiors $\in \mathbf{V}_g$ have been considered. The same process repeats for all $\mathbf{V}_g \in \mathbf{V}_{\text{groups}}$.

Table 4-1: Valley-adaptive clearing scheme.

<p>Begin</p> <p>1 $\mathbf{V}_{\text{groups}} = \text{Identifying Valleys}(\mathbf{P}, k, n)$</p> <p>2 $\mathbf{P} = \{ \}$</p> <p>3 For each $\mathbf{V}_g \in \mathbf{V}_{\text{groups}}$</p> <p>3.1 $\mathbf{x} = \mathbf{V}_g[1]$;</p> <p>3.2 If ($\mathbf{x} \in \mathbf{X}_T$)</p> <p>3.2.1 $\mathbf{P} = \mathbf{P} \cup \text{Valley Replacement}(\mathbf{V}_g)$</p> <p>3.3 Else</p> <p>3.3.1 $\mathbf{P} = \mathbf{P} \cup \text{Valley Clearing}(\mathbf{V}_g)$</p> <p>3.4 End</p> <p>3.5 End for</p> <p>End</p>

Table 4-2: Pseudo code of valley identification phase.

```

Vgroups = Identifying Valleys (P, k, n)
Begin
1. Sort population individuals (P) according to their fitness in ascending order.
2. q = zeros (n,1); V0 = { }; id = 0;
3. Vgroups = {}
4. Elite = {}; Inferior = {}
5. For i=1 to n
5.1. If P[i] is assigned to valley group: continue
5.2. if ( $F(\mathbf{P}[i]) > f_{max}$ )
5.2.1. inferior = inferior  $\cup$  {P [i]}
5.2.2. continue
5.3. end if
5.4. Vid = {P[i]}.
5.5. w = 1
5.6. For j = i+1 to n
5.6.1. If (  $\|\mathbf{P}[j] - \mathbf{P}[i]\| < Q_{clear}$  )
5.6.1.1. Vid = Vid  $\cup$  {P[j]}.
5.6.2. End if
5.7. End for
5.8. Vgroups = Vgroups  $\cup$  [80]
5.9. id = id + 1
6. End for
7. For each Vg in Vgroups do :
7.1. s = size (Vg)
7.2. Elite = Elite  $\cup$  { Vg [1:k]}
7.3. For i = k+1:s
7.3.1. If hill-valley-detected(Vg[i], Vg[j])
7.3.1.1. Vg' = {Vg[j] | j  $\geq i$  and  $\|\mathbf{V}_g[j] - \mathbf{V}_g[i]\| \leq \|\mathbf{V}_g[j] - \mathbf{V}_g[1]\|$ }
7.3.1.2. Vgroups = Vgroups  $\cup$  Vg'
7.3.2. Else
7.3.2.1. Inferior = Inferior  $\cup$  { Vg[i]}
7.3.3. End if
7.4. End for
8. End For
End

```

B. Valley Clearing Phase

Valley clearing (see Table 4-3) is a process where less fit individuals (or inferiors) are relocated out of the same basin of attraction, leaving valleys to be further exploited by the fittest individuals (elites). In the valley clearing process, each inferior member of the valley group (**V_g**) is relocated randomly in the range of [Q_{clear} and $3 * Q_{clear}$] whereas other individuals (or elites) are left unchanged for the purpose of exploiting within the basin of attraction.

Table 4-3: Pseudo code of valley clearing phase.

V_g =Valley Clearing (V_g)	
Begin	
1.	For each individual (x) $\in$ V_g
1.1.	if $F(\mathbf{x}) \geq \xi$
1.1.1.	$r = \text{random}(Q_{\text{clear}}, 3 * Q_{\text{clear}})$
1.1.2.	$\mathbf{n} = \text{random unit vector}$
1.1.3.	$\mathbf{x} = \mathbf{x} + r \mathbf{n};$
1.2.	End if
2.	End for
End	

C. Valley Replacement Phase

The motivation behind valley replacement process (see Table 4-4) is to reduce any potential computational resources that may be wasted on rediscovering of valleys where the local optima solutions have already been uncovered. Individuals of the populations falling in previously encountered valleys are replaced with individuals in new basin of attractions, so as to bias the search towards previously unexplored region of the landscape. Previously encountered valleys refer to the valleys where the stationary points have been uncovered and archived. Archive (see Section 4.2.3) can thus be used as a reference for these valleys. To determine whether an individual, $\mathbf{x}$, lies on uncovered valley or not, the archive is searched for a stationary point lies within Q_{clear} of $\mathbf{x}$. This search in a tree-based archive costs a maximum of $\log(a_n)$ where a_n is the number of archived solutions. If no stationary point has satisfied this condition, $\mathbf{x}$ is considered for further investigation. The cost of this process will end up with $O(\log(a_n))$. Otherwise, the valley detection procedure is considered. The cost of valley detection in the worst case is m where m is the number of sample points. The maximum cost would be equal to $O(\log(a_n)+m)$.

Table 4-4: Pseudo code of valley replacement phase.

V_g =Valley Replacement (V_g)	
Begin	
1.	Generate x random vector within the boundary conditions
2.	Find closest item in the archive, say w
3.	If $\ \mathbf{w}-\mathbf{x}\ > Q_{\text{clear}}$:
3.1	For each individual (y) $\in V_g$
3.1.1	$\mathbf{r} = \text{rand}(0, Q_{\text{clear}}/2)$
3.1.2	v = random vector
3.1.3	$\mathbf{v} = \mathbf{v}/\ \mathbf{v}\ $
3.1.4	$\mathbf{y} = \mathbf{x} + \mathbf{r} * \mathbf{v},$
3.2	End for
4.	Else:
5.	If no hill-valley-detected(V_g [<i>i</i>], V_g [<i>j</i>]) : GOTO 3.1
6.	End if
End	

4.2.3 Complexity of Valley-adaptive Clearing Scheme

To assign individuals to a given valley V_1 , its dominant, the first individual of the sorted population, is compared to all valley-unassigned individuals. It is compared to the $n - 1$ other individuals to attain valley V_1 . Assume that this valley contains v_1 individuals. To detect if there is a sub valley, v_1 individuals undergo valley-detection procedure with cost $v_1 * m$ in total, where m is the number of sample points.

Next, the dominant in valley V_2 is the first individual of the population not yet assigned. It is compared to the $n - 1 - v_1$ individuals that are not yet assigned. By induction reasoning, the dominant of valley V_i has to be compared to the $n - 1 - \sum_1^i miv_1$ individuals not yet assigned. Let v be the number of valleys, the total number of comparisons ρ is:

$$\rho = (\sum_i^v miv_i) - v \quad (\text{Eq. 4.6})$$

The maximal value of term $\sum_i^v miv_i$ is obtained when the valley $\mathbf{V}_1$ contains $(n - v + 1)$ individuals, while each other valley contains only one individual. Its minimal value is obtained when valley $\mathbf{V}_v$ contains $(n - v + 1)$ individuals and each other valley contains only one individual. Consequently, ρ belongs to the interval:

$$m \left(n + \frac{(v(v-3))}{2} \right) \leq \rho \leq m \left(vn + \frac{(v(v_1-1))}{2} \right) \quad (\text{Eq. 4.7})$$

Therefore, the complexity in the worst case is $O(vmn)$.

4.2.4 Experimental Study on Clearing Schemes

In this section, we study the efficacy of the valley-adaptive clearing scheme (AVAC) in comparison with two other existing clearing methods, particularly the canonical clearing (C) and modified clearing (MC) schemes, using the same multi-modal benchmark test problems described and considered in Chapter 3 for the sake of consistency. The experimental settings and results are reported in what follows.

The following algorithmic configurations are also considered: a mutation probability of 40% to increase the level of diversity and a crossover probability of 60%. The number of sample points (m) in the hill-valley detection, the number of elites (k), and the archiving-distance threshold are configured at 5, 2 and 0.5, respectively. A population size of 100 is considered. All the search terminates at a maximum generations of 100. The average is taken from 50 execution runs.

In this section, performance analysis of clearing schemes with and without archive is presented in Subsection A . The sensitivity of the archiving clearing schemes to the niche radius change is subsequently investigated in Subsection B.

A. Performance Analysis of Clearing Schemes

We study the search performance of the clearing schemes with and without archiving in terms of percentages of uncovered optima. Experiments were, therefore, executed on different landscapes of varying degree of complexity and dimensionality. The performance of the clearing schemes (C-Canonical Clearing, and MC -Modified Clearing) and archiving clearing schemes (AC- Archiving Canonical Clearing, AMC – Archiving Modified Clearing, and AVAC- Valley-Adaptive Clearing) on the benchmark problems is summarized in Figure 4-5.

Figure 4-5: Percentages of uncovered optima by clearing schemes, on the three 2D benchmark problems.

Figure 4-5 shows the percentages of optima uncovered by different clearing schemes on the four benchmark problems. The plotted results show that the archiving clearing

schemes have rates of improvement ranging from 2-10 times over non-archiving schemes in the percentage of uncovered optima. It is also shown that the valley-adaptive clearing scheme maintains the highest percentages of uncovered local optima on each of the different benchmark problems. The percentages of optima adaptive uncovered by AVAC, for example, vary in the ranges of 90-100%, whereas those uncovered by others clearing compeers vary largely in the range of 10-95%, respectively. Due to the large search space, the gaps between the percentages of the proposed scheme and other clearing schemes are observed to be very wide on Sines and Schwefel benchmark problems.

Since the archiving clearing schemes outperformed non-archiving clearing schemes in maintaining multiple optima, we will consider only the archiving clearing schemes for further experimental analysis.

B. Sensitivity Analysis of Niche Radius

Next, we study the sensitivity of the niche radius or Q_{clear} on the search performance of the archiving clearing schemes (AVAC, ACM, and AC) as well as the fitness sharing (FS). In Figure 4-6, the percentages of local optima uncovered by the respective archiving clearing schemes as well as fitness sharing (FS), for different configurations of niche radius Q_{clear} , on the four benchmark problems are summarized to vary largely in the range of 15-100%, on the four benchmarks. Among niching operators, fitness sharing was observed to converge the lowest number of optima, varying between 15-45%, whereas the archiving clearing schemes was observed to converge to higher percentage of optima, varying between 40-100%. Among all archiving clearing scheme, the valley-adaptive clearing scheme fares significantly better in the range of 85-100%. The percentage

standard deviations of local optima uncovered by AVAC, FS, AC AMC vary in the ranges of 1.5-5.6, 5.3-12.7, 5.5-20 and 7-23.5, respectively, demonstrated high sensitivity of the latter three methods to Q_{clear} , the niche radius.

Figure 4-6: Percentages of optima uncovered by archiving clearing schemes and fitness sharing on the benchmark problems, for different niche radii (Q_{clear}).

4.3 Memetic Operator: Life-time Learning in Memetic Algorithms and Valley Elitism

Motivated by Dawkin's notion of a meme, Moscato and Norman have presented the term memetic algorithm [29]. In Moscato's works, memetic algorithms are described as genetic algorithms with a local search which plays a central role in the search process. However, the key difference between genetic and memetic lies in the difference between the gene and meme. In particular, the meme, before being passed on, typically evolve as the person, transmitting it, thinks and understands whereas genes exchanged intact. The main objective of memetic algorithm is to exploit a population-based global search

methodology to broadly locate good areas of the search space, combined with repeated usage of local search heuristic by individual solutions to locate local optimum. Ideally, memetic algorithms encompass the duality of genetic and cultural evolution, permitting the diffusion, selection, inheritance and variation of memes as well as genes.

Extending Moscato and Norman work, Radcliffe and Surrey [81], afterward, provided a formal description of memetic algorithms. The description provides a homogeneous formal framework, unifying memetic and genetic algorithms. Based on this framework, if a local optimizer is considered with genetic algorithms and every child in the population undergo the local search operations, the memetic algorithm can then be simply considered as a special type of genetic exploration over the local region of the optima. Reproduction operator will typically generate solutions that are outside this region, but a local optimizer can mend such solutions to make final children who fall on this region. Such a procedure yields a memetic algorithm.

One of the important design issues in memetic algorithms is to decide upon which individuals would undergo the life-time learning procedure. Blindly applying life-time learning procedure is neither conceivable nor rational and may lead to unnecessary consumption of resources. Valley-adaptive clearing scheme may provide a potential solution to this challenge. As can be noted from the previous chapter, the individuals after valley-adaptive clearing are segregated into valley groups shall they share a common valley. Population individuals can be classified into valley groups as shown in the scratched figure below. As landscape transformed according to (Eq. 4.8), first-order saddle points are now lying at the bottom of the valley. Valley elites are the candidate

solutions, possessing the best quality among its valley group to converge to first-order saddle points. Their chance to converge to a solution is higher than any other individual per valley groups. Since they possess such a quality, we consider them for life-time learning procedure to ensure faster convergence and increase the chances of the local search or life-time learning procedure to succeed. Such a strategy is called valley elitism local refinement, shown in Figure 4-7 while Figure 4-8 shows the whole-valley local refinement strategy where all individuals are allowed to undergo local refinement procedure.

Figure 4-7: Valley-elitism-local refinement strategy.

Figure 4-8: Whole-valley-local refinement strategy.

In order to confirm this hypothesis, we compare the success rate—the percentage of successfully converged solutions—of the local search or life-learning procedure when applied on valley elites and when applied on the whole valley group members and number of gradient calls needed on different landscapes (Sines, Multi-function, Rastrigin and Schwefel)- discussed in Chapter 3. According to Table 4-5, the success rate of the valley elitism strategy is more than 50% while the success rate of the whole valley strategy is below 30%. Moreover the average gradient calls incurred by different strategy (as shown in Table 4-6) highlighted the efficiency of valley elitism local refinement strategy over the whole valley local refinement strategy.

Table 4-5: Success rate of different local search strategies.

	Valley-Elitism-Local Refinement Strategy	Whole-Valley-Local Refinement Strategy
Sine Function	%75	%15
Multi-Function	%80	%25
Rastrigin Function	%70	%20
Schwefel Function	%55	%10

Table 4-6: Average number of gradient calls incurred by different local search strategies.

	Valley-Elitism-Local Refinement Strategy	Whole-Valley-Local Refinement Strategy
Sine Function	75	1560
Multi-Function	90	1798
Rastrigin Function	60	1235
Schwefel Function	250	1723

4.4 *The Multi-modal Memetic Framework*

In this section, a real-coded multi-modal memetic framework for locating first-order saddle points (see Figure 4-9) is introduced. In this framework, a population of individuals is first randomly generated. Once population individuals are initialized, they undergo the evolutionary process. To comply with the principle of the natural selection in the evolutionary theory, the population individuals are evaluated to define their fitness on solving the problem, using (Eq. 4.8).

$$f_{ts}(\mathbf{x}) = \frac{f(\mathbf{x}) - f_{max}}{(|\mathbf{g}| + \varepsilon)(|n-1| + \varepsilon)} \quad (\text{Eq. 4.8})$$

where f_{max} denotes the energy threshold below which the saddle point of interest lies. $f(\mathbf{x})$ donates the energy at current configuration $\mathbf{x}$, $|\mathbf{g}|$ is the gradient norm at $\mathbf{x}$, n is the number of negative eigenvalues, ε is chosen small to prevent division by zero error. First-order saddle points lie at the optima of the fitness function $f_{ts}(\mathbf{x})$ where the gradient norm vanishes and $n = 1$.

Putting the focus on (Eq. 4.8), $f_{ts}(\mathbf{x})$ is noted here to consist of two portions: numerator and denominator. The numerator part or $(f(\mathbf{x}) - f_{max})$ serves to favor solutions having quality that meet the user defined minimum criteria of f_{max} . The denominator term or $(|\mathbf{g}| + \varepsilon)(|n-1| + \varepsilon)$, on the other hand, is designed to bias the

search towards first-order saddle points or transition states. Hence, for a minimization problem, the lower energy of a first-order saddle point or the greater the gap between f_{max} and $f(\mathbf{x})$, the fitter is the first-order saddle point.

Individuals are subsequently selected to undergo the evolutionary process of crossover and mutation. Individual then undergoes the valley-adaptive clearing where individuals are segregated into valley groups shall share a common valley. A valley group then undergoes a valley replacement phase if it has a member satisfying first-order saddle point criteria in (Eq. 2.3). In such a case, the individuals are replaced into different basins of attraction in the effort toward discovery of new first-order saddle points. Otherwise, the valley group undergoes a valley clearing phase, where the best members (valley elites) are allowed to survive while others are relocated in order to explore the nearby first-order saddle points in the search space. Adapting valley-elitism-local refinement strategy, valley elites are thus allowed to undergo a life-time procedure. In this study, we consider Schlegel method as a life-time procedure for finding first-order saddle points. Identified saddle points undergo duplication checking and archiving. During duplicate identification and replacement process, all non-duplicated discovered saddle points are archived using a multi-maps, associative data structures that store elements indexed by keys to ensure that comparisons and duplicate checking are performed efficiently. Multi-maps, in particular, allows for quick access and rapid retrieval of archival elements based on key values. From the computational cost prospective, the worst case access time is $O(\tau + 1)$ where τ is the number of keys and worst case insertion time is $O(\log s_n)$ where s_n is the number of archived solutions.

Once the identification of non-duplicate solutions and archival has been achieved, the same process repeats until a maximum number of generations or other stopping criteria have been satisfied.

Figure 4-9: A Multi-modal Memetic Framework for locating first-order saddle points. [Note that the components that will be customized later for the problem at hand are highlighted by <<Customized>>. The evolutionary operators will be customized for the water clusters and ring-deficient covalently bonded small molecules. The memetic operator will be customized for biochemical systems in Chapter 5].

4.5 Experimental Study on Saddle Point Optimization

In this section, we study the efficacy of multi-modal memetic framework (MMA) for locating first-order saddle points and pit it against several existing state-of-the-art first-order saddle point optimization methods, including the stochastic multi-start local

search(SMLS) [82] as representative of single-ended methods, while Chaudhury GA (ChGA) [21] and Bungay GA (BGA) [19] as representatives of the genetic algorithm. The sequential niching memetic algorithm (SNMA) [83], the adaptive-valley clearing GA (AVAC) [84] and Mahalanobis with self-adaptive-niche radius CMA-ES (M-S-CMA-ES)[85], as representatives of recent advances in niching and memetic algorithms, are also considered for comparison in the present study. Brief descriptions of the algorithms considered and experimental settings are given in Table 4-7 and Table 4-8, respectively.

This section is outlined in the following manner. Section 4.4.1 represents results on the percentages and precision accuracy of uncovered first-order saddle points while Section 4.4.2 details the convergence speed of different algorithms on different landscapes.

Table 4-7: Descriptions of the state-of-the-art algorithms considered.

Category	Algorithm	Description
Single-ended Methods	Stochastic Multi-start Local Search (SMLS)	A typical stochastic multi-start solver with the Schlegel algorithm considered for locating first-order saddle points.
Genetic Algorithm	Bungay Genetic Algorithm (BGA [19])	A binary-coded GA proposed by Bungay to find the transition states of some chemical reactions such as HNC isomerization.
	Chaudhury Genetic Algorithm (ChGA [21])	A real-coded genetic algorithm proposed to find the transition states on Lennard Jones clusters.
Recent Advances in Niching and/or Memetic Algorithms	Sequential Niching Memetic Algorithm (SNMA) [82]	An advanced niching technique that extends the effort of Beasley et al. [83] on sequential niching technique to locate multiple optimal using a local search to improve solution precision. SNMA incorporates a gradient-based local search process, a derating function and the basic clearing technique.
	M-S-CMA-ES, Mahalanobis Self-adaptive niche radii CMA-ES	An advanced niching method which extend CMA-ES with adaptive niche radius strategy coupled with Mahalanobis distance for solving complex, non-isotropic high-dimensionality multi-modal optimization problems.
	Adaptive-valley clearing GA (AVAC)	An advanced multi-modal optimization genetic algorithm that consider the same evolutionary operators as well as valley-adaptive clearing [84].

Table 4-8: Experimental settings.

Algorithm	Description
$P_{mutation}$	50%
$P_{crossover}$	50%
N	10
g_{max}	100,000
m	5
k	2
Q_{clear}	0.1

4.5.1 Precision Accuracy of Uncovered First-order Saddle Points

As explained in chapter 3, the percentage of uncovered first-order saddle points is considered as a crucial performance measure for studying the efficacy of any optimization algorithms. Here, several algorithms are investigated in the experimental study on discovery of multiple first-order saddle points where a maximum computational budget of 10^6 gradient calls is considered. During the search, all discovered first-order saddle points are archived. The experimental results are summarized in Figure 4-10. As observed from the figure, the traditional stochastic multi-start local search locates higher percentages of first-order saddle points than traditional genetic algorithms of BGA and CGA- The percentages roughly maintained for accuracy up to -10. Multi-modal evolutionary algorithms attain several folds more than the stochastic multi-local search. However, one with increasing accuracies, the number of saddle points significantly decreases to approach zero which sometimes exponentially, as the accuracies go below -5. On contrary, the multi-modal memetic algorithms attain the highest percentage of the uncovered first-order saddle points which is maintained across higher accuracies as witnessed from the standard deviation of the first-order saddle points uncovered under different accuracies and landscapes on Table 4-9. MMA was observed to maintain the largest percentage of uncovered saddle points with more than 90% across different landscapes.

4.5.2 Convergence Speed

Convergence speed is another important performance measure that shows how fast a given method can converge to a solution. The number of gradient calls is considered as the convergence speed indicator. The fewer the number of the gradient calls, the more efficient is the algorithm.

Here, the number of gradient call is reported as the computational budget incurred by the method upon uncovering a single first-order saddle point with the gradient precision accuracy threshold of 10^{-1} . Experimental results tabulated in Table 4-10 shows that canonical GA always outperforms the recent multi-modal niching EA in the convergence speed. Imbalance between search space exploration and exploitation decreases the efficiency of the recent multi-modal Niching EA of converging as fast as canonical EA. Such imbalance has been resolved in the multi-modal memetic evolutionary algorithm such SNMA and MMA. Among multi-modal memetic algorithm, the proposed MMA attains more than 10 folds convergence speed when compared to its multi-modal memetic compeer, the SNMA. The number of folds increases even more when compared to others methodologies. For example, the improvement rate, when compared to AVAC arrives, on the multi-function landscape, to 100 times. Clearly, the multi-modal memetic framework proposed incurred significantly small number of gradient calls as compared to conventional stochastic search method-the SMLS, the canonical GA and the recent advances in multimodal niching and memetic algorithms for the gradient precision criterion considered.

Table 4-9: Standard deviation of uncovered first-order saddle points using different algorithms with varying gradient precision accuracy, on all benchmark test problems.

	SMLS	BGA	CGA	M-S-CMA-ES	AVAC	SNMA	MMA
Sine function	0	3.12	1.43	13.6	44.611	2.59	1.96
Multi-function	1.198	3.25	1.436	22.9	44.611	2.45	0.506
Rastrigin function	1.187	3.25	1.436	9.66	42.6	0.5	1.96
Schwefel function	0.72	3.25	1.43	17.03	44.61	0.51	1.96
Average	0.77625	3.2175	1.433	15.7975	44.108	1.5125	1.5965

Figure 4-10: Percentages of uncovered first-order saddle points using different algorithms with varying gradient precision accuracy, on all benchmark test problems.

Table 4-10: Number of gradient calls incurred by the algorithms considered on different landscapes.

		Sines Function	Multi-Function	Rastrigin Function	Schwefel Function
SMLS	SMLS -Schlegel	1,187	1,313	1,207	1,086
Canonical EA	BGA	3,758	2,750	3,294	4,532
	CGA	4,246	3,147	3,972	5, 241
Recent Multi-Modal Niching EA	M-S-CMA-ES	7,598	5,479	6,654	15,470
	AVAC	4,598	4,010	4,427	6,470
Multi-Modal Memetic EA	SNMA	950	780	1,265	1,725
	MMA	70	90	60	250

4.6 Conclusion

In this chapter, a novel multi-modal memetic framework for finding first-order saddle points has been proposed. The proposed framework composed of evolutionary operators such as crossover, mutation, selection and niching operators, or the valley adaptive clearing scheme, and the memetic operator or Schlegel algorithm. Some important aspects for implementing the algorithm were discussed. These include the following points. Like evolutionary algorithm, memetic algorithms manipulate a collection of potential solutions in parallel. Unlike canonical EAs that merely consider stochastic search operators, memetic algorithms consider, in addition to the stochastic search operators, an individual learning stage that refines the prospective solutions. In the proposed framework, we also proposed valley-adaptive clearing scheme to maintain multiple first-order saddle points. Performance of the clearing schemes with and without archiving was compared. The results showed that the archiving clearing schemes generally maintained much more optima than existing non-archiving clearing schemes. Among the archiving clearing schemes, experimental results showed that the proposed valley-adaptive clearing scheme maintained the highest percentage of uncovered optima with varying radii and landscapes of different degrees of complexity and dimensionality. These results highlighted the adaptability and robustness of the proposed scheme. Valley-

elitism-local refinement strategy is considered to tide up the multimodal operator or the valley adaptive clearing and the memetic operator for applying the life-time learning procedure on individuals with the highest potential to converge to first-order saddle points. Experimental studies on benchmark problems show the efficacy of the proposed method in terms of solution quality and convergence speed as compared to other counterparts.

Thus far, we have proposed a multi-modal memetic framework capable of finding more than 90% of saddle points in several synthetic landscapes with gradient precision lower than 10^{-9} . Next, we explore the biochemical applications to find a set of appropriate applications of finding first-order saddle points. Chapter 5 describes possible applications for the proposed framework. Chapter 5 also exploits the challenges that the current framework faces when dealing with complex biochemical systems. In particular, several molecular rules and constraints shall be maintained across the search process. Due to the importance of water clusters and covalently-bonded molecules, the rest chapters propose a customization of the proposed framework especially the evolutionary operators that incorporate domain knowledge to improve the search process. These chapters constitute the second part of the thesis, discussed next.

PART II: Biomolecular Applications:

Water Cluster and Ring-deficient Covalently-bonded Small Molecules

Biomolecular systems are very important constituents of biological systems, with a significant impact on environment and our daily life. Living organisms have an incredible number of biomolecular systems, which at a small scale can be classified into few biomolecular systems, including water, amino acids, peptides, proteins, DNA and RNA. In this part, different types of biomolecular systems would be presented Chapter 5. A concise introduction about the computational and experimental methods has been presented where pros and cons of each method have been highlighted. Due to the limited ability of the experimental approach when dealing with small molecules, computational approaches have been considered. Among computational approaches, quantum-mechanical models have been considered due to their sufficient accuracy needed to attain the saddle points. Due to their abundance in nature and importance in our daily life, two main biomolecular systems have been considered in this study: water clusters and ring-deficient covalently-bonded small molecules. The problem of finding saddle points or transition states in a biomolecular context has been formulated. In Chapter 5, two extensions have been introduced to the general framework proposed in Chapter 4: a molecular local search operator and structural similarity matrix. The extended framework has been considered on biomolecular systems such as water and covalently-bonded systems, where the framework has shown to attain few solutions. Further investigation into evolutionary operators has highlighted that the need for specially tailored operators, which incorporate the domain knowledge of the biomolecular systems. For this purpose, two extensions have been introduced to further empower the capability of the framework proposed in Chapter 4 and extended in Chapter 5. The extensions are reported in Chapter 6 and 7. In chapter 6, molecular operators especially designed for water clusters have been

introduced to the extended multimodal memetic framework proposed in chapter 4. The water-cluster extended framework has attained better performance in terms of computational cost and number of transition states uncovered compared to its compeers. A number of newly uncovered transition states in water clusters have been reported, which would help researchers and scientists to understand the thermodynamic properties of water clusters. Similarly, in chapter 7, two molecular operators have been introduced to the extended framework in Chapter 5 for tackling covalently-bonded molecular systems. The extended framework has attained the largest set of transition states of a biomolecular system with the minimal computational cost among its compeers. A number of web-services have been proposed to enable scientists of finding transition states of biomolecular systems. To conclude, this study highlights the importance of domain knowledge incorporation in locating saddle points when dealing with biomolecular systems.

Chapter 5: Biomolecular Systems

Biomolecular system is a special type of chemical systems—aggregations of atoms forced together with natural forces, which participate in biological functions. Living organisms have an uncountable number of biomolecular systems, including water, amino acids, peptides, proteins, DNA and RNA. These systems can be classified according to how the aggregation of the system takes place. In one hand, biomolecular systems are called covalently-bonded systems if their atoms are synergized together with strong covalent bonds. On the other hand, the systems are called atomic and molecular clusters if the system is aggregated by a weak hydrogen bond or any other weak forces.

Biomolecular systems have been studied both experimentally and computationally. Experimentally, biomolecular systems have been studied using nuclear magnetic resonance (NMR) spectroscopy, Infrared spectroscopy and X-ray crystallography. Such technologies are, however, expensive to own and only available at few number of laboratories worldwide. Alternatively, biomolecular systems can be studied computationally using quantum mechanics (QM), molecular mechanics (MM) and hybrid quantum mechanics/molecular mechanics (QM/MM). Small biomolecular systems, up to a few hundred of atoms, can be studied using quantum approaches, which are distinguished with high accuracy and the capability to describe reactive reactions in terms of bond breaking and formation. The larger biomolecular systems can be studied, on the other hand, using molecular mechanics by which the thermodynamic and kinetic properties of the system can be estimated, using molecular dynamics simulation. The molecular mechanics are extremely fast when comparing to quantum mechanics.

However, it cannot study transition states of systems where bonds break and form. Moreover, it requires extensive parameterization. Having the advantage of speed in MM and accuracy in the QM, the hybrid QM/MM approach allows the study of the complex system up to several 100,000 of atoms, while maintaining the accuracy at a certain part of the systems.

Finding saddle points in biomolecular systems are very important as indicated earlier in Chapters 1 and 2. Accurate estimation of the thermodynamic and kinetic properties of biomolecular systems necessitates first-order saddle points identification. In this chapter, we redefine saddle points in terms of biomolecular systems as transition states where the structure properties of the systems are considered as constraints that should be maintained.

Finding transition states in biomolecular systems using experimental methods is difficult. It becomes even extremely challenging, when dealing with small biomolecular systems. This is due to the fleeting nature of the transition states. Moreover, the short time span that a biomolecular system stays at the transition state makes it very challenging to capture. It is only few specialized experimental tools and equipment that are capable of capturing transition states. These equipments are only available at finger countable laboratories world-wide.

Alternatively, the biomolecular systems can be studied with help of modeling tools. Therefore, the algorithm proposed in Chapter 4 is extended to adapt the biomolecular systems. In particular, the algorithm shall incorporate an efficient molecular transition state local solver, duplicate checking and archival procedure with a structural similarity metric to enhance the ability of the algorithm to deal with biomolecular systems. Two

biomolecular systems are considered in this study: water clusters and small biomolecular systems, due to its biological importance.

This chapter is organized in the following manner. Section 5.1 shows experimental approaches used to discover the biomolecular systems structure and dynamics. Section 5.2 shows modeling approaches for studying biomolecular systems while different types of systems are reported in Section 5.3. A formalized definition of saddle points in the context of biological systems is given in Section 5.4. Section 5.5 extends the multi-modal memetic algorithm in Chapter 4 to adapt the biomolecular system. Experimental studies on the multi-modal molecular memetic framework for finding transition states on water clusters and covalently-bonded molecular systems show that the modifications are not sufficient. In particular, 90% of the population individuals after the initialization procedure are inappropriate or invalid molecular structures. This percentage increases while applying the evolutionary operators. This highlights a need for a specialized initialization and crossover operators for water clusters and covalently-bonded systems.

5.1 Experimental Approaches for Studying Biomolecular Systems

As highlighted previously, biomolecular systems can be studied experimentally using a different type of spectroscopies: Nuclear Magnetic Resonance (NMR) spectroscopy, Infrared (IR) spectroscopy and X-ray crystallography. NMR is an experimental technique built based on the phenomenon of nuclear magnetic resonance to study the physical and chemical properties of biomolecular systems. In particular, NMR provides exhaustive details about dynamics, structure, reaction state and chemical environment of molecules. NMR has also been widely used for inspecting the properties of organic molecules.

Another system for studying biomolecular systems experimentally is infrared (IR) spectroscopy [86]. IR spectroscopy uses the infrared region of the electromagnetic spectrum to characterize biomolecular systems. Based on the region of the infrared portion of electromagnetic spectrum, IR spectroscopy can be further classified into near, mid, far-infrared. The near-infrared spectroscopy uses frequency between 14000-4000 cm^{-1} and wave length between 0.8–2.5 μm wavelengths to excite or overtone harmonic vibration of biomolecular systems. The mid-infrared spectroscopy uses frequency 4000 and 400 cm^{-1} as well as 2.5-25 μm to study the fundamental vibration as well as associated rotational vibrational structures in bio-molecular systems. Finally, far-infrared spectroscopy has a frequency between 400 and 10 cm^{-1} and wavelength of 25-100 μm to study the rotational aspects of biomolecular systems.

Last, but the most common approach for characterizing the organization of atoms within a crystal structure is X-ray crystallography [87, 88]. It uses a beam of x-rays to strike at the crystal structure. As it crosses the crystal structure, the beam disperses into many different directions. From these directions and intensities of the diffracted beams, three-dimensional image of the density of the electrons can be fully developed. In particular, the mean position of the atoms in the crystallized structure can be precisely resolved as well as their chemical bonds, their orders and other information. Since its development, it has widely used to reveal the structures and functions of many biomolecular systems, including proteins and DNA.

5.2 *Modeling Approaches for Biomolecular Systems*

Alternatively, biomolecular systems can be studied through mathematical modeling of the interaction between constituting atoms and their subatomic structures. Three different modeling tools have been proposed in the literature, the first is the quantum-mechanical modeling, which is grounded on the quantum theory, in which the atomic and subatomic structures are modeled according to the duality of matter and wave. This can, practically speaking, model small system up to a few hundreds of atoms with computational tools available today. It can model how bonds form and break. The second is the molecular mechanical modeling, which can also model large complex systems, but it cannot model the bonds' breaking or formation. To model complex systems while maintaining focus upon the details of how bonds break and form, a hybrid QM/MM may be adapted. The modeling tools are detailed below.

5.2.1 Quantum Mechanics Modeling

The quantum-mechanical modeling (QM) [89] is a mathematical approach for describing both structural and dynamics of biomolecular systems. It considers the duality of matter and wave of the constituent atoms and subatomic structures. Its name originates from the observation that some physical quantities can switch only in discrete amounts. The angular momentum of an electron bound to an atom or molecule, for instance, is quantized. In the context of quantum mechanics, the dynamics and structure of photons, electrons, and other atomic-scale objects are unified under the wave–particle duality of energy and matter as well as the uncertainty principle.

QM is necessary for describing the chemical reaction and other processes, including electronic excitation and charge transfer. QM is very powerful and the most accurate model, proposed so far. However, it is extremely computationally-demanding and restricted to systems with a few hundreds atoms. It scales with $O(N^q)$ where q is the order larger than or equal to three, and N is the number of atoms.

5.2.2 Molecular Mechanics Modeling

Molecular mechanical modeling (MM) [2, 90] aims at describing the molecular structure and properties as practical as possible. The purpose is to study large molecules containing thousands of atoms for a long period of timescale with a reasonable computational cost. In this model, both nuclei and electrons are lumped into an atom-like particle, represented by a spherical object with radii attained from measurements and theory. In this model, the interaction between particles must be pre-assigned to a specific set of atoms and modeled as springs with the help of classical potentials.

The total energy in this system is described as a sum of Taylor's expansions for springs for a pair of bonded atoms, considering additional energy terms: bending, torsional energy, van der Waals energy, and electrostatics. By extricating out the van der Waals and electrostatic terms, molecular mechanic tries to craft the remaining constants more transferable among molecules.

MM has the advantage of great computational speed, which scales in form of $O(N^2)$. This subsequently allows molecular dynamic simulation, conformational energy searching and docking of complex biomolecular systems, which necessitate a large figure of

evaluations. In addition, MM allows the calculation of thermodynamics and kinetic properties via molecular dynamics simulation.

However, MM has also several pitfalls. Firstly, it requires extensive parameterization. Secondly, MM cannot be used for simulating chemical reaction where covalent bonds are broken or formed. Thirdly, the energy estimates obtained are not very precise. Fourthly, MM has a limited accuracy regarding the chemical environments. Finally, it can model the system only at the ground or stable states.

5.2.3 Hybrid Quantum Mechanics and Molecular Mechanic Modeling

QM/MM [91] is a combined approach of quantum and molecular mechanics. It is the perfect choice for modeling reactions in complex biomolecular systems. It enables the modeling for reactive biomolecular systems at a rational computation effort while maintaining the needed accuracy. QM/MM simply treats the chemically active regions as co-factors and substrates using QM while treating the surrounding solvent and proteins classically with MM. It has been found useful in studying hydrogen tunneling and has been used for calculating the hydride transfer in the alcohol dehydrogenase – an enzyme in liver. This approach can model system with up to a few hundreds of thousands atoms and allow simulations over time scales of tens of nanoseconds.

5.3 *Types of Biomolecular Systems*

Biomolecular systems are aggregates of chemical elements that participate in biological function. They can be classified into atomic and molecular clusters and covalently-bonded biomolecular systems. Atomic and molecular clusters are aggregation of some molecular units, enforced by weak forces. Among the most atomic and molecular

clusters, which of a great biological impact, is the water clusters. Covalently-bonded biomolecular systems can be classified according to the size of the system into small molecules such as drugs and amino acids; medium molecular systems such as peptides and small RNAs, and large molecular systems such as proteins. In this study, we focus on water clusters as well as small molecular systems, detailed in what follows.

5.3.1 Water Cluster

Water is the most abundant substance on earth. Almost seventy percent of the planet is covered by water. Cells, organs and any life forms are constantly in need for water. Without water, many chemical reactions could not take place, biological systems would not function, and life may not even have originated. However, the detailed, molecular-level role of water in the function of a chemical or biological system is an enigma and still a matter of great scientific debate [92, 93]. One of the central problems in biochemistry, for example, is to understand the structure and function of proteins at the molecular level. Many enzymes also catalyze biochemical reaction through water molecules. Protein synthesis requires the presence of several layers of water, as well. The explicit inclusion of water is essential in the unveiling of many problems in biochemistry [92]. Recent studies on climate changes have revealed that water vapors play a key role in amplifying global warming [94, 95], absorbing the emitted earth infrared radiation and remitted it back to earth. Therefore, a detailed knowledge about water is a prerequisite for a better understanding of many phenomena in many fields of science including biology, chemistry and climate sciences [92].

To investigate such fundamental knowledge of water, our research study starts with the simplest water systems, the water clusters. Water cluster is a discrete assembly of water molecules synergized together by hydrogen bonds [96], which has been a subject of increasing research investigations in the last few decades due to their importance in cloud formation, ice formation and melting, solution chemistry and a large number of biochemical processes [97]. The fundamental knowledge of water clusters can be investigated experimentally or computationally. Dealing with water clusters experimentally, however, necessitate sophisticated tools such as far-infrared and vibration-rotation-tunneling spectroscopy (an infrared spectroscopy technique) which is prohibitively expensive and in most cases unavailable. Moreover, when dealing with smaller clusters, such as those under 20 water molecule clusters, or hazardous chemicals/compounds, proves nearly impossible through experimental means [98, 99]. In such cases, computational study fills the gap perfectly.

5.3.2 Covalently-bonded Molecular Systems

Covalently-bonded molecular structures [100] are abundant in biology. They constitute important biological systems such as RNA, DNA, amino acids and proteins. Covalently-bonded systems refer to a group of biomolecular systems where atoms are aggregated together with covalent bonds. These systems can be classified according to their size in terms of atoms into small, medium and large systems. Small systems represent systems with a number of atoms below 100 atoms. Small systems include amino acids and small peptides. Such systems can be modeled with the quantum mechanical modeling while it is difficult to study such systems using experimental techniques. Medium covalently-bonded systems include those systems with a number of atoms larger

than 100 and less than 1,000. Such systems include large peptides as well as small protein. Large systems, on the other hand, include proteins possess atoms larger than 1,000. Both medium and large systems can be studied both experimentally using X-ray crystallography and computationally using molecular mechanical modeling and hybrid mechanical and quantum modeling.

In this study, we focus on the small ring-deficient molecular systems due to the important roles that they play in biology, medicinal, organic and computational chemistry, alongside their use as combinatorial building blocks in chemical synthesis.

5.4 *Transition States Identification in Biomolecular Systems*

The fitness or energy landscape has proven to be a useful conceptual framework in various fields, from biological evolution and protein folding to combinatorial and molecular optimization [101-103]. A landscape can be formally defined as an ordered set of three components $\mathbf{L} = (\mathbf{X}; f; \varphi)$, where $\mathbf{X}$ is the set of possible solutions or structural configurations, $f(\mathbf{x})$ is the fitness function, and φ denotes the distance measure between two structural configurations in $\mathbf{X}$.

5.4.1 Structural Configuration/Representation Space and Potential Energy Function

The structural configuration space $\mathbf{X}$ is the set of physically consistent 3D structures. Each structure $\mathbf{x} \in \mathbf{X}$ is a vector of $3n$ real numbers representing each atom's geometrical coordinates in three dimensional space, measured in Angstroms ($\text{\AA}$), that is, $\mathbf{x} \in \mathbf{R}^{3n}$, where n is the number of atoms in the molecule.

The potential energy function, $f = f(\mathbf{x}): \mathbf{X} \rightarrow \mathbf{R}$, gives an indication on the height of the energy landscape. Since our goal is to locate transition states, which may be associated with bond breaking and formation, we will thus work with quantum-mechanical modeling approach called Hock-Flock Tree that calculates the binding energies of the given molecule at STO-3G basis set [104].

5.4.2 Transition State Sampling and Optimization

Transition state denotes the configuration where the minimum energy barrier is needed to cross from a low-energy vicinity of the reactant to other low-energy vicinity of the product. Transition states play a key role in theoretical and physical chemistry. Their molecular structure and energy help in the identification of reaction rates [105] and minimum energy paths [106].

Mathematically, transition state can be defined as a stationary point ($\mathbf{x}_i$) on landscape $\mathbf{L}$ with only one negative eigenvalue in the Hessian matrix [14], and can be expressed as:

$$\mathbf{X}_T = \left\{ \mathbf{x}_i \mid \left(\left\| \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{x}_i} \right\| = 0 \right) \& (\gamma_{ik} < 0) \& f(\mathbf{x}_i) < f_{max} \right\} \quad (\text{Eq. 5.1})$$

where $\mathbf{x}_i \in \mathbf{R}^d$, d is the dimensional size, $f(\mathbf{x}_i) \in \mathbf{R}$, $\mathbf{X}_T$ is the set of transition states, and γ_{ik} is the only negative eigenvalue of Hessian matrix $\mathbf{H}_i$, while ‘&’ denotes a logical AND operator.

More often, scientists are generally interested on specific transition states that are deem as “good”. To formalize the concept of “good” transition states, we define the ideal sample set $\mathbf{X}_T$ having the following properties:

- contains all structures with energies below a user-defined threshold, i.e. ($f(\mathbf{x}_i) <$

f_{max}).

- contains no duplicates, i.e., there do not exist any $(\mathbf{x}_i, \mathbf{x}_j) \in \mathbf{X}_T$ s.t. $\|f(\mathbf{x}_i) - f(\mathbf{x}_j)\| < j$ and $\varphi(\mathbf{x}_i, \mathbf{x}_j) < \vartheta$.
- has a precision of $\left\| \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{x}_i} \right\| < \delta$.

where j and ϑ are the maximum acceptable similarities in the energy and configuration spaces, respectively, δ defines the precision of the gradient norm and $\varphi(\mathbf{x}_i, \mathbf{x}_j)$ denotes a structure similarity metric.

In the present study, the Ultrafast Shape Recognition (USR) is considered. USR measures the similarity between molecular structure shapes with a signature vector of twelve atomic distance statistics, $\mathbf{U}$. This signature captures the mean, standard deviation and asymmetry of the distances from each atom in the structure to four anchor points. The anchor points are a (the structure's centroids), b (the atom closest to a), c (the atom furthest from a) and d (the atom furthest from c). This signature gives an effective measure since it is invariant to translational and rotational symmetries. Hence, the similarity $s(\mathbf{x}_i, \mathbf{x}_j)$ between two molecular structures $\mathbf{x}_i$ and $\mathbf{x}_j$, as inverses of distances between the signatures, can be easily defined by the inverse-scaled Manhattan distance as follows:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{1 + \frac{1}{12} \sum_{k=1}^{12} |\mathbf{u}_k^{\mathbf{x}_i} - \mathbf{u}_k^{\mathbf{x}_j}|} \quad (\text{Eq. 5.2})$$

where $\mathbf{u}_k^{\mathbf{x}_i}$ denotes the k^{th} component of $\mathbf{x}_i$'s USR signature. Using the signature in Eq.

5.2, the distances between structures can be derived as:

$$\varphi(\mathbf{x}_i, \mathbf{x}_j) = 1 - s(\mathbf{x}_i, \mathbf{x}_j) \quad (\text{Eq. 5.3})$$

$\varphi(x_i, x_j)$ maps $[0; 1)$ with 0 indicating maximum similarity. $\varphi(x_i, x_j)$ is also symmetric since $s(x_i, x_j) = s(x_j, x_i)$. Note for convenient description of structure similarity, we consider $s(x_i, x_j)$ in percent. In case that $s(x_i, x_j) = 0.73$, one can say there are 73% structure similarity between x_i and x_j .

Also, It follows naturally that the binding energies of structures in the ideal sample are bounded by f_{max} and j . A good sample set ($\mathbf{X}_k$) should thus be one that approximates the ideal set $\mathbf{X}_T$ sufficiently well. Hence, one can test for closeness to $\mathbf{X}_T$ by measuring the cardinality of $\mathbf{X}_k \cap \mathbf{X}_T$.

5.5 *Molecular Extensions for Multi-Modal Memetic Framework*

Here the multimodal memetic framework proposed in Chapter 4 has been extended with the Berny algorithm — a modified version of Schlegel algorithm which is more efficient when dealing with complex molecular systems — and a duplicate checking procedure which is updated to use the structure similarity measure instead of the Euclidian distance, as detailed below.

5.5.1 Life-Time Learning Procedure

The Berny algorithm is considered here as the life-time refinement procedure in this study due to its ability in converging to precise transition states efficiently. Considering the valley groups V_{groups} identified previously in valley-adaptive clearing step, detailed in Chapter 4, the elite individual ($\mathbf{x}_{elite}$) of each valley group will undergo the Berny algorithm to arrive at a precise transition state found in the near vicinity of $\mathbf{x}_{elite}$. The Berny geometry optimization algorithm is a modified version of the Schlegel algorithm

[12], which operates by altering the signs of eigenvalues, in the event, more than one negative eigenvalue is encountered. All negative eigenvalues are then changed to their absolute value, excluding the smallest [41]. On the other hand, there is no negative eigenvalue; the sign of the least positive eigenvalue is negated.

For each $\mathbf{x}_{elite}$ vector, the Berny algorithm proceeds based on the energy and its derivatives at $\mathbf{x}_{elite}$ and iterates with the following cycle of steps:

- 1) Hessian matrix is checked whether it has the right inertia (i.e., having only one negative eigenvalue)
- 2) If so, the quadratic step is determined,
- 3) Otherwise, the rational function optimization approach [25] is employed,
- 4) If the quadratic step exceeds the trust radius, the quadratic step is scaled down to the trust radius,
- 5) Eliminate frozen variables and their corresponding gradient and step components,
- 6) Determine $\mathbf{g}_{k+1}$ and approximate Hessian matrix $\mathbf{H}_{k+1}$ from $\mathbf{H}_k$, $\Delta\mathbf{x}_k$, $\mathbf{g}_k$, and $\mathbf{g}_{k+1}$ using Bofill updates.
- 7) Update trust radius using the method of Fletcher [26].

The cycle or life-time learning procedure iterates until a convergence in term of maximum number of iterations or precision configured is fulfilled.

5.5.2 Identification of Duplicate Molecular Structure Configurations and Archival Procedure

During duplicate identification and replacement process, all non-duplicated discovered structures are archived using a multi-map data structure. Multi-maps are associative data structures that store elements indexed by keys to ensure that comparisons and duplicate checking are performed efficiently. Multi-maps, in particular, allows for quick access and rapid retrieval of archival elements based on key values. From the computational cost prospective, the worst case access time is $O(m + 1)$ where m is the number of keys and worst case insertion time is $O(\log s_n)$ where s_n is the number of archived solutions. In this study, the archive is indexed based on the number of significant figures of binding energies as well as USR.

5.6 *Experimental Study*

To assess the efficacy of the proposed algorithm, empirical study was performed using several examples of biomolecular systems detailed in Section 5.6.1. The potential energy as well as its gradients and eigenvalues were calculated using an ab-initio approach at the Hartree-Fock level of approximation with STO-3G basis set, which is available in the Gaussian 09 software package. With 10% mutation probability chosen to balance between exploration and exploration during the search, a population of 10 individuals is allowed to explore the energy landscape of several benchmark molecules described in Section 5.6.1. The algorithmic parameter settings used are as tabulated in Table 5-1. Empirical results on the biomolecular systems considered are reported in Section 5.6.2.

5.6.1 Benchmark Example Biomolecular Systems

Here, we consider two different sets of biochemical systems. One represents the atomic and molecular clusters and the other represents the small covalently-bonded molecular systems, as detailed below.

Table 5-1: Molecular parametric settings of the algorithms.

Parameter	Value	Justification
ϑ	0.05	This structure similarity threshold is widely used structure similarity threshold to distinguish between structurally-similar and structurally-different molecules regardless of the molecules orientation. i.e. this the structure similarity threshold that should be maintained between two different molecules in the population.
j	0.0001	The minimum Euclidean distance that should be between two different molecules.
δ	3.5×10^{-4}	Widely used gradient precision for characterizing stationary points in biomolecular systems.
f_{max}	-100	Energy upper bound below which the search algorithm would explore for solutions.

A. Water Clusters

In this study, we consider three different water cluster problems, namely, Dimer, Trimer, and Tetramer (See Figure 5-1).

- 1) Dimer (See Figure 5-1.A) consists of two water molecules or six atoms with 18 degrees of freedom.
- 2) Trimer (See Figure 5-1.B) consists of three water molecules with 27 degrees of freedom,
- 3) Tetramer (See Figure 5-1.C) consists of four water molecules and has a total of 36 degrees of freedom.

Figure 5-1: An illustrative example of the A) Dimer water cluster, B) Trimer water cluster, and C) Tetramer water cluster.

B. Covalently-bonded Molecular systems

Here, we study several important ring-deficient covalently-bonded small molecules, with significant impact on human health. These systems as shown in Figure 5-2 include GABA, GABOB, Leucine, Tromethamine, Valine, and Glutamic acid which are listed and detailed as follows:

- 1) **Gamma-Aminobutyric Acid (GABA):** GABA is a chief inhibitory neurotransmitter in the central nervous system of mammals, regulating neuronal excitability in the system. It is also well-known for its pharmacological actions on GABAergic systems.
- 2) **GABOB:** GABOB is a derivative of the neurotransmitter GABA, and known as gamma- aminohydroxybutyric acid or Buxamin. It works mainly as an anti-epileptic drug, offering a number of significant de-stressing and anti-aging benefits.
- 3) **Leucine:** Leucine is an essential amino acid, which is essential for hemoglobin formation. It also activates the mammalian target of rapamycin kinase, regulating cell growth. Leucine, on the other hand, may be toxic. In particular, Leucine toxicity may lead to delirium and neurologic compromise, and can even be life-threatening.

- 4) **Valine:** Valine is another essential amino acid which is characterized with stimulant activity. It regulates the immune function of the body and helps in the development and growth of the muscle as well as tissue repair. It is also a main precursor in the penicillin biosynthetic pathway. It is also useful in treatments involving emotional upsets, insomnia and nervousness.
- 5) **Tromethamine:** Tromethamine is a basic compound which is typically used as an alkalizing agent and as a buffer in enzymatic reactions and affects the balance of water and electrolytes in the body. It is broadly used in the synthesis of surface-active agents and pharmaceuticals, in particular, the treatment of metabolic acidosis.
- 6) **Glutamic acid:** Glutamic acid is the most abundant amino acid in the mammalian central nervous systems (CNS) where serving as an excitatory neurotransmitter. Recent studies have shown its important role in functions of learning and memory alongside its major play in the pathogenesis of many neurodegenerative disorders. Unlike most neurotransmitters, glutamic acid is characterized by a high degree of chemical flexibility resulting in the presence of many low-energy conformations. These allow selective interactions with various transporters and receptor proteins and altogether incorporate to compose the glutamatergic system within the CNS. Identifying such low-energy stable conformations and their associated transition states of glutamic acid may open new doors to better understand the interactions taking place inside the glutamatergic system in a way to know more about cognitive functions and learning processes. Moreover, this identification can be a step forward towards the design of new, yet effective drugs for many neurodegenerative disorders.

Figure 5-2: Ring-deficient covalently-bonded small molecules.

5.6.2 Results

In this section, relative performance of the proposed memetic algorithm as compared to the state-of-the-art Sequential Niching Memetic Algorithm (SNMA) and the

conventional Stochastic Multi-start Local Search (SMLS) is presented. Figure 5-3 quantifies the average performance over ten independent runs in terms of success rate (SR), number of discovered transition states (#TS) during the course of finding the transition states. With 30% success rate, the multi-modal molecular memetic framework (MMA) has shown to outperform the other algorithms significantly, discovering as many as up to two folds more transition states discovered by the other competing algorithms.

Although the algorithm outperforms other algorithms in terms of success rate and number of transition states uncovered, many rooms are still there to improve the performance. To identify these potential areas to investigate, further study is investigated on the initialization procedure as well as crossover and mutation operators. In particular, we are interested on the validity of the molecular structures generated. Valid molecular structure is the one that avoid the steric effect, i.e. no overlap between atoms, and maintains bonds, while the nonsensical molecular structure is the one that does maintain either one of these conditions. Figure 5-4 shows example valid molecular structures and nonsensical structures.

To study the effect of the initialization procedure, a population of 100 individual are randomly sampled from the rectangular search space in the range of $-10\text{\AA}$ and $10\text{\AA}$. The number of valid molecular structures and nonsensical structures for each biomolecular system is depicted in Figure 5-5. The figure clearly indicates that the random initialization procedure seldom generate a valid molecular structure, with one valid structure for 10 nonsensical structure in the best case.

Figure 5-3: Performance of different algorithms.

Also, the crossover operator may have room to improve the proposed algorithm. To confirm this hypothesis, a population of a hundred valid structures is examined for different biomolecular systems. The population is then allowed to undergo only the crossover operations. The number of valid molecular structures as well as the nonsensical structures is then counted and reported in Figure 5-6. According to the figure, less than one fifth of the population has been maintained in valid molecular structures while the rest become nonsensical structures. Similar percentage has been also observed for mutation operator in Figure 5-6 as well.

Figure 5-4: Nonsensical molecular structures and valid GABA structures.

Figure 5-5: Percentage of valid and nonsensical molecular structure generated after initialization.

Figure 5-6: Percentage of valid and nonsensical molecular structure generated after both crossover and mutation.

These results highlight the role of the initialization, crossover and mutation operators in the evolutionary search. Implementing those operators in a way that consider domain knowledge would be our immediate interest.

5.7 Conclusion

In this chapter, we have discussed the biomolecular systems and the approaches used to study it both experimentally and computationally. We highlighted the issues with experimental approaches when dealing with small biomolecular systems and how computational methods could be ideal solutions. Different modeling frameworks have been discussed. The most accurate modeling tool, the quantum mechanical modeling, has been employed, targeting small water cluster and covalently-bonded systems. A formal definition of the transition states in biomolecular systems has been presented. Two modifications to the multi-modal memetic framework proposed in Chapter 4 have been proposed to locate transition states in biomolecular systems. The modified algorithm has been employed on Water clusters and covalently-bonded systems. The results have shown that the initial population contains more the 80% of invalid molecular structures, which increases as the evolutionary process proceeds. This highlights the need for a specialized initialization procedure and evolutionary operators.

In the following chapters, we propose specialized algorithms that incorporate domain knowledge to tackle the water cluster in Chapter 6 and covalently-bonded molecular systems in Chapter 7. Last but not the least, thesis is concluded in Chapter 8 and several future research investigations are also provided in Chapter 8.

Chapter 6: Evolutionary Discovery of Transition States in Water Clusters

As shown previously in Chapter 5, water clusters are very important biomolecular systems. The difficulty in studying such systems using experimental approaches has led us to the computational studies of water clusters. At the core of the computational studies, optimization plays a key role, particularly in the identification of the basic fundamental knowledge on water molecular structures, including the stable, meta-stable and transition states [101, 102, 107]. In the literature, however, the major attention has been focused on identifying water global minima and isomers [101, 102, 107, 108]. Global optimization of water clusters seeks for the water cluster configuration with the lowest potential energy, while the isomer optimization task involves locating meta-stable states (multiple local minima) of water clusters. Both tasks are generally formulated as minimization search tasks. Transition state discovery, on the other hand, has been established as a much more arduous optimization task than the formers. The core challenges emerge from the nature of transition state optimization, which involves both maximization and minimization at the same time, making it more demanding than traditional optimization of maxima or minima. The presence of a large number of critical points, including maxima and minima, further hampered the progress of search for transition states. Finding transition states of water clusters using high fidelity first-principles calculations, specifically, also makes the search even more challenging due to the high computational cost of solving the Schrödinger equations. The challenge escalates exponentially with increasing water clusters size, since any addition of a water molecule (H_2O), which is composed of two

hydrogen and one oxygen atoms, increases the dimensionality of the search by nine degrees of freedom based on a Cartesian coordinate representation.

To efficiently identify transition states in water clusters, this chapter extends the multi-modal memetic framework, proposed in Chapter 4 and molecularly-extended in Chapter 5, with a water-specialized molecular cluster initialization procedure to generate favorable initial water cluster structures and several novel molecular reproduction operators designed for water molecular clusters evolution, where the rules and constraints of molecular water clusters are incorporated. In particular, the proposed molecular crossover is devised to enable good quality molecular substructures to replicate across generations. The mutation operators, on the other hand, are designed to enhance solution diversity at different levels of granularities as inspired by the natural dynamics of water molecules and their constituting atoms.

We demonstrated the efficacy of the extended framework (MOL-MA) against other recent algorithms proposed in the literature on first-principles calculations of water cluster, and performed extensive numerical studies (totaling more than 60 days of CPU time) for pure water cluster $(\text{H}_2\text{O})_n$, $n = 2-4$. The transition states for water clusters up to four water molecules uncovered using MOL-MA are then reported. In particular, the search performances of MOL-MA and other counterpart algorithms considered for transition states discovery are summarized in terms of solution precision, computational effort and number of uncovered transition states. The results obtained show that MOL-MA not only reproduced previously found transition states in water clusters but also

discovered new transition states for sizes 2-4 water molecules, compared to those known to date [102, 109].

The chapter is organized as follows: The water-specially-designed operators are presented in Section 6.1, while Section 6.2 gives the results obtained from our computational study. Section 6.3 shows cases of the transition states uncovered, with a novel methodology for the case of transition states clustering based on morphological structure similarities. The brief conclusion is then stated in Section 6.4.

6.1. Water-Specially-Designed Operators

In this section, we extend the general framework proposed in Section 4.4 and depicted in Figure 4-9, considering the molecular operators introduced in Section 5.5 with water specially-designed initialization and evolutionary operators. We donate this extended form of the framework as (MOL-MA) [110]. In MOL-MA, a population of molecular water clusters is first randomly generated using a specialized water cluster initialization procedure, which will be detailed in Section 6.1.1. The water clusters are then assessed as candidate solutions for the problem at hand using the proposed fitness function, which is described in Section 4.4. Water clusters in the population survive to the next generation according to the stochastic universal sampling procedure. The surviving water clusters then undergo the specially designed molecular-based water evolutionary operators of crossover and mutation (as detailed in Section 6.1.2). To facilitate diversity in the search, evolved water clusters then undergo the valley-adaptive clearing scheme (Section 4.2) where population individuals belonging to the same valley are segregated from others into niche or valley groups. The elite individual per valley group is then refined using the

recent development of Schlegel algorithm (Berny algorithm) and Lamarckian inheritance [111], presented in Section 5.5.1. Valley elites that satisfy as transition states are archived and all other members of the valley groups are relocated randomly to other regions of the energy landscape to facilitate search exploration towards the discovery of other potential high quality transition states that may exist. In the search process, any newly uncovered transition states are archived and checked against possible duplicate, presented in Section 5.5.2. This process repeats for a maximum number of generations or until other convergence criteria are satisfied.

6.1.1 Molecular Water Clusters Initialization

Using the domain knowledge and experiences of physicists and chemists pertaining to water molecular clusters [112], a set of initial water clusters consisting of n water molecules is generated to form the initial MA population. Particularly, a water molecule with centroid vector (x,y,z) is first randomly generated within a pre-defined sphere bounded by radius r . An example structure of the water molecule is depicted in Figure 6-1, with bond length bounded by $[b_{min}, b_{max}]$ and angle α in the range $[\alpha_{min}, \alpha_{max}]$. All subsequent molecules of the MA population are then generated randomly such that the distances to previously generated molecules fulfill the following constraints imposed:

1. A minimum (O-O) distance between two neighboring oxygen atoms of $O-O_{min}$.
2. A minimum (O-H) distance between oxygen and hydrogen atoms of two neighboring water molecules is $O-H_{min}$ [113].

An example initial population for four water cluster or $(H_2O)_4$, consisting of 12 water clusters (W_c), is depicted in Figure 6-2.

$$0.95 \mp 0.1 \text{ \AA}$$

Figure 6-1: Basic water monomer.

6.1.2 Molecular-Based Water Evolutionary Operators

The novel molecular reproduction operators are also designed specially based on the experiences of physicists and chemists for water clusters evolution, where the rules and constraints of water clusters are incorporated into the MOL-MA, which are described in details in the present subsection.

C. Molecular Crossover

In this subsection, we introduce a novel crossover operator specially designed for molecular water clusters as outlined in Table 6-1. The crossover process between two water clusters is depicted in Figure 6-3. The aim is to exchange some of the water molecular substructures between clusters, thus allowing successful substructures to replicate across populations and generations. Two water clusters in the population are, therefore, selected and then aligned such that they share a common center of mass. Two random perpendicular vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ crossing the center of mass are subsequently identified. Each cluster is then randomly rotated around $\mathbf{v}_i$. A surface of slicing is then calculated based on the center of mass and $\mathbf{v}_i$. Each water cluster is subsequently sliced into two sections: above and below the surface of slicing, forming chromosome segments or molecular substructures. Such segments are then exchanged between the parent water clusters.

Figure 6-2: An example initial population consisting of 12 water chromosomes (W_c), on the four water cluster problem or $(H_2O)_4$.

Table 6-1: Molecular water clusters crossover.

Input: Two water clusters or chromosomes: W_{c1} and W_{c2}	
Begin:	
1.	Calculate center of mass for W_{c1} and W_{c2}
2.	Rotate W_{c1} and W_{c2} around v_1 and v_2 , respectively
3.	Move water clusters such that center of mass lies at the origin (0.0, 0.0, 0.0)
4.	Calculate surface of slicing ³
5.	Dividing the chromosomes/ water clusters into segments/ substructures
6.	Exchanging substructures between clusters
End	

³ Surface of slicing is a plane that crosses the center of mass and v_i .

Figure 6-3: An illustration of the molecular crossover operator on two water clusters or (H_2O)₂.

D. Molecular and Atomic Mutations

Analogous to biological mutation, the mutation operator in EAs serves to introduce subtle changes to a randomly chosen subset of chromosomes in the MA population or water clusters in the present context. In this study, we propose four molecular and atomic mutation schemes for water clusters evolution and categorized them into rotation-based and translation-based operators, which are detailed in what follows:

1. Rotation-based operators

- 1.1 Rotation of a molecule (See Figure 6-4.A) around the center of mass of the entire water cluster (or chromosome) (See Figure 6-4.B). In particular, the entire molecule is rotated by random angle σ_1 around the center of mass of the water clusters. Such rotation serves to break the hydrogen bonds of the rotated water molecule, thus leading to the identification of transition states associated with hydrogen bond breaking. In a similar manner, transition states of hydrogen bond formation between far apart water molecules may be uncovered in the process.
- 1.2 Rotation of a molecule around the center of mass of the molecule (or sub-segment of the chromosome) (See Figure 6-4.C). Rotating a single water molecule by a random angle (σ_2) around its center of mass allows diversified configuration at the scale of water molecule. In addition, it also reconfigures the hydrogen bonds interaction of the water molecule with the neighboring molecules.

2. Molecular and atomic translation-based operators

- 2.1 Translation of a molecule: Linear translation of a water molecule in the cluster (See Figure 6-5.A) by random vector $\mathbf{t}$ such that $\|\mathbf{t}\| \leq \tau \text{ \AA}$ (See Figure 6-5.B).

2.2 Translation of an atom: The coordinates of a randomly chosen hydrogen atom in the molecular cluster is perturbed (See Figure 6-5.C).

Figure 6-4: An illustration of rotation-based mutation operators for water cluster A) a two water cluster molecular structure, B) upon rotation of a molecule, around the center of mass of the water cluster, and C) upon rotation of a water molecule in the water cluster, around the center of mass of the water molecule.

Figure 6-5: An illustration of translation-based mutation operators for water cluster A) a two water cluster molecular structure, B) upon translation of a random molecule in the water cluster, and C) upon translation of a random atom in the water cluster.

6.2. *Experimental Study*

In this section, we study the efficacy of the proposed molecular memetic algorithm (MOL-MA) for locating transition states and pit it against several existing state-of-the-art transition states optimization methods, including the stochastic multi-start local search Berny Algorithm (SMLS-BA) [82] as representative of single-ended methods, GRRM [114] as representative of doubly-ended methods, while Chaudhury GA (ChGA) [21] and Bungay GA (BGA) [19] as representatives of the genetic algorithm. The sequential niching memetic algorithm (SNMA) [83] and Mahalanobis with self-adaptive-niche radius CMA-ES (M-S-CMA-ES)[85], as representatives of recent advances in niching and memetic algorithms, are also considered for comparison in the present study. Brief descriptions of the algorithms considered are given in Table 6-2.

To ensure a fair comparison, the advanced fitness function in Eq. 4.8 is considered for all the algorithms. All methods act as baselines for comparison on realistic water clusters transition states optimization of up to four molecules, using a comprehensive set of performance measures that are detailed in Section 6.2.2. The energy, gradient and eigenvalues are calculated using first-principles approach at the Hartree–Fock level, with the STO-3G basis set [104], and all computations are executed on the NTU-HPC cluster⁴. For the purpose of statistical significance, the experimental results summarizing 10 independent experimental runs are reported in Section 6.2.3. Note that more runs are deemed to be impractical due to the intractable computational resources that would be required on the computational expensive high-fidelity water models considered in the present study.

⁴ NTU-HPC cluster has 2 *headnodes* (x3650M2) with 300 *compute nodes* (2 x *Quad-Core Nehalem x5570*, 2400 *cores*, 2.93Ghz, 16GB RAM *per node*), *SMP Server* (x3950M2, 48 *cores*, 2.66Ghz, 256GB RAM) and *Power Server* (IBM 550, 16 *cores*, 4.2Ghz, 64GB RAM)

6.2.1 Water Cluster Test Problems

In this study, we consider three different water cluster problems, namely, Dimer, Trimer, and Tetramer (See Figure 6-6). Dimer consists of two water molecules or six atoms with 18 degrees of freedom; Trimer consists of three water molecules with 27 degrees of freedom, while Tetramer consists of four water molecules and has a total of 36 degrees of freedom. The complexity of first-principles calculation is $O(N^4)$, where N is the number of atoms.

Table 6-2: Descriptions of the state-of-the-art algorithms considered.

Category	Algorithm	Description
Single-ended Methods	Stochastic Multi-start Local Search (SMLS)	A typical stochastic multi-start solver with the Berny algorithm considered for locating transition states.
Doubly-ended Methods	GRRM [114]	An advanced doubly-ended method that assumes the availability of knowledge about the reactant and product of molecular systems when finding the transition states that lie in-between.
Genetic Algorithm	Bungay Genetic Algorithm (BGA [19])	A binary-coded GA proposed by Bungay to find the transition states of some chemical reactions such as HNC isomerization.
	Chaudhury Genetic Algorithm (ChGA [21])	A real-coded genetic algorithm proposed to find the transition states on Lennard Jones clusters.
Recent Advances in Niching and/or Memetic Algorithms	Sequential Niching Memetic Algorithm (SNMA) [82]	An advanced niching technique that extends the effort of Beasley et al. [83] on sequential niching technique to locate multiple optimal using a local search to improve solution precision. SNMA incorporates a gradient-based local search process, a derating function and the basic clearing technique.
	M-S-CMA-ES, Mahalanobis Self-adaptive niche radii CMA-ES [85]	An advanced niching method which extend CMA-ES with adaptive niche radius strategy coupled with Mahalanobis distance for solving complex, non-isotropic high-dimensionality multi-modal optimization problems.

To provide an estimate on the computational cost of deriving the potential energy, and hence the fitness value $f_{ts}(\mathbf{x})$ (based on Eq. 4.8) of the water clusters when employing high fidelity first principle calculations, the wall clock time incurred by the different stochastic algorithms considered are summarized in Figure 6-7. From the figure, it is worth noting the significantly high resources used for conducting the present

computational studies, where the incurred wall clock time per run, per algorithm, and per $(\text{H}_2\text{O})_n$ can extend to several hundreds of hours. Thus, the entire experimental study in total took more than 60 days of computations to complete, for pure water clusters $(\text{H}_2\text{O})_n$, $n=2-4$, despite the use of high performance computing facilities.

6.2.2 Performance Measures

Performance measures are quantitative tools that enable fair comparisons among algorithms. Three different performance measures that are widely used in the context of optimization include solution precision, percentage of uncovered transition states, and computational cost, are considered here in the present study. These performance measures are briefly described in Table 6-3.

Figure 6-6: An illustrative example of the A) Dimer water cluster, B) Trimer water cluster, and C) Tetramer water cluster.

Figure 6-7: Mean wall clock time of 10 independent runs incurred by MOL-MA, SNMA, SMLS, M-S-CMA-ES, BGA, and CHGA on the Dimer, Trimmer and Tetramer water clusters.

Table 6-3: Descriptions of the performance measures considered.

Performance Measure	Description
Percentage of Uncovered Transition States	Percentage of uncovered solutions (%TS) or water transition states [84] is defined here as the ratio of, the number of water transition states (n_{ts}) successfully uncovered by an algorithm, to the maximum number of water transition states (n_{max}), uncovered throughout the experimental studies or $\%TS = \frac{n_{ts}}{n_{max}} * 100$. n_{max} is used instead since no knowledge on the true number of transition states that exists in the water clusters is available to date. From our studies, n_{max} of transition states for Dimer, Trimmer and Tetramer are found to be 5, 10 and 25, respectively.
Computational Costs	The resources required by an algorithm in identifying m transition states. As first-principles calculations represent the most computationally expensive part of computation, here the computational costs is reported in terms of the number of gradient evaluations incurred.

6.2.3 Numerical Results

In this section, the experimental settings and numerical results obtained are presented and analyzed. Beside the algorithmic parameters considered in Table 4-8 and Table 5-1, the parameteric configurations for MOL-MA used in the present study are summarized in Table 6-4. Next, we present on the numerical results of MOL-MA pitted against other state-of-the-art evolutionary and memetic approaches for the discovery of transition states in the Dimer, Trimmer and Tetramer water cluster problems. For the sake of brevity, the

unique properties of the algorithms considered in the present study, i.e., in terms of life time learning, specialized molecular based water evolutionary operators, niching operators and domain knowledge required are summarized in Table 6-5.

Table 6-4: Parametric settings of the algorithms. These settings are obtained from the domain knowledge of scientists and physicists on water clusters.

Parameter	Value
b_{min}	0.85 Å
b_{max}	1.1 Å
α_{min}	80 ⁰
α_{max}	120 ⁰
τ	5 Å
$P_{mutation}$	10%
$P_{crossover}$	90%
$O-O_{min}$	2.3 Å
$O-H_{min}$	1.5
ϑ	0.05

Table 6-5: Properties of the algorithms considered (Y denotes the existence of a property while N refers to its absence).

	SMLS	GRRM	BGA	ChGA	M-S-CMA-ES	SNMA	MOL-MA
Life-time learning	Y	Y	N	N	N	Y	Y
Specialized Molecular-based Water Evolutionary Operators	N	N	N	N	N	N	Y
Niching Operator	N	N	N	N	Y	Y	Y
Domain Knowledge Required	Initial guesses	Reactant and product	Initial guesses	Global Optima	N	N	N

A. Percentage of Uncovered Water Transition States

The performance efficacy of the algorithms assessed according to their ability in uncovering multiple high quality transition states is reported in this subsection. In Figure 6-8, the mean percentages of transition states uncovered by the respective algorithms are depicted. From the results, BGA, ChGA, and M-S-CMA-ES are observed to fail in locating any transition states on the Dimer, Trimer and Tetramer water problems. This is a result of the poor performance of BGA, ChGA, and M-S-CMA-ES in converging to highly precise solutions or transition states efficiently.

Algorithms including GRRM, SMLS, SNMA, and MOL-MA, in contrast, include a life-time learning procedure in their design, which enable them to attain high precision solutions efficiently, thus leading to the larger percentage of transition states uncovered. Among them, the memetic algorithms (SNMA and MOL-MA) attain relative higher percentage of transition states uncovered. MOL-MA, among all, achieved the highest percentages of precise transition states uncovered, by means of the life-time learning procedure, specialized water operators and valley-adaptive clearing scheme, at 100%, 98%, and 95% on the Dimer, Trimmer and Tetramer water cluster problems, respectively.

Further, GRRM, SMLS, SNMA, and MOL-MA, despite showing some success in locating transition states, are observed to vary in percentage of transition states uncovered or (%TS) on the different water problems. Such variation indicates the degree of robustness in the algorithms. Clearly, the lower is the variation exhibited by an algorithm on the different water cluster problems, the more robust is the algorithm. From Figure 6-8, MOL-MA is deemed as most robust due to the low variation of the standard deviation in %TS. Note that the observed lower quartile on the percentage of transition states for MOL-MA is higher than any upper quartile on the uncovered percentage of transition states observed on all the counterpart algorithms.

Figure 6-8: Box plot on the percentage of transition states uncovered (%Ts) by each algorithm on the Dimer, Trimmer, and Tetramer water cluster problems.

B. Computational Costs

The computational requirements of the algorithms considered in the present study are also summarized in Figure 6-9, except for BGA, ChGA, and M-S-CMA-ES, since these latter algorithms failed to uncover at least one transition state, despite using up the maximum computational budget of 2.5×10^5 gradient evaluations allowable. The three subplots A, B and C, of Figure 6-9 depict the computational requirements of the algorithms on the Dimer, Trimmer and Tetramer water cluster problems, respectively. From the results, SMLS is observed to incur the most computational effort on the water problems and exhausted all computational budget allowable on Tetramer (at the expense not even locating a single solution or transition state). Apart from Dimer, GRRM is observed to be computationally more efficient than SNMA and SMLS, under the assumption that *a priori* knowledge on where the reactant and product lie is available.

Note that such information is often scarce in reality. In spite of the largest percentage of precise transition states uncovered as discussed in the previously two subsections, MOL-MA remains to display superiority in terms of computational efficiency over the other state-of-the-art algorithms, on all the water cluster problems considered. In particular, the cost reductions of MOL-MA over the other counterpart algorithms on the water cluster problems when summarized in Table 6-6 indicate that MOL-MA maintains an average cost reduction of 84%, 161% and 214% over GRRM, SNMA and SMLS, respectively.

Table 6-6: Cost reduction in (%) by MOL-MA over the other counterpart algorithms (N/A indicates not applicable as the algorithm fails to locate any transition states).

	SNMA	SMLS	GRRM
Dimer	212.5%	462.5%	N/A
Trimmer	116%	190%	50%
Tetramer	157%	N/A	118.5%
Average	161%	298.5%	84.25%

Figure 6-9: Box plot on the number of gradient evaluations count incurred by the algorithms considered on the Dimer, Trimmer, and Tetramer water cluster problems.

6.3. Visualization of Discovered Water Cluster Transition States

In the present study, MOL-MA is shown not only to reproduce previously known transition states in in Dimer, Trimmer and tetrammer water clusters, but also established newly discovered transition states. To present these uncovered transition states for further study by scientists and physicists, typical approaches based on graph theory can be used [115]. In contrast to existing approaches that requires *a priori* knowledge on the classes of water clusters, we used agglomerative hierarchical clustering group the water transition states based on their structural similarity. Dendrograms of the hierarchical clustering for the $(\text{H}_2\text{O})_n$ transition states are then derived in Figure 6-10:12 to depict the water cluster transition states.

Figure 6-10 depicts the dendrogram of the Dimer problem (using hierarchical clustering), where the transition states of water clusters are classified into two main groups: the single hydrogen bonded group (SHBG) and double hydrogen bonded groups (DHBG). The double hydrogen bonded group can be further segregated into parallel hydrogen bonded group (PHBG) (where hydrogen bonds lie in parallel to each other as shown in Table 6-7) and intersecting hydrogen bonded group (IHBG) (where two hydrogen bonds share a common oxygen atom). SHBG involves transition states 2WTS1 and 2WTS2 which posses the lowest relative energy under 4.06209 KJ/MOL. PHBG includes transition states 2WTS4 and 2WTS5. The connection between PHBG and SHBG is through IHBG as represented by transition state 2WTS3.

On the Trimmer problem, a total of 10 transition states was discovered as tabulated in Table 6-8, where the globally low-lying transition state 3WTS1 has a relative energy of

7.163406 KJ/MOL. The Trimmer transition states are clustered using agglomerative hierarchical clustering to arrive at the dendrogram in Figure 6-11 where the Trimmer transition states are grouped together into: triangle or linear structures. Triangle transition states generally have lower relative energy than their linear counterparts. In particular, the average relative energy of the triangle transition structures is 0.006 KJ/MOL lower.

On the Tetramer problem, MOL-MA discovered 25 transition states tabulated in Table 6-9. These hierarchically clustered transition states are depicted in Figure 6-12 where they are organized structurally together into two main categories: ring and linear shape. The ring shape category includes transition states with closed ring, which can be further segregated in square and extended triangle ring shapes. The square ring group includes those transition states that resemble a square. The extended triangle ring group involves those transition states that extend triangle Trimmer with a water molecule. Transition states in the square ring group are most often the most energy favorable, to which the global Tetramer transition state belongs, while linear transition states generally possess higher relative energy.

Figure 6-10: Dendrogram for the hierarchical clustering of Dimer transition states.

Table 6-7: Dimer transition states.

ID	CLASS	TS	ΔE
2WTS1	Single Hydrogen Bonded Group (SHBG)		2.63E-06
2WTS2			4.06208
2WTS3	Intersecting Hydrogen Bonded Group (IHBG)		10.51329
2WTS4	Parallel Hydrogen Bonded Group (PHBG)		15.14526
2WTS5			16.5835

Figure 6-11: Dendrogram for the hierarchical clustering of Trimmer transition states.

Table 6-8: Trimmer transition states.

ID	CLASS	TS	ΔE	ID	CLASS	TS	ΔE
3WTS1	Triangle		7.163406	2WTS6	Triangle		32.92446
3WTS2	Triangle		11.90267	3WTS7	Triangle		37.55444
3WTS3	Linear		16.52455	3WTS8	Linear		39.36816
3WTS4	Linear		17.20039	3WTS9	Linear		40.11884
3WTS5	Linear		28.6958	3WTS10	Triangle		41.26749

Figure 6-12: Dendrogram for the hierarchical clustering of Tetramer transition states.

Table 6-9: Tetramer transition states.

ID	CS	TS	ΔE	ID	CS	TS	ΔE	ID	CS	TS	ΔE
4WTS1	C1		15.00662	4WTS10	C2		86.49286	4WTS19	C1		98.37601
4WTS2	C1		43.76194	4WTS11	C2		88.12224	4WTS20	C2		99.12469
4WTS3	C2		73.59111	4WTS12	C2		89.04555	4WTS21	C3		101.6754
4WTS4	C1		73.76608	4WTS13	C1		90.23651	4WTS22	C3		102.647
4WTS5	C2		73.95853	4WTS14	C2		90.87253	4WTS23	C1		107.3868
4WTS6	C1		78.01182	4WTS15	C2		91.18099	4WTS24	C2		108.1849
4WTS7	C2		82.91528	4WTS16	C2		93.52	4WTS25	C1		117.4129
4WTS8	C2		85.59123	4WTS17	C3		95.21847				
4WTS9	C2		86.03999	4WTS18	C2		97.13153				

6.4. Conclusion

Due to the importance of water transition states in science, this chapter, in the entirety, has been dedicated to the discovery of transition states in water clusters. The framework was extended with water-specially-designed operators, leading to the valley-adaptive clearing molecular memetic algorithm (MOL-MA) that requires no *a priori* knowledge on the initial guess, reactant and product or the global optima, for the discovery of the transition states in water clusters. The molecular memetic algorithm is composed of several core components, namely a specialized procedure for water cluster structure initialization, a fitness function, molecular-based water evolutionary operators, valley-adaptive clearing scheme and the Berny-based life time learning procedure. Assessments made against several state-of-the-art approaches in the field, in terms of percentage of uncovered transition states and computational cost incurred, demonstrated the excellent efficacy of MOL-MA over existing approaches.

Among all the counterpart algorithms considered, MOL-MA attained the largest percentage of uncovered transition states of highest precision, despite the lowest computational requirements. MOL-MA, moreover, does not require any *a priori* knowledge on the landscape such as initial guesses, reactant and products and global optima, which is always scarce in reality. In addition, the variations on the solution set uncovered by MOL-MA is always lower than the counterpart algorithms, highlighting the robustness or reliability of MOL-MA in the discovery of transition states.

MOL-MA has been employed on Dimer, Trimer and Tetramer water clusters using high fidelity computational models based on first principles calculations. The results obtained indicated that MOL-MA not only successfully discovered the previously known transition states, but also found newly established transition states.

These uncovered transition states were presented or visualized on dendrograms that are morphologically categorized using agglomerative clustering procedure and Ultrafast Shape Recognition structural similarity metric. The insights gained through the transition states uncovered and their morphological categorization could provide better understanding of the thermodynamics properties, hydrogen bond behavior as well as the isomerization processes in water clusters.

The success achieved through MOL-MA on water clusters encourages the authors for further investigate applicability for other molecular systems such as drug molecules, amino acids and protein folding. Although MOL-MA is applied to water clusters in this study, it can be easily adapted for subsequent studies in computational chemistry and biology. Moreover, the analyses conducted in this study can be replicated for other molecular systems to uncover landscape properties and provide insights to both physical chemists and evolutionary algorithmists.

Chapter 7: Evolutionary Discovery of Transition States in Ring-deficient Covalently-bonded Small Molecules

In the previous chapters 5 and 6, we discussed that water clusters are aggregation of water molecules, which are held together by easy-to-form and easy-to-break bonds called the hydrogen bonds. Unlike water clusters, covalently-bonded molecules are aggregation of atoms bonded together with a sequence of covalent bonds. This sequence of covalent bonds shall be maintained across the search. Therefore, this chapter describes an extended version of the general framework proposed in Section 4.4 and Figure 4-9 and extended in Section 5.5, named the Molecular Memetic Computing methodology (MMC), which maintains covalent bonds across the search and requires no *a priori* knowledge on the initial guess, reactant and product or the global optima, for the discovery of the transition states. The essential backbone of our framework is an evolutionary algorithm coupled with a Berny transition states local solver in the spirit of Lamarckian learning. In contrast to earlier works, we introduce a tree representation of the molecular system with embedded chemical bonds knowledge. In addition, the subtree concept in tree coincides with the concept of substructure in the molecular context. This allows development of effective, yet meaningful reproduction operators. Based on this representation, we introduced specially designed tree-based reproduction operators that pass on noteworthy chemical bonds in the evolutionary process. We also consider a transition state fitness function, proposed in Section 4.4, to guide the memetic search towards low-energy transition states. Moreover, valley-adaptive clearing and duplicate checking schemes are also incorporated to enhance diversity of the search and discovery of multiple unique low-energy transition states, respectively. Empirical results are then presented for a number

of small ring-deficient molecules based on the Hartree Fock method with STO-3G basis set. Comparisons to recent advancements in related state-of-the-art algorithms showed that the proposed methodology not only uncovered the biggest set of transition states, but also attained the achievements at significantly lower computational costs.

The chapter is organized as follows: Section 7.1 provides the amendments of the general framework proposed in Chapter 4, necessary for efficient identification of transition states in covalently-bonded molecular systems, while Section 7.2 gives the results obtained from our computational study. Section 7.3 introduces a web service developed based on the extended framework for scientific society to use. Section 7.4 applies the proposed algorithm on locating glutamate isomers and study its impact on the function and dysfunction of the glutamatergic system – a communication system in the nervous system which mediate most of the communication involved in higher level functions of learning and memory. The brief conclusion is then stated in Section 7.5.

7.1. Covalent-bond-driven operators

In this section, we extend the general framework discussed in Chapters 4 and 5 to efficiently deal with covalently-bonded molecular systems. The extended form of the framework is donated as chemical bond-embedded memetic computational methodology (MMC)[116]. In MMC, tree-based chromosome representation is developed based on knowledge about the molecule structure, as details in Section 7.1.1, where information regarding atoms and bonds of the molecule are extracted and maintained in a tree structure for the whole process of evolution. Then, a population of potential transition state structure is randomly generated using a specialized initialization procedure, described in Section 7.1.2. The population individuals are then assessed as candidate transition states considering the fitness function, which is

described in Section 4.4. Population individuals survive to the next generation according to the stochastic universal sampling procedure [117]. The surviving structures then undergo the specially designed tree-based evolutionary operators of crossover and mutation (as detailed in Section 7.1.3). To facilitate diversity in the search, evolved transition states then undergo the valley-adaptive clearing scheme (see Section 4.2) where population individuals belonging to the same valley are segregated from others into niche or valley groups. The elite individual per valley group is then refined using Berny algorithm and Lamarckian inheritance [111], as described previously in Section 5.5.1. Valley elites that satisfy as transition states are archived and all other members of the valley groups are relocated randomly to other regions of the energy landscape to facilitate search exploration towards the discovery of other potential low-energy transition states that may exist. In the search process, any newly uncovered transition states are archived and checked against possible duplicates, as detailed in Section 5.5.2. This process repeats for a maximum number of generations or until the convergence criteria are satisfied.

7.1.1 Chromosome Representation

Chromosome representation in Evolutionary algorithms (EA) and MAs plays a key role in the success of the search process. It is widely accepted for the chromosome to be represented as a string of bits or real numbers, leading to binary and real-coded EA, respectively. However, a string representation may not fully gratify the requirements of molecular systems due to the existence of various rules and constraints, among which, the most important, the bond constraints. Unlike the string representation which represents decision variables in the form of a concatenated string, tree representation encode decision variables in a form of tree-like interconnected

nodes where node is constituted of coupled decision variables while interconnection between nodes is used for maintaining chemical bonds. In such away, tree representation is autonomously capable of maintaining chemical bonds of the molecule during the evolutionary process, especially for ring-deficient molecules. Tree representation also allows easy access to nodes while being aware of their structure aspects within the molecule. Each non-leaf node in tree can be considered as a root node for its descendent nodes. Both a non-leaf node and its descendant nodes constitute a sub tree. Since the subtree concept in the tree implements the concept of the substructure, tree representation allows easy identification of and dealing with meaningful subunits or substructure in the molecular system. All these features allow the development of effective, yet meaningful reproduction operators in evolutionary computation while maintaining the chemical bonds of the molecules unchanged, and subsequently effective search methodology that explore only the region of interest in the search space.

Tree node shown in Figure 7-1 is proposed to represent atoms and their bonds with its parent nodes. In particular, it maintains information such atoms index or id in the molecule (AtomID), coordinates ($\mathbf{x}$) and element as well as the type and order of the bond that synergize it with its parent. An example chromosome tree representation exemplified on GABA is shown in Figure 7-2, using C_α or C_6 as a pivot atom or root node. It is also worth-emphasizing that Reference Set (Ω_R) (see Figure 7-2) which includes references to all non-leaf nodes is also created during the construction of the tree as a substructure identifier and an index to speed up the accessibility to deep tree nodes.

Figure 7-1: Tree node in UML class notation.

7.1.2 Knowledge Extraction and Embedding

To develop efficient and effective discovery of low-energy transition states, knowledge about the structure and chemical bond of molecules shall be exploited to enhance the search. The activities required to extract and embed this knowledge on tree structure are depicted Figure 7-3. The process starts by identifying the molecule subject of study, using NCBI, drug data banks or any other chemistry databases. From a molecular structure file obtained, the molecule constituting atoms are then identified in terms of their elements and Cartesian coordinates. Further efforts are invested on bond identification to limit the search on low-energy transition states. In particular, the algorithm shall identify all bonds within a given molecule. By definition, a bond physically exists if the distance between its participating atoms within the limit of bond lengths [118] $[(\alpha_{\mathbf{b}_T(\mathbf{y}-\mathbf{x})_{\min}}), (\alpha_{\mathbf{b}_T(\mathbf{y}-\mathbf{x})_{\max}})]$, as detailed in Table 7-1. Its type $(\mathbf{b}_T(\mathbf{y} - \mathbf{x}))$ is recognized by its participating atoms ($\mathbf{y}$ and $\mathbf{x}$) and its length by $\|\mathbf{y} - \mathbf{x}\|$ while its order depends on the participating atoms and the bond length. To encode a molecule into a representative tree, the construction process traverse the molecule from a given pivot atom Σ through finding a set of atoms ($\Omega_{\mathbf{A},\Sigma}$) and set of bonds ($\Omega_{\mathbf{B},\Sigma}$) bonded to Σ . A child node is created for each atom in $\Omega_{\mathbf{A},\Sigma}$, and updated accordingly. Bond features for the child nodes are also updated according to the bond

specifications in $(\Omega_{B,\Sigma})$ between Σ and current child node. The same procedure repeats for this node and continues until there are no bonds or atoms left.

Table 7-1: Bond lengths constraints [118].

Bond Type ($b_T(y - x)$)	Min length ($\alpha_{b_T(y-x)_{\min}}$) Å	Max length ($\alpha_{b_T(y-x)_{\max}}$) Å
O-H	0.95	1.15
C-H	0.95	1.15
N-H	1.06	1.12
N-C	1.47	2.10
O-C	1.43	2.15
C-C	1.20	1.54

Once a tree-representation based chromosome has been generated as detailed in Section 7.1.1, a random set of mutations (See Section 7.1.3 (B)) is then employed on the seed structure. The randomly mutated variants of the seed molecule structure are then considered as an initial population. An example initial population of 10 individuals exemplified on GABA is depicted in Figure 7-4.

7.1.3 Tree-Based Molecular Evolutionary Operators

In evolutionary algorithms, reproduction operators are generally meant for inducing random, yet effective change on the population. We thus propose tree-based reproduction operators as change-inducing methodologies that use the tree structure to induce meaningful changes on the molecule while maintaining tree structure and chemical bonds unchanged. The proposed tree-based reproduction operators are composed of a tree-based molecular crossover which allows chromosomes to exchange subtrees/substructures and tree-based mutation operators which induce subtle changes on tree nodes and substructures. These are briefly described in what follows.

Figure 7-2: Chromosome tree-structure exemplified on GABA, taking C₆ or alpha carbon as a pivot atom or root/t node.

Figure 7-3: Knowledge extraction and embedding initialization procedure.

Figure 7-4: An example initial population of GABA structures.

A. Tree-Based Molecular Crossover

Crossover operates by interchanging genetic information between individuals or chromosomes. The crossover process between two molecules is depicted in Figure 7-5. The aim is to exchange some of the molecular substructures between generated molecules, thus allowing successful substructures to replicate across populations and generations. Therefore, two molecular structures (**P** and **Q**) in the population are first selected and then aligned according to Kabsch procedure [119], as detailed in what follows:

Step 1. (**P** and **Q**) are translated such that both molecules shares the center of mass at the origin of the coordinate system as indicated in (Eq. 7.1:2) below,

$$\mathbf{P}_t = \mathbf{P} - \mathbf{p}_{\text{cm}} \quad (\text{Eq. 7.1})$$

$$\mathbf{Q}_t = \mathbf{Q} - \mathbf{q}_{\text{cm}} \quad (\text{Eq. 7.2})$$

where $\mathbf{p}_{\text{cm}}$ and $\mathbf{q}_{\text{cm}}$ are the centre of masses of **P** and **Q**, respectively.

Step 2. The covariance matrix **C** between (**P_t** and **Q_t**) is then calculated as

$$\mathbf{C} = \mathbf{P}_t^T \times \mathbf{Q}_t \quad (\text{Eq. 7.3})$$

Step 3. Singular Value Decomposition (SVD) is then employed on **C**, leading

$$\mathbf{C} = \mathbf{V} \times \mathbf{S} \times \mathbf{W}^T \quad (\text{Eq. 7.4})$$

Step 4. The optimal rotation matrix **R_{opt}** is then calculated as

$$\mathbf{R}_{\text{opt}} = \mathbf{W} \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} \times \mathbf{V}^T \quad (\text{Eq. 7.5})$$

where $d = \text{sign}(\det(\mathbf{C}))$ to ensure that the rotation matrix is correct as a right-handed coordinate system.

Step 5. Both molecules (**P_t** and **Q_t**) are then rotated using **R_{opt}**, leading to (**P'** and **Q'**), as shown below,

$$\mathbf{P}' = \mathbf{R}_{\text{opt}} \times \mathbf{P}_t \quad (\text{Eq. 7.6})$$

$$\mathbf{Q}' = \mathbf{R}_{\text{opt}} \times \mathbf{Q}_t \quad (\text{Eq. 7.7})$$

Once the two chromosomes ($\mathbf{P}$ and $\mathbf{Q}$) are aligned as in ($\mathbf{P}'$ and $\mathbf{Q}'$), a random atom or node is then chosen and identified from the ReferenceSet ($\Omega_{\mathbf{R}}$) in chromosomes ($\mathbf{P}'$ and $\mathbf{Q}'$). Its details as well as the subtree structure are thus exchanged between the parent structures.

B. Tree-Based Molecular and Atomic Mutations

Analogous to biological mutation, the mutation operators in EAs bring in subtle changes to a randomly chosen subset of chromosomes in the MA population. We propose the following three molecular and atomic mutation strategies for structural evolution.

ST-1. Sub-Tree/substructure Rotation (See Figure 7-6): a node $\mathbf{s}$ is randomly chosen from the ReferenceSet ($\Omega_{\mathbf{R}}$). Their substructure or subtree is rotated randomly around the axis of the chosen node ($\mathbf{s}$) and its parent ($p(\mathbf{s})$) as shown in (Eq. 7.8), below.

$$\forall \mathbf{x} \in \mathbf{a}(\mathbf{s}), \mathbf{x} = \mathbf{R}_{(\vartheta_1, \langle p(\mathbf{s}) - \mathbf{s} \rangle)} \times \mathbf{x} \quad (\text{Eq. 7.8})$$

where $\mathbf{a}(\mathbf{s})$ set of atoms falls on nodes precedents to $\mathbf{s}$, which can be easily identified from the tree structure, $\mathbf{x}$ is a vector containing a selected atom coordinates and $\mathbf{R}_{(\vartheta_1, \langle p(\mathbf{s}) - \mathbf{s} \rangle)}$ is the rotation matrix where rotation is performed around $\langle p(\mathbf{s}) - \mathbf{s} \rangle$ axis with a random angle ϑ_1 .

ST-2. Leaf Nodes/Atoms Rotation (See Figure 7-7): a set of nodes are randomly chosen from ReferenceSet ($\Omega_{\mathbf{R}}$). Their child leaf nodes ($\Omega_{\mathbf{L}}$) are then arbitrarily rotated. In particular, each atom with coordinates ($\mathbf{x}$) in $\Omega_{\mathbf{L}}$ is then rotated around the bond between the grandparent $p(p(\mathbf{x}))$ and parent $p(\mathbf{x})$ of the node as follows:

$$\forall \mathbf{x} \in \Omega_L, \mathbf{x} = \mathbf{R}_{(\vartheta_2, \langle p(p(\mathbf{x})) - p(\mathbf{x}) \rangle)} \times \mathbf{x} \quad (\text{Eq. 7.9})$$

where $\mathbf{R}_{(\vartheta_2, \langle p(p(\mathbf{x})) - p(\mathbf{x}) \rangle)}$ is the rotation matrix where rotation is performed around $\langle p(p(\mathbf{x})) - p(\mathbf{x}) \rangle$ axis with a random angle ϑ_2 .

ST-3. Nodes/Atoms Translation (See Figure 7-8): a set of randomly chosen atoms/nodes (Ω_T) are arbitrarily translated as in (Eq. 7.10) such that their bonds with neighbors are maintained within length constraints detailed in Table 7-1.

$$\forall \mathbf{x} \in \Omega_T, \mathbf{x} = \alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))} (p(\mathbf{x}) - \mathbf{x})$$

$$s. t. \alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))} \in \left[\alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))_{min}}, \alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))_{max}} \right] \quad (\text{Eq. 7.10})$$

where $p(\mathbf{x})$ is parent atom of $\mathbf{x}$ and $\alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))}$ is the bond between the atom $\mathbf{x}$ and its parent atom where $E(\mathbf{x})$ and $E(p(\mathbf{x}))$ are elements of the atom $\mathbf{x}$ and its parent, respectively. $\alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))}$ is a random number between the minimum ($\alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))_{min}}$) and maximum ($\alpha_{b_T(E(p(\mathbf{x})), E(\mathbf{x}))_{max}}$) bond lengths, as tabulated in Table 7-1.

Figure 7-5: Tree-based crossover operator, exemplified on GABA.

Figure 7-6: Rotation of randomly chosen substructure, exemplified on GABA.

Figure 7-7: Rotation of randomly chosen leaf nodes, exemplified on GABA.

Figure 7-8: Translation of a random set of atoms, exemplified on GABA.

7.2. *Experimental Study*

In this section, we study the efficacy of the proposed memetic methodology (MMC) for locating transition states of several ring-deficient covalently-bonded molecules and pit it against several existing state-of-the-art optimization methods, including the stochastic multi-start local search Berny Algorithm (SMLS) [82] as representative of single-ended methods, GRRM [114] as representative of doubly-ended methods, while Chaudhury GA (ChGA) [21] and Bungay GA (BGA) [19] as representatives of the genetic algorithm. The sequential niching memetic algorithm (SNMA) [83], as representatives of recent advances in niching and memetic algorithms, are also considered for comparison in the present study.

All methods work as baselines for comparison on the transition states optimization problem on a set of ring-deficient molecules detailed in section 5.6.1 (B), using number of transition states uncovered and the number of gradient evaluation as performance measures. The energy, gradient and eigenvalues are calculated using first-principles approach at the Hartree–Fock level, with the STO-3G basis set, and all computations are executed on the PDCC cluster⁵.

We present on the numerical results of (MMC) pitted against other state-of-the-art evolutionary and memetic approaches for the discovery of transition states. For the sake of brevity, the unique properties of the algorithms considered in the present study, i.e., in terms of life time learning, specialized molecular evolutionary operators, niching operators and domain knowledge required are summarized in Table 7-2.

⁵ PDCC is 16 nodes cluster at the parallel and distributed computing center at NTU

Table 7-2: Properties of the algorithms considered (Y denotes the existence of a property while n refers to its absence).

Properties	MMC	SNMA*	SMLS	GRRM	BGA	ChGA
Life-time learning	Y	Y	Y	N	N	N
Tree-Based Molecular Operators	Y	Y	N	N	N	N
Niching Operator	Y	Y	N	N	N	N
Domain Knowledge Required	N	N	Initial guesses	Reactant and Product	Initial guesses	Global Optima

The performance efficacy of the algorithms assessed according to number of transition states uncovered depicted in Table 7-3. According to the table, it is observed that only algorithms that consider local search operations have attained at least a transition states. Among those successfully attained a transition state, algorithms with tree-based molecular operators has observed to attain the largest set of transition states uncovered. Among all, MMC has been observed to attain the largest number of transition, with significant improvement over other algorithms ranging from 120-802% improvement rate.

The computational requirements of the algorithms considered in the present study are also summarized in Table 7-4. From the results, SMLS is observed to incur the most computational effort among algorithms configured with local searcher or life-time learning procedure. In spite of the largest percentage of transition states uncovered as discussed previously, MMC remains to display superiority in terms of computational efficiency over the other state-of-the-art algorithms.

Table 7-3: Number of transition states uncovered (improvement rate in percentage) [Note that the best is highlighted in bold].

Molecule Algorithm	GABA	GABOB	LEUCINE	TROMTHAMINE	VALINE	Glutamic Acid
MMC	256	640	305	519	159	490
SNMA*	198 (129%)	449(142%)	136 (224%)	382 (135%)	123(129%)	339(144%)
SMLS	37 (691%)	81(790%)	38 (802%)	69 (752%)	25(636%)	71(690%)
GRRM	89 (287%)	89 (719%)	102 (299%)	99 (524%)	75 (212%)	103(475%)
BGA	0	0	0	0	0	0
ChGA	0	0	0	0	0	0

Table 7-4: Number of gradient evaluations (the best is highlighted in bold).

Molecule Parameters	GABA	GABOB	LEUCINE	TROMTHAMINE	VALINE	Glutamic Acid
MMC	321,307	430,621	675,321	141,300	259,139	527,745
SNMA*	421,560	510,256	725,621	479,456	545,945	745,456
SMLS	954,561	980,451	985,126	987,565	951,458	978,879
GRRM	478,568	437,561	696,890	476,451	426,569	625,345
BGA	10 ⁶	10 ⁶	10 ⁶	10 ⁶	10 ⁶	10 ⁶
ChGA	10 ⁶	10 ⁶	10 ⁶	10 ⁶	10 ⁶	10 ⁶

With the significantly larger number of uncovered transition states, the use of evolutionary MMC has furthermore allowed knowledge discovery from data generated over the course of the evolution rather than requiring prior knowledge as with the case of the other methods. Table 7-5 tabulates the characteristics of the uncovered transition states in terms of number of oxygen, carbon, and hydrogen atoms as well as number of the total atoms and the rotatable bonds. Observation suggests oxygen atoms play a key role in the flexibility of the molecule, allowing larger number of transition states to be uncovered. Carbon atoms, in contrast, seem to limit the flexibility of the molecule, decreasing the number of transition states to be uncovered. Meanwhile, hydrogen atoms are observed not to play important role in the number of transition states uncovered. Additionally, it is observed that the number of rotatable bonds play an obvious role in the number of transition states uncovered while the total number of atoms has not shown any significant impact. The large number of transition states of GABOB, thromethamine and glutamic acid indicates the flexibility of the molecules, allowing them to interact with broader range of molecular systems. GABOB as an anti-epileptic drug, for example, is a more effective neuro-inhibitor than GABA from which it is derived. The flexibility that GABOB can adopt allows it to interact more efficiently with larger set of GABA-ergic system components in the central nervous systems.

Table 7-5: Characteristics of uncovered transition states in terms of number of atoms, number of rotatable bonds, and fitness-distance correlation

	GABA	GABOB	Leucine	Thrometamine	Valine	Glutamic Acid
#transition_state	256	643	305	519	159	490
#oxygen	2	3	2	3	2	4
#carbon	4	4	6	4	5	5
#hydrogen	9	9	13	11	11	9
#total_atoms	16	17	22	19	19	19
#rotatable_bonds	3	3	3	3	2	3
FDC	26.57	3.37	54.48	31.83	92.20	5.2

In Table 7-5, characteristics of the energy landscape of the molecules are also tabulated as Fitness-Distance Correlation (FDC) [101] to measure the difficulty of the landscape. FDC is essentially the Pearson product moment correlation between the energy differences and the structural differences of the solutions to the lowest-energy transition states as shown below

$$\text{FDC} = \frac{\text{cov}(\partial f; \partial \theta)}{\sigma(\partial f)\sigma(\partial \theta)}$$

where $\text{cov}(\cdot; \cdot)$ and $\sigma(\cdot)$ are the covariance and the standard deviation, respectively, while ∂f and $\partial \theta$ represent the energy difference and the USR structure dissimilarity between each transition state and the lowest-energy transition state, respectively. The FDC categorizes energy landscapes into well-ordered, rough, or deceptive. High correlation with FDC value above 60 indicates well-ordered landscape such that optimization methods can locate the transition states quite easily whereas low correlation indicates rough landscape in which optimization techniques could be misled to sub-optimal solutions. Negative correlation, meanwhile, indicates a deceptive landscape where the lowest-energy transition state is located among the high energy transition states. From Table 7-5, the ring-deficient molecules we investigate, except Valine, are found to possess rough landscapes. Of remarkable difficulty is the landscape of GABOB with FDC value

of 3.37. The MMC has however successfully uncovered around six to eight times more of its transition states than those found by the other methods, owing the success to the tree-based representation and the covalent-bond-driven evolutionary operators. This further showcases the strengths of the evolutionary memetic computing paradigm when coupled with the appropriate representation and operators such that little or no effort would be spent exploring the region where no solution may be located.

With the large number of transition states uncovered, we further our analysis to include the distribution of uncovered transition states by the MMC across the relative energy and the structure dissimilarity, making reference to the lowest-energy transition states, as shown in Figure 7-9. Distribution of the uncovered transition states of the molecules under investigation are depicted in separate subgraphs using the scatter plots. Of remarkable achievement is the broad range of transition states uncovered by the MMC as can be concluded from observing the x-axis of the scatter plots. In other words, the MMC has been capable of discovering unique transition states with high degree of structural dissimilarities. Observing the y-axis, on the other hand, it is noted that the uncovered transition states are distinguishable into low- and high-energy transition states. The presence of high-energy transition states in the region of low dissimilarities could easily hinder the discovery of low-energy transition states in the region of high dissimilarities. Fortunately, this is not the case with the MMC. The valley-adaptive clearing scheme used to maintain the population diversity has indeed been an effective measure for the successful and efficacious exploration of the energy landscape. Together with the tree-based representation and the covalent-bond-driven evolutionary operators, the MMC has incurred the least amount of computational costs in addition to uncovering

the largest set of transition states. This shows the evolutionary memetic computing paradigm is indeed a powerful and efficient tool for solving complex problems when equipped with appropriate representation and operators.

Figure 7-9: Distribution of uncovered transition states across relative energy and structure dissimilarity with reference to the lowest-energy transition state

7.3. Transition State Web Server

As highlighted in Chapter 1 and 2, the determination of the transition state ensemble (TSE) of small molecules can improve the understanding of the physicochemical properties of protein folding. To identify the transition state ensemble (TSE) of small molecules, many optimization approaches have been proposed in the last decade. Despite these efforts, the existing methods are plagued with several pitfalls such as slow

convergence and the low quality as well as the number of transition states uncovered. Moreover, the freely available transition states generator available for use by scientists remains limited to date. Further, programming knowledge on how to compile, install or even modify these freely available codes are often part of the prerequisites required to use them correctly. This is often tough on scientists that lack such specialized skills.

To identify the transition state ensemble (TSE), in this study, we design a freeware web-service: TranSGen: <http://c2istig.sce.ntu.edu.sg/ts/TranSGen.html> - a web service generator of transition states in small molecule [120], see Figure 7-10. Formulated with the memetic search methodology, proposed in Section 7.1, the structural configuration space is explored for the discovery of transition states using first-principle calculations. When first invoked, the web service develops structural awareness about the queried molecule. Molecules are subsequently represented in the tree, forming chromosomes. They are then evolved according to the basic memetic framework, while embedding knowledge on all molecular bonds within the search initialization process, evolutionary, diversity maintaining and local search operators. All unique transition states uncovered during the search are archived and finally sent to the user for their perusal by email.

To overcome the absence of freely available, convenient-to-use tools for transition state discovery, we fill the gap by designing a web service which uses the recent emerging advances of computational intelligence, particularly, the memetic algorithm (MA). This web service is presented to the scientific community as a service that generates a database of transition states from the small molecule of interests using first-principle calculations. Discovered transition states shall then be sent to the user via the email provided. The web

service has been employed on several small molecules, discovering dozens of not only previously uncovered but also newly found transition states.

TransGen: Transition states Generator for Small Non-Cyclic Molecules

TransGen

TransGen is the first cloud [services](#) that allow scientists, drug designers, chemists and bioinformaticians to generate transition states of small non-cycle molecules online, using first-principle calculations.

This cloud service uses the memetic computation paradigm. In this paradigm, a population of individuals with a specific "Population Size" explores the [search](#) space across a "Number of Generations". The individuals are evaluated as potential transition states using First-Principle methods and basis sets.

To submit a job to TransGen, the user needs to first identify and upload the structural file of the molecule to the cloud. "XYZ", "PDB", "SDF", "MOL", and "CML" are the current file extensions that the cloud service supports. Users are then required to fill-in the "Transition state Generation Submission Form" by including a user name and a valid [email](#) address to be later on informed about the status of their job. Job details such as method and basis set, number of generations and population size are set by default. However, users can easily change these by directly choosing an option from the down lists.

When the job is successfully completed, an email will be sent to the user with a database of transition states generated and a link to visualize them online. The Job ID, generated after submission, can be easily entered in the "Job Status Inquiry" field to check the job status online.

For trial, you may attempt with:

1. [Glutamic acid](#)
2. [4-amino-3-hydroxybutyric acid](#)
3. [gamma-Aminobutyric Acid](#)
4. [Arginine](#)
5. [Valine](#)
6. [Iedonal](#)

Users are most welcome to contact us our e-mail: mostafa.mhashim@gmail.com for help or suggestions.

[FAQ](#)

Job Status Inquiry

Job Status Enquiry

Job ID

Transition state Generation Submission Form

Your Information Details

User Name

Email

Job Details

Method and Basis Set

Number of Generations

Population size

Notifications

Upload Molecule File

Upload File ***

*** This cloud service support the following file formats or extensions: "XYZ", "PDB", "SDF", "MOL", and "CML"

Figure 7-10: TransGen webserver main page.

7.4. Toward Better Understanding of Glutamatergic Systems: Glutamate Stereoisomers Discovery and Receptor Binding Affinity

In the brain, nerve cells (neurons) communicate via a chemical messenger glutamic acid which mediates most of the brain activity and functions such as cognition, learning and memory. Abnormalities in glutamate-dependent neural activity are also the cause for many neurological diseases, including Alzheimer's disease and stroke/brain trauma-induced nerve cell death. Glutamic acid has traditionally been considered as a simple, uniform molecule, existing at all brain regions and at all times as the same structure and function. We intend to establish that Glutamic acid can exist in multiple different structure

forms and thus with different levels of activity. This will not only provide new insights to our understanding of how brain functions, but may also open new doors for the design of novel strategies in the prevention and therapeutical management of neurological abnormalities such as epilepsy, neurodegenerative diseases and neuropsychological disorders.

In this study, the proposed methodology presented in Section 7.1 has also been employed for the discovery of glutamic acid stereoisomers on the potential energy landscape, using energy as a fitness function and DIIS as a local search algorithm [121]. The energies as well as first and second derivatives of glutamic acid structure, used in this algorithm during the evolution process, are calculated using the ab-initio approach. Glutamic acid stereoisomers generated are then transformed to *PDBQT* format in their preparation for protein docking using Vina where the backbone hydrogen atoms have been deleted and partial charges have been added. Against, the ionotropic glutamate receptor 2 (GluR2) ligand-binding domain (Figure 7-11), the uncovered stereoisomers are docked according to the rigid-ligand and rigid-receptor docking approach to show the instantaneous effect of glutamate stereoisomers while interacting with GluR2.

Our computational studies have led to the discovery of 524 glutamate stereoisomers. Receptor binding analysis (Figure 7-12) has revealed that the binding of these glutamate stereoisomers varies greatly between -3.5 and -7.5 KCAL/MOL, indicating significant difference in the activation of the glutamate receptors. These findings encourage us to explore further the environmental conditions that lead to those stereoisomers that can over-activate the glutamate receptors. Similar studies can be replicated for the interaction

of the glutamate stereoisomers with other glutamate receptors and transporters. Knowing the environmental conditions and stereoisomers that can over-activate glutamate receptors and/or inhibit glutamate transporters can assist scientists in the design of novel strategies in the prevention and therapeutical management of neurotoxicity - the main component of several fatal neurodegenerative diseases, including Alzheimer's disease.

Figure 7-11: Examples of glutamate stereoisomers identified. AMPA receptor binding domain (right) is enlarged (right bottom) to show the structural details of the receptor and the glutamate molecule (red). Distinct properties in AMPAR interaction with varied glutamate stereoisomers is the focus of our study.

(A)

(B)

Figure 7-12: (A) the binding energy distribution of glutamate stereoisomers against iGluR2. (B) statistical summary on the binding energy of glutamate stereoisomers.

7.5. *Conclusion*

With the importance of small molecules in biology, chemistry, and medicine, this chapter has been utterly dedicated to the discovery of the transition states of small ring-deficient molecules. The multi-modal memetic framework proposed in Chapter 4 has been extended with the tree-based representation of ring-deficient molecules and the covalent-bond-driven evolutionary operators. The extend form of the multimodal memetic framework shapes a novel Molecular Memetic Computing (MMC) methodology that requires no prior knowledge for the discovery of transition states. In the spirit of Lamarckian learning, the specialized representation and operators are used to complement the memetic computing technique comprising a genetic algorithm for the population-based global search, the Berny algorithm for the life-time individual learning, and the valley-adaptive clearing scheme for maintaining the population diversity.

The evolutionary MMC methodology shows promising application of memetic computing techniques in the field of molecular optimization by demonstrating excellent efficacy against several state-of-the-art approaches. Not only did the MMC uncover the largest number of transition states, but it also incurred the least amount of computational costs. The use of evolutionary MMC, furthermore, allows knowledge discovery from data generated over the course of the evolution rather than requiring prior knowledge as with the case of many earlier methods. It has been found out that oxygen atoms play a key role in the flexibility of the molecule, allowing larger number of transition states to be uncovered. In contrast, carbon atoms have been observed to limit the flexibility of the molecule, decreasing the number of transition states to be uncovered. Meanwhile, hydrogen atoms have not been observed to play any important role in the number of

transition states uncovered. Additionally, it has also been found out—as one would expected—that the number of rotatable bonds play an obvious role in the number of transition states uncovered while the total number of atoms has not shown any significant impact. On a side note, large number of the transition states of GABOB, Tromethamine and glutamic acid suggests the flexibility of the molecules, allowing them to interact with broader range of molecular systems. GABOB as an anti-epileptic drug, for instance, is a more effective neuro-inhibitor than GABA, from which it is derived. The flexibility that GABOB can adopt allows it to interact more efficiently with larger set of GABA-ergic system components in the central nervous systems.

Based on MMC, we have proposed a web server for finding transition states of small molecules, providing a freeware and easy-to-use service for scientific community. Also we had employed the algorithm for finding stereoisomers of glutamic acid, finding more than 524 stereoisomers which have been examined against glutamate receptor 2, showing varying modulation effect on the glutamate receptors. This not only will help scientists to better understand the glutamatergic system, the major communication systems in the central nervous systems, but also may help scientists in developing therapeutical management tools for many neurodegenerative disorders including Alzheimer's disease.

With the generic tree-based representation, the MMC can be easily adapted for subsequent studies of other small ring-deficient molecules that are important in biology, chemistry, and medicine. To cover molecules that involve ring structures, an extension to the tree-based representation shall be studied. Extensions to the current covalent-bond-driven evolutionary operators shall also be devised. To allow working with larger

molecules, techniques to intelligently restrict the execution of individual learning procedure shall be explored.

Chapter 8: Concluding Remarks and Future Research

Investigations

8.1. Conclusion

Saddle points represent important landmarks in the problem landscape, which are very important for several applications in various fields of science and engineering. Estimation of chemical reaction rate, image segmentation, path-planning and robotics navigation represent interesting applications. Identifying such points using evolutionary and memetic approaches is a field that has remained yet to be well investigated. In this dissertation, the interest is thus to embark on a novel research study to locate multiple saddle points accurately and efficiently using evolutionary approaches. In this dissertation, a detailed review on finding first-order saddle points has been provided. In the review, methods are categorized as either conventional or stochastic. Conventional method has the advantage of converging quickly and precisely provided that the search starts from near vicinity of the saddle points. Stochastic methods, on the other hand, are capable of converging well even if the search were to start far from the saddle points. A motivating example for us is Evolutionary Algorithms (EAs). Nevertheless, it is worth noting that existing EAs do not serve our problem of interest sufficiently well. In particular, existing EAs experience significant difficulties in maintaining multiple saddle points during the search. Furthermore, EAs are often deemed to converge too slowly, leading to poor search efficiency.

In this dissertation, we proposed a multi-modal memetic framework, composing of 1) a fitness function that transforms the problem landscape to favor saddle points instead of

local or global optima, 2) evolutionary operators of crossover and mutation for efficient exploration of the problem landscape, 3) valley-adaptive clearing scheme for maintaining multiple saddle points during the evolutionary process, and 4) memetic operator consisting of saddle point local searcher for attaining precise saddle points. Experimental study on synthetic benchmark functions of varying complexity demonstrates the efficacy of the proposed framework for locating multiple first-order saddle points of higher precision, when compared to its evolutionary compeers.

From synthetic benchmark problems to real-world problems, the multi-modal memetic framework is applied to systems representing the two main streams of the molecular systems, namely, water clusters and ring-deficient covalently-bonded small molecules. In particular, the framework has been extended to efficiently explore the landscape of biomolecular systems for transition states, especially when using first-principle calculations which describe mathematically the interaction among atoms and their subatomic structures. For water cluster, the proposed framework is extended with water-specially-designed reproduction operators. The application of the proposed framework on water cluster has shown not only to uncover those previously reported but also to establish newly discovered transition states of size 2-4 water molecule, with the minimal computational cost among to its compeers.

In ring-deficient molecules, the multi-modal memetic framework is extended with tree-based evolutionary operators, to induce configurational changes on the molecules while maintaining both the tree-structure and the molecular properties unchanged. Comparisons to recent advancements in related state-of-the-art algorithms on several

small, yet very important, molecules showed that the multi-modal memetic framework not only uncovered the largest set of transition states, but also attained the achievements at significantly lower computational costs. With further analysis on the transition states uncovered by the framework, it has been found out that the number of oxygen atoms and rotatable bonds play a key role in the flexibility of the ring-deficient covalently-bonded small molecule, allowing larger number of transition states to be uncovered while carbon atoms have been observed to limit the flexibility of the molecule, decreasing the number of transition states to be uncovered. The number of hydrogen atoms as well as total number of atoms, in contrast, has not shown any significant impact.

The framework, extended for small ring-deficient molecules, has been, equally important, considered for locating glutamic acid stereoisomers, which later uncovered to have a modulation effect on the glutamatergic systems. Such findings open new door for neuroscientist to study the impact of the glutamic acid stereoisomers on the function and dysfunction of the glutamatergic systems. Moreover, several web-services have been developed based on these algorithms to help scientist, scholars and bioinformaticians in their research requiring both stable and transition states using accurate approaches of first-principle calculations.

From this study, we finally conclude that the memetic computing paradigm is the most efficient evolutionary paradigm ever applied for saddle point identification, and incorporation of domain knowledge into the evolutionary operators improves search performance significantly, constraining search process to explore only the feasible region of the landscape.

8.2. *Potential Future Research Investigations*

In this subsection, some of potential future research directions are briefly described. Here, the research directions are categorized as either focusing on the context of algorithmic issues or the applications. The first form of investigation targets for more effective, efficient and robust methodologies for multi-modal saddle point evolutionary search. Applying the proposed algorithms to locate transition and meta-stable states of complex systems and processes such as water cluster, β -amyloid self-assembly and proto-cell formation is the second core research interest. The details of each are presented in the rest of the section.

8.2.1. Algorithmic Issues

Future research investigations involve the amendment that can improve the current algorithmic framework. In particular, the algorithmic issues would include the incorporations of machine learning or graph mining technologies for identifying the potential solution individuals that shall proceed with individual learning. This stage is highlighted by C in Figure 8-1. Another immediate potential algorithmic improvement that is worth considering lies within the individual learning procedure. Instead of considering a single individual learning procedure, meta-Lamarckian learning [111] or the modeling of meme complexes [122], can be considered. This stage is noted as “symbioses individual learning stage” as denoted by B in Figure 8-1. Last but not the least, shown in Figure 8-1 (C, D), it would be interesting to consider the incorporations of graph indexing and management tool for efficient organization of the archival of saddle points or transition states uncovered throughout the search process.

A. Machine Learning

Classification, the machine learning approach, can improve the search performance of the proposed methodologies. In particular, classification could cut down the number of the first-principle calculations in the local search operations, which is a computationally-demanding bottleneck. Applying local search operations selectively on configurations that would most likely converge to solutions would highly enhance search performance. In this direction, the future framework may incorporate a classifier. This classifier would be trained on both the configurations that undergo the local search operations and successfully converge to solutions and those otherwise. Subsequently, the classifier would recommend configurations resemble the characteristics or the features of the first to undergo local search operations. In such a way, incorporation of such a classifier would eventually improve search efficacy considerably.

B. Symbiosis of Local Searchers

Finding transition states is very difficult. Some local searchers, despite we have not encounter such a case, can be effective in locating some saddle points than others on the same or different problem landscape. Considering symbiosis of local searchers represents an additional potential area of interest. Deciding when and where one local searcher, rather than the other, should be applied may also be incorporated in the machine learning approach, mentioned previously in A. This design can mend the search performance of the proposed framework.

Figure 8-1: Algorithmic-issues depicted in the framework: A) shows where the machine learning stage should be incorporated; B) shows where the symbioses life-time learning procedure should be considered; C and D) show where graph mining and indexing should be incorporated.

C. Graph-Based Indexing and Management

Archive is an essential part of the proposed algorithm. In particular, it is a very important tool to enhance search toward unexplored region in the landscape. Since molecular structures can be represents in a graph form, considering graph indexing and management would represent an interesting direction for future research. Recent advancement in graph databases and management would help in providing efficient way to archive and index molecular structures. This would eventually have impact on the search performance.

D. Graph Mining

Graph mining is one of the most active areas in computer science. Several methods have suggested in the literature, in the last decade, which can boast the proposed algorithm. In particular, graph mining can assist in mining graph building blocks to work as memes in optimization of the transition states in complex biomolecular systems.

E. Multi-Scale Incorporation

The usage of the first-principle calculations represents one of the main pitfalls of the current algorithms. The nature of transition states, however, imposes such a requirement. In this part, we hypothesis that optimization of transition states in complex systems can be considered first using approximated potential energy which shall be less computationally demanding. Once a good estimate of the transition state has been uncovered, a more accurate landscape based on the first principle calculations can then be used according to Hybrid MM/QM modeling. In this way, the algorithm can first explore the inexpensive energy landscape to identify the highly potential first-order saddle points. First order saddle points uncovered during this stage can then be used as initial conjecture or useful biases to the same algorithm when considering higher scale approaches (i.e. more accurate landscapes based on first-principle calculations), thus enhancing subsequent searches.

Figure 8-2: The multi-scale framework for the current algorithmic framework proposed.

To find first-order saddle points at higher resolutions, several strategies can be employed. The first and most simple, is to apply local search strategy on the solution uncovered. The challenge with this approach is that the current derived biases might not be good starting points for the local search to embark. The second strategy may consider applying the same evolutionary paradigm proposed in the thesis, as shown in Figure 8-2. For every uncovered solution, a number of points would be sampled within a certain region of the uncovered solution. This population of points would undergo the same evolutionary process until an improved solution is uncovered or the maximum number of

gradient evaluations has been incurred. Since our concern in the next stage is to find a single solution, valley-adaptive clearing may be ignored at this stage.

8.2.2. Domains of Application

In this subsection, some challenging-future applications of the proposed algorithms are described.

A. Optimization of β -Amyloid Isomers and Transition State Structures

β -amyloid is a peptide with 43 amino acids, which is the building block of the amyloid plaques in the brains of the people who are suffering from Alzheimer's disease. Due to β -amyloid role in the pathogenic process of Alzheimer's disease, it has become one of the most actively studied molecules in Alzheimer's disease research projects. Understand the mechanism of amyloid formation requires identifying different isomers (local optima with same sequences of amino acids) as well as the transition states. In this study, all possible isomers (including those lying at the global minima) as well as the transition state structures of β -Amyloid peptides is to be located. Such findings will help scientist understand the dynamics of β -Amyloid peptides assemblies.

B. Convex Region Identification

Convex region identification presents an importance component of machine learning. Although the link between the convex region and saddle points have been mathematically proven in [35, 123], the use of saddle point in identifying the convex region remains yet to be well-investigated. Saddle points, especially first-order saddle points, can pose as border check point for convex regions. In this study, the aim is to use the identified saddle points with the aid of curve fitting approaches for convex region identifications.

C. Minimum Energy Pathway Approach for Designing Artificial Oxygen Carrier Proto-Cells as Potential Replacement for Red Blood Cells

Blood diseases such as anemia, Sickle-cell, thalassemia, spherocytosis and leukemia are becoming common in human today. The common feature among these diseases is that they negatively affect the blood's efficiency of carrying oxygen. In this study, the interest lies in using the approaches to identify the minimum energy pathway for transition of inanimate to animate oxygen-carrier proto-cells as follow:

Material

1. Hemoglobin (~ 2700 atom)
2. For Proto-Cell Material, the material given in [124,125] will be considered.

Approach:

1. Locating stable-states of both Hemoglobin formation and proto-cell assembling.
2. Locating transition states of both Hemoglobin and the proto-cell
3. Using Information available in 1 and 2, identify the minimum energy pathway of the proto-cell assembly from disorder state to vascular state around Hemoglobin.

Author's Publications

- [1] M. Ellabaan, Y.S. Ong, S.D. Handoko, C.K. Kwoh and H. Man "Discovering Unique, Low-energy Transition States of Small, Non-cyclic Molecules Using Evolutionary Molecular Memetic Computing", accepted in *the Special issue of IEEE Computational Intelligence Magazine on "Complex Engineering Optimization"*.
- [2] M. Ellabaan, Y. S. Ong, Q. C. Nguyen and J.-L. Kuo, "Evolutionary Discovery of Transition States in Water Clusters", *Journal of Theoretical and Computational Chemistry*, Vol. 11, No. 05, Oct 2012.
- [3] M. Ellabaan, S.D. Handoko, Y. S. Ong, C.K. Kwoh, S. Bahnassy, F. Ellassawy, and H.Y. Man, "A Tree-structured Covalent-bond-driven Molecular Memetic Algorithm for Optimization of Ring-deficient Molecules", *Computers and Mathematics with Applications*, Vol. 64, No. 12, 2012.
- [4] M. Ellabaan, Y.S.Ong, Valley-Adaptive Clearing Scheme for Multimodal Optimization Evolutionary Search, in: the 9th International Conference on Intelligent Systems Design and Applications (ISDA'09), IEEE Computer Society, Pisa, Italy, 2009, pp. 1-6.
- [5] M. Ellabaan, Y.S. Ong, M.H. Lim, J.-L. Kuo, Finding Multiple First Order Saddle Points Using a Valley Adaptive Clearing Genetic Algorithm in: IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA'09), Daejeon, Korea 2009, pp. 457-462.
- [6] M. Ellabaan, Y.S. Ong, S.D. Handoko, C.K. Kwoh, S.A. Bahnassy, F.M. Ellassawy, Toward Better Understanding of Glutamatergic Systems: Glutamate isoforms discovery and Receptor Binding Affinity, in: ISMB 2012 (Poster). Long Beach, California, USA, 2012.
- [7] M. Ellabaan, Y.S. Ong, S.D. Handoko, C.K. Kwoh, S.A. Bahnassy, F.M. Ellassawy, TranSGene, 1.0: A Web Service For Transition States Discovery in Small Ring-exclusive Molecules Using First-Principles Calculation, in: ISMB 2012 (Poster). Long Beach, California, USA, 2012.
- [8] M. M. H. Ellabaan and Y. S. Ong, 'Experiences on Memetic Computation for Locating Transition States in Biochemical Applications', ACM Genetic and Evolutionary Computations (GECCO 2012), July 2012.
- [9] M. Ellabaan, et. al. "Multi-Modal Valley-adaptive Clearing Memetic Algorithm for Efficient Discovery of First-order Saddle Points", The 9th International Conference on Simulated Evolution and Learning (SEAL'2012)

References

- [1] M. Lassonde, *Approximation, optimization and mathematical economics*, Springer, 2001.
- [2] G. Keserü and I. Kolossváry, *Molecular mechanics and conformational analysis in drug design*, Blackwell Science Inc, 1999.
- [3] D. Jones and B. Sleeman, *Differential equations and mathematical biology*, CRC Press, 2003.
- [4] C. Sminchisescu and B. Triggs, "Building roadmaps of minima and transitions in visual models," *International Journal of Computer Vision*, vol. 61, pp. 81-101, 2005.
- [5] K. Lim, Y. Ong, M. Lim, X. Chen, and A. Agarwal, "Hybrid ant colony algorithms for path planning in sparse graphs," *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, vol. 12, pp. 981-994, 2008.
- [6] M. Toda, "Transition State Theory Revisited," in *Advances in chemical physics* S. A. R. Ilya Prigogine, Wiley-Interscience, 2002.
- [7] D. Mackay, P. Berens, K. Wilson, and A. Hagler, "Structure and dynamics of ion-transport through gramicidin-a," *Biophysical Journal*, vol. 46, pp. 229-248, 1984.
- [8] K. Kim, H. Nguyen, P. Swaminathan, and E. Clementi, "NA⁺ AND K⁺ ION-transport through a solvated gramicidin transmembrane channel - molecular dynamics studies using parallel processors," *Journal of Physical Chemistry*, vol. 89, pp. 2870-2876, 1985.
- [9] M. Toller, G. Jacucci, G. Delorenzi, and C. Flynn, "Theory of classical diffusion jumps in solids," *Physical Review B*, vol. 32, pp. 2082-2095, 1985.
- [10] D. Oxtoby, "Nucleation of first-order phase transitions," *Accounts of Chemical Research*, vol. 31, pp. 91-97, 1998.
- [11] C. Smith and E. Wood, "Protein denaturation " in *Biological Molecules (Molecules and Cell Biochemistry)*, 1991.
- [12] M. Ellabaan, Y. Ong, M. Lim, and J. Kuo, "Finding multiple first order saddle points using a valley adaptive clearing genetic algorithm " presented at the IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA'09), Daejeon, Korea 2009.
- [13] R. Olsen, G. Kroes, G. Henkelman, A. Arnaldsson, and H. Jonsson, "Comparison of methods for finding saddle points without knowledge of the final states," *Journal of Chemical Physics*, vol. 121, pp. 9776-9792, 2004.
- [14] S. Trygubenko and D. Wales, "A doubly nudged elastic band method for finding transition states," *Chemical Physics*, vol. 120, pp. 2082-2094, 2004.
- [15] C. Reddy and H. Chiang, "Stability boundary based method for finding saddle points on potential energy surfaces," *Journal of Computational Biology*, vol. 13, pp. 745-766, 2006.
- [16] A. Goodrow, A. T. Bell, and M. Head-Gordon, "Transition state-finding strategies for use with the growing string method," *Journal of Chemical Physics*, vol. 130, 2009.
- [17] J. del Campo and A. Koster, "A hierarchical transition state search algorithm," *Journal of Chemical Physics*, vol. 129, p. 12, 2008.

- [18] G. Henkelman and H. Jónsson, "A Dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives," *Journal of Chemical Physics*, vol. 111, p. 7010, 1999.
- [19] S. Bungay, R. Poirier, and R. Charron, "Optimization of transition state structures using genetic algorithms," *Journal of Mathematical Chemistry*, vol. 28, pp. 389-401, 2000.
- [20] P. Chaudhury and S. Bhattacharyya, "A simulated annealing based technique for locating first-order saddle points on multidimensional surfaces and constructing reaction paths: several model studies," *Journal of Molecular Structure: Theochem*, vol. 429, pp. 175-186, 1998.
- [21] P. Chaudhury, S. Bhattacharyya, and W. Quapp, "A genetic algorithm based technique for locating first-order saddle point using a gradient dominated recipe," *Chemical Physics*, vol. 253, pp. 295-303, 2000.
- [22] D. Goldberg and J. Richardson, "Genetic algorithms with sharing for multimodal function optimization," in *the Second International Conference on Genetic Algorithms on Genetic algorithms and their application*, Cambridge, Massachusetts, United States, 1987, pp. 41-49.
- [23] X. Yu, K. Tang, T. Chen, and X. Yao, "Empirical analysis of evolutionary algorithms with immigrants schemes for dynamic optimization," *Memetic Computing*, vol. 1, pp. 3-24, 2009.
- [24] J. Bacardit, E. Burke, and N. Krasnogor, "Improving the scalability of rule-based evolutionary learning," *Memetic Computing*, vol. 1, pp. 55-67, 2009.
- [25] R. Lung and D. Dumitrescu, "A new evolutionary model for detecting multiple optima," in *the 9th Annual Conference on Genetic and Evolutionary Computation*, London, England, 2007, p. 1303.
- [26] G. Singh and K. Deb, "Comparison of multi-modal optimization algorithms based on evolutionary algorithms," in *GECCO '06*, Seattle, Washington, USA, 2006, p. 1312.
- [27] Y. Jin, "Knowledge Incorporation in Initializaion, Recombination and Mutation," in *Knowledge Incorporation in Evolutionary Computation*, Springer, 2005.
- [28] Y. Ong, M. Lim, and X. Chen, "Research Frontier: Memetic Computation - Past, Present & Future," *IEEE Computational Intelligence Magazine*, vol. 5, pp. 24-36, 2010.
- [29] P. Moscato, "On evolution, search, optimization ,genetic algorithms and martial arts: Towards memetic algorithms," California Institute of Technology, Pasadena, California, USA, 1989.
- [30] F. Neri, J. Toivanen, G. Cascella, and Y. Ong, "An adaptive multimeme algorithm for designing HIV multidrug therapies," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 4, pp. 264-278, 2007.
- [31] S. Handoko, C. Kwok, and Y. Ong, "Feasibility Structure Modeling: An Effective Chaperon for Constrained Memetic Algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 14, pp. 740-758, 2010.
- [32] L. M., Y. Ong, Y. Jin, and B. Sendhoff, "Lamarckian memetic algorithms: local optimum and connectivity structure analysis," *Memetic Computing*, vol. 1, pp 175-190, 2009.

- [33] Q. Nguyen, Y. Ong, and M. Lim, "A Probabilistic Memetic Framework," *IEEE Transactions on Evolutionary Computation*, vol. 13, pp. 604-623, 2009.
- [34] P. Moscato, "Memetic algorithms: A short introduction," *Mcgraw-Hill's Advanced Topics In Computer Science Series*, pp. 219-234, 1999.
- [35] H. Lai and S. Yang, "Saddle point and duality in the optimization theory of convex set functions," *The ANZIAM Journal*, vol. 24, pp. 130-137, 2009.
- [36] J. Bofill and J. Anglada, "Finding transition states using reduced potential-energy surfaces," *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, vol. 105, pp. 463-472, 2001.
- [37] J. Bofill, "Updated Hessian matrix and the restricted step method for locating transition structures," *Journal of Computational Chemistry*, vol. 15, pp. 1-11, 2004.
- [38] J. Baker, "An algorithm for the location of transition states," *Journal of Computational Chemistry*, vol. 7, pp. 385-395, 1986.
- [39] J. Simons, P. Joergensen, H. Taylor, and J. Ozment, "Walking on potential energy surfaces," *The Journal of Physical Chemistry*, vol. 87, pp. 2745-2753, 1983.
- [40] M. powell, *Math Prog*, p. 1:26, 1971.
- [41] H. Schlegel, "Optimization of equilibrium geometries and transition structures," *Journal of Computational Chemistry*, vol. 3, pp. 214-218, 1982.
- [42] P. Culot, G. Dive, V. Nguyen, and J. Ghuysen, "A quasi-Newton algorithm for first-order saddle-point location," *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, vol. 82, pp. 189-205, 1992.
- [43] R. Malek and N. Mousseau, "Dynamics of Lennard-Jones clusters: A characterization of the activation-relaxation technique," *Physical Review E*, vol. 62, pp. 7723-7728, 2000.
- [44] M. Yun, R. Lavery, N. Mousseau, K. Zakrzewska, and P. Derreumaux, "ARTIST: An activated method in internal coordinate space for sampling protein energy landscapes," *Proteins-Structure Function and Bioinformatics*, vol. 63, pp. 967-975, 2006.
- [45] X. Wu, J. Yang, J. Hou, and Q. Zhu, "Defects-enhanced dissociation of H-2 on boron nitride nanotubes," *Journal of Chemical Physics*, vol. 124, 2006.
- [46] L. Xu, G. Henkelman, C. Campbell, and H. Jonsson, "Pd diffusion on MgO(100): The role of defects and small cluster mobility," *Surface Science*, vol. 600, pp. 1351-1362, 2006.
- [47] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical recipes in C*. Cambridge MA, USA: Cambridge Univeristy Press, 1992.
- [48] G. Henkelman, G. Jóhannesson, and H. Jónsson, "Methods for finding saddle points and minimum energy paths," *Progress on Theoretical Chemistry and Physics*, vol. 111, pp. 269-300, 2000.
- [49] M. Dewar, E. Healy, and J. Stewart, "Location of transition states in reaction mechanisms," *Journal of the Chemical Society, Faraday Transactions 2*, vol. 80, pp. 227-233, 1984.
- [50] S. Fischer and M. Karplus, "Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom," *Chemical Physics Letter*, vol. 194, pp. 252-261, 1992.

- [51] G. Henkelman, B. Uberuaga, and H. Jónsson, "A climbing image nudged elastic band method for finding saddle points and minimum energy paths," *The Journal of Chemical Physics*, vol. 113, p. 9901, 2000.
- [52] H. Jonsson, G. Mills, and K. Jacobsen, "Nudged elastic band method for finding minimum energy paths of transitions," *Classical and Quantum Dynamics in Condensed Phase Simulation*, pp. 385-404, 2012.
- [53] C. Cerjan and W. Miller, "On finding transition states," *The Journal of Chemical Physics*, vol. 75, p. 2800, 1981.
- [54] A. Shtub, L. LeBlanc, and Z. Cai, "Scheduling programs with repetitive projects: a comparison of a simulated annealing, a genetic and a pair-wise swap algorithm," *European Journal of Operational Research*, vol. 88, pp. 124-138, 1996.
- [55] L. Ingber and B. Rosen, "Genetic algorithms and very fast simulated reannealing: A comparison," *Mathematical and Computer Modelling*, vol. 16, pp. 87-87, 1992.
- [56] L. Davis, *Genetic algorithms and simulated annealing*, Morgan Kaufmann Publishers, 1987.
- [57] M. Zheng, "Ground water management optimisation using genetic algorithms and simulated annealing: Formulation and comparison," *Journal of American. Water Research. Association.*, vol. 34, pp. 519-530, 1998.
- [58] R. Lung and D. Dumitrescu, "Roaming optimization: A new evolutionary technique for multimodal optimization," *Studia Universitatis Babes-Bolyai-Informatica*, vol. 49, pp. 99-109, 2004.
- [59] Z. Zhu, Y. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognition*, vol. 40, pp. 3236-3248, 2007.
- [60] Q. Nguyen, Y. Ong, H. Soh, and J. Kuo, "Multiscale Approach to Explore the Potential Energy Surface of Water Clusters (H₂O)_n n=8," *The Journal of Physical Chemistry A*, vol. 112, pp. 6257-6261, 2008.
- [61] Z. Zhu, Y. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics*, vol. 37, pp. 70-76, 2007.
- [62] Z. Zhou, Y. Ong, M. Lim, and B. Lee, "Memetic algorithm using multi-surrogates for computationally expensive optimization problems," *Soft Computing*, vol. 11, pp. 957-971, 2007.
- [63] K. Wong, Y. Ong, R. Eren, and C. Fung, "Hybrid fuzzy modelling using memetic algorithm for hydrocyclone control," in *International Conference on Machine Learning and Cybernetics*, Shanghai, China, 2004, pp. 4188-4193.
- [64] J. Tang, M. Lim, Y. Ong, and M. Er, "Parallel memetic algorithm with selective local search for large scale quadratic assignment problems," *International Journal of Innovative Computing Information and Control*, vol. 2, pp. 1399-1416, 2006.
- [65] J. Tang, M. Lim, and Y. Ong, "Diversity-adaptive parallel memetic algorithm for solving large scale combinatorial optimization problems," *Soft Computing*, vol. 11, pp. 873-888, 2007.
- [66] Y. S. Ong, P. B. Nair, and A. J. Keane, "Evolutionary optimization of computationally expensive problems via surrogate modeling," *AIAA journal*, vol. 41, pp. 687-696, 2003.

- [67] Y. Ong, M. Lim, N. Zhu, and K. Wong, "Classification of adaptive memetic algorithms: a comparative study," *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics*, vol. 36, pp. 141-152, 2006.
- [68] Q. Nguyen, Y. Ong, M. Lim, and N. Krasnogor, "Adaptive Cellular Memetic Algorithms," *Evolutionary Computation*, vol. 17, pp. 231-256, 2009.
- [69] G. Lin, Y. Liang, L. Kang, X. Yao, and R. Sarker, "An Analysis the Exploration Abilities of Crossover Operators," in *3rd International Conference on Intelligence Computation and Applications*, Wuhan, China, 2008, pp. 334-338.
- [70] D. Gouwanda and S. Ponnambalam, "Evolutionary Search Techniques to Solve Set Covering Problems," in *Conference of the World-Academy-of-Science-Engineering-and-Technology*, Bangkok, Thailand, 2008, pp. 20-25.
- [71] C. Zhu, H. Kitagawa, and C. Faloutsos, "Example-based robust outlier detection in high dimensional datasets," in *5th IEEE International Conference on Data Mining*, Houston, Texas, USA, 2005, pp. 829-832.
- [72] J. Lee, S. Oh, C. Choi, and H. Jeong, "Flow estimation using genetic algorithm and neural network," presented at the ICIT, IEEE International Conference on Industrial Technology, 2009.
- [73] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, 1996.
- [74] S. Mahfoud, "Crowding and preselection revisited," *Urbana*, vol. 51, 1992.
- [75] O. Mengshoel and D. Goldberg, "Probabilistic crowding: Deterministic crowding with probabilistic replacement," *Urbana*, vol. 51, 1992.
- [76] F. Streichert, G. Stein, H. Ulmer, and A. Zell, "A clustering based niching EA for multimodal search spaces," *Lecture Notes in Computer Science*, pp. 293-304, 2004.
- [77] G. Harik, "Finding multimodal solutions using restricted tournament selection," in *Proceedings of Sixth International Conference on Genetic Algorithms*, 1995.
- [78] A. Petrowski, "A clearing procedure as a niching method for genetic algorithms," in *Proceedings of IEEE International Conference on Evolutionary Computation*, 1996, pp. 798-803.
- [79] J. Gan and K. Warwick, "Dynamic niche clustering: a fuzzy variable radius niching technique for multimodal optimization in GAs," in *Proceedings of the 2001 Congress on Evolutionary Computation*, 2001, pp. 215-222.
- [80] D. Lagorce, T. Pencheva, B. Villoutreix, and M. Miteva, "DG-AMMOS: A New tool to generate 3D conformation of small molecules using Distance Geometry and Automated Molecular Mechanics Optimization for in silico Screening," *BMC Chemical Biology*, vol. 9, p. 6, 2009.
- [81] P. Surry, "A Prescriptive Formalism for Constructing Domain-specific Evolutionary Algorithms," Ph.D. Thesis, University of Edinburgh, 1998.
- [82] S. Kok and C. Sandrock, "Locating and Characterizing the Stationary Points of the Extended Rosenbrock Function," *Evolutionary Computation*, vol. 17, pp. 437-453, 2009.
- [83] J. Vitela and O. Castaños, "A real-coded niching memetic algorithm for continuous multimodal function optimization," in *IEEE Congress on Evolutionary Computation, 2008. CEC 2008.*, Autonoma de Mexico, Mexico City, Mexico, 2008, pp. 2170-2177.

- [84] M. Ellabaan and Y. Ong, "Valley-Adaptive Clearing Scheme for Multimodal Optimization Evolutionary Search," in *9th International Conference on Intelligent Systems Design and Applications (ISDA'09)*, Pisa, Italy, 2009, pp. 1-6.
- [85] O. Shir, M. Emmerich, and T. Bäck, "Adaptive niche radii and niche shapes approaches for niching with the CMA-ES," *Evolutionary Computation*, vol. 18, pp. 97-126, 2010.
- [86] H. Marel and H. Beutelspacher, *Atlas of infrared spectroscopy of clay minerals and their admixtures*, Elsevier Publishing Company, 1976.
- [87] M. Woolfson, "X-ray Crystallography," *Cambridge University Press*, 1970.
- [88] A. Wilson, *Elements of X-ray Crystallography*, Addison-Wesley Menlo-Park, 1970.
- [89] E. Kemble, *The fundamental principles of quantum mechanics: with elementary applications*, vol. 472, Dover New York, 1958.
- [90] J. Bayma, *The elements of molecular mechanics*, Macmillan and co., 1866.
- [91] P. Atkins and R. Friedman, *Molecular quantum mechanics*, vol. 3, Oxford University Press Oxford, 1997.
- [92] G. Robinson, *Water in biology, chemistry, and physics: experimental overviews and computational methodologies*, World Scientific Pub Co Inc, 1996.
- [93] K. Liu, J. Cruzan, and R. Saykally, "Water clusters," *Science*, vol. 271, p. 929, 1996.
- [94] D. Hartmann, "Climate change: tropical surprises," *Science*, vol. 295, p. 811, 2002.
- [95] T. Wigley and S. Raper, "Thermal expansion of sea water associated with global warming," *Nature*, vol. 330, pp. 127-131, 1987.
- [96] R. Ludwig, "Water: from clusters to the bulk," *Angewandte Chemie International Edition*, vol. 40, pp. 1808-1827, 2001.
- [97] S. Maheshwary, N. Patel, N. Sathyamurthy, A. Kulkarni, and S. Gadre, "Structure and Stability of Water Clusters (H₂O)_{n=8-20}: An Ab Initio Investigation," *Journal of Physical Chemistry A*, vol. 105, pp. 10525-10537, 2001.
- [98] C. Gruenloh, J. Carney, C. Arrington, T. Zwier, S. Fredericks, and K. Jordan, "Infrared spectrum of a molecular ice cube: the S4 and D2d water octamers in benzene-(water)₈," *Science*, vol. 276, p. 1678, 1997.
- [99] M. Viant, J. Cruzan, D. Lucas, M. Brown, K. Liu, and R. Saykally, "Pseudorotation in water trimer isotopomers using terahertz laser spectroscopy," *Journal of Physical Chemistry A*, vol. 101, pp. 9032-9041, 1997.
- [100] F. Fontaine-Vive, *From dynamics to structure and function of model biomolecular systems*, Delft University Press, 2007.
- [101] H. Soh, Y. Ong, C. Quoc, H. Quang, M. Habibullah, T. Hung, and J. Kuo, "Discovering Unique, Low-Energy Pure Water Isomers: Memetic Exploration, Optimization, and Landscape Analysis," *IEEE Transactions on Evolutionary Computation*, vol. 14, pp. 419-437, 2010.
- [102] D. Wales, *Energy Landscapes: With Applications to Clusters, Biomolecules and Glasses*, Cambridge University Press Cambridge, 2003.
- [103] C. Brooks, J. Onuchic, and D. Wales, "Statistical thermodynamics. Taking a walk on a landscape," *Science*, vol. 293, p. 612, 2001.
- [104] F. Jensen, *Introduction to computational chemistry*, vol. 2, Wiley New York, 1999.

- [105] T. Truong and D. Truhlar, "Ab initio transition state theory calculations of the reaction rate for $\text{OH} + \text{CH}_3\text{OH} \rightarrow \text{CH}_3\text{O} + \text{H}_2$," *The Journal of Chemical Physics*, vol. 93, p. 1761, 1990.
- [106] G. Henkelman, G. Jóhannesson, and H. Jónsson, "Methods for finding saddle points and minimum energy paths," *Theoretical Methods in Condensed Phase Chemistry*, pp. 269-302, 2002.
- [107] J. Hernández-Rojas, J. Bretón, J. Gomez Llorente, and D. Wales, "Global potential energy minima of $\text{C}_{60}(\text{H}_2\text{O})_n$ clusters," *Journal of Physical Chemistry B*, vol. 110, pp. 13357-13362, 2006.
- [108] M. Day, K. Kirschner, and G. Shields, "Global search for minimum energy $(\text{H}_2\text{O})_{n=3-5}$ clusters," *Journal of Physical Chemistry A*, vol. 109, pp. 6773-6778, 2005.
- [109] Y. Choi, C. Pak, and K. Kim, "Electric field effects on water clusters ($n = 3-5$): systematic ab initio study of structures, energetics, and transition states," *The Journal of Chemical Physics*, vol. 124, p. 094308, 2006.
- [110] M. Ellabaan, Y. Ong, Q. Nguyen, and J. Kuo, "Evolutionary Discovery of Transition States in Water Clusters," *Journal of Theoretical and Computational Chemistry*, vol. 11, 2012.
- [111] Y. Ong and A. Keane, "Meta-Lamarckian learning in memetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 8, pp. 99-110, 2004.
- [112] J. Kuo and M. Klein, "Structure of protonated water clusters: Low-energy structures and finite temperature behavior," *The Journal of Chemical Physics*, vol. 122, p. 024516, 2005.
- [113] W. Swope, H. Andersen, P. Berens, and K. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *The Journal of Chemical Physics*, vol. 76, p. 637, 1982.
- [114] S. Maeda, Y. Matsuda, S. Mizutani, A. Fujii, and K. Ohno, "Long-Range Migration of a Water Molecule To Catalyze a Tautomerization in Photoionization of the Hydrated Formamide Cluster," *The Journal of Physical Chemistry A*, vol. 114, pp. 11896-11899, 2010.
- [115] Q. Nguyen, Y. Ong, and J. Kuo, "A Hierarchical Approach to Study the Thermal Behavior of Protonated Water Clusters $\text{H}^+(\text{H}_2\text{O})_n$," *Journal of Chemical Theory and Computation*, vol. 5, pp. 2629-2639, 2009.
- [116] M. Ellabaan, S.D. Handoko, Y. S. Ong, C.K. Kwok, S. Bahnassy, F. Ellassawy, and H.Y. Man, "A Tree-structured Covalent-bond-driven Molecular Memetic Algorithm for Optimization of Ring-deficient Molecules", *Computers and Mathematics with Applications*, Vol. 64, 2012.
- [117] O. Mishima and H. Stanley, "The relationship between liquid, supercooled and glassy water," *Nature*, vol. 396, pp. 329-335, 1998.
- [118] R. Weast, *Handbook of Chemistry and Physics*, CRC Press, 1984.
- [119] W. Kabsch, "A solution of the best rotation to relate two sets of vectors " *Acta Crystallographica* vol. 32, 1976.
- [120] M. Ellabaan, Y. Ong, S. Handoko, C. Kwok, S. Bahnassy, and F. Ellassawy, "TransGene, 1.0: A Web Service For Transition States Discovery in Small Ring-

- exclusive Molecules Using First-Principles Calculation," presented at the ISMB 2012 (Poster). Long Beach, California, USA, 2012.
- [121] M. Ellabaan, Y. Ong, S. Handoko, C. Kwoh, S. Bahnassy, and F. Ellassawy, "Toward Better Understanding of Glutamatergic Systems: Glutamate isoforms discovery and Receptor Binding Affinity," presented at the ISMB 2012 (Poster). Long Beach, California, USA, 2012.
- [122] X. S. Chen, Y. S. Ong, "A Conceptual Modeling of Meme Complexes in Stochastic Search," *IEEE Transactions On Systems, Man and Cybernetics - Part C*, Vol. 42, 2012.
- [123] R. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Mathematics of Operations Research*, vol. 1, pp. 97-116, 1976.
- [124] S. Rasmussen, L. Chen, M. Nilsson, and S. Abe, "Bridging nonliving and living matter," *Artificial Life*, vol. 9, pp. 269-316, 2003.
- [125] H. Fellermann, S. Rasmussen, H. Ziock, and R. Solé, "Life cycle of a minimal protocell-a dissipative particle dynamics study," *Artificial Life*, vol. 13, pp. 319-345, 2007.