Implementation of machine learning and other simulation protocols for the representation of G-Quadruplex stabilizer

Chiew, Kang Jing

2019

Chiew, K. J. (2019). Implementation of machine learning and other simulation protocols for the representation of G-Quadruplex stabilizer. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/83241>

<https://doi.org/10.32657/10356/83241>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**IMPLEMENTATION OF MACHINE LEARNING AND
OTHER SIMULATION PROTOCOLS FOR THE
REPRESENTATION OF G-QUADRUPLEX STABILIZER**

CHIEW KANG JING

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

2019

**IMPLEMENTATION OF MACHINE LEARNING AND
OTHER SIMULATION PROTOCOLS FOR THE
REPRESENTATION OF G-QUADRUPLEX STABILIZER**

CHIEW KANG JING

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

A thesis submitted to the Nanyang Technological
University in partial fulfilment of the requirement for the
degree of Master of Science

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

[Input Date Here]
17 September 2019
.....
Date

[Input Signature Here]
CHIEW KANG JING
.....
[Input Name Here]

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

[Input Date Here]

17 September 2019

.....
Date

[Input Supervisor Signature Here]

... Asst Prof Chen Gang ...

.....
[Input Supervisor Name Here]

Dr. LU YUNPENG

Authorship Attribution Statement

Please select one of the following; *delete as appropriate:

*(A) This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

~~*(B) This thesis contains material from [x number] paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.~~

[Input Date Here]
17 September 2019
.....
Date

[Input Signature Here]
CHIEW KANG JING
.....
[Input Name Here]

Abstract

Human chromosomal telomeres are capable of forming G-Quadruplex structures, which can inhibit the activity of telomerase, an enzyme commonly found in cancerous cells and largely responsible for their immortality. The inhibition may therefore lead to cell apoptosis in these cancer cells. As such, there have been growing interest in the research for G-Quadruplex stabilizers. This thesis focusses on the exploration and development of new representations for the structure-function relationship of G-Quadruplex-ligand biomolecular system using machine learning (ML) techniques. Two main models, namely Element Specific Persistent Homology (ESPH) and Rigidity Index-Score (RI-Score), were adapted due to their successes in representing other biomolecular systems. It was discovered that both methodologies similarly presented strong average positive correlation for the representation of the studied system. (Pearson Correlation: 0.6770-0.7871, RMSE: 0.4621-0.5811) In addition, most of the models developed performed admirably when compared to the well-established Quantitative Structure-Activity Relationship (QSAR) method. In particular, Ligand-based model using ESPH firmly outperformed all other models, new and existing alike. (Pearson Correlation: 0.7614-0.7871, RMSE: 0.4621-0.4866)

Acknowledgements

Firstly, I would like to express my most sincere appreciation to both my supervisor, Asst Prof Chen Gang, and my co-supervisor, Dr Lu Yunpeng, for their continuous guidance and support during the entire duration of my MSc research. It is my privilege to be given the opportunity to take on this interesting topic and learn from both of them.

Secondly, I would like to thank Asst Prof Xia Kelin and my mentor, Dr Meng Zhenyu, for their kind patience in guiding me through this research field. Their advice and knowledge helped me tremendously. I am grateful to have learn from them through my MSc research.

I would also like to thank Nanyang Technological University and the School of Physical and Mathematical Sciences for allowing me the chance to be in the integrated program to perform my MSc research.

Lastly, I would like to thank my family and friends for the support and encouragement in this journey.

Table of Contents

Abstract	1
Acknowledgements	2
Table of Contents	3
List of Tables	5
List of Figures	6
Abbreviations.....	8
Chapter 1 Introduction	10
1.1 G-Quadruplex	10
1.1.1 Structure and Topology	11
1.1.2 Human Telomeric G-Quadruplex	13
1.2 Artificial Intelligence and Machine Learning	16
Chapter 2 Methodology	19
2.1 Element Specific Persistent Homology (ESPH)	19
2.1.1 Ligand-based model	25
2.1.2 Complex-based model	30
2.1.3 Results	36
2.2 Rigidity Index-Score (RI-Score)	42
2.2.1 Results	46
Chapter 3 Comparison with Existing Methodologies	53
Chapter 4 Internal Validity Test	58
Chapter 5 Limitations	61

5.1	All Models	61
5.2	Complex-based and RI-Score Models	62
Chapter 6	Conclusion and Future Work	64
6.1	Conclusion	64
6.2	Future Work	66
References		68
Appendix A		73
Appendix B		101
Appendix C		109
Appendix D		121
Appendix E		133
Appendix F		145
Appendix G		152
Appendix H		154
Appendix I		157
Appendix J		165

List of Tables

Chapter 2: Methodology

Table 2-1:	Comparison between various modifications of the Ligand-based model	36
Table 2-2:	Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the Ligand-based model.....	37
Table 2-3:	Comparison between various modifications of the Complex-based model.....	39
Table 2-4:	Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the Complex-based model.....	40
Table 2-5:	Comparison between Ligand-based and Complex-based model	41
Table 2-6:	Comparison between various modifications of the exponential kernel model.....	46
Table 2-7:	Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the exponential kernel model.....	47
Table 2-8:	Comparison between various modifications of the Lorentz kernel model.....	48
Table 2-9:	Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the Lorentz kernel model.....	50
Table 2-10:	Ranking across all developed models.....	51

Chapter 3: Comparison with Existing Methodologies

Table 3-1:	Average standard deviations (SD) and coefficient of variation (CV) within and between Ligand-based ESPH and CORAL QSAR models.....	56
------------	--	----

Chapter 4: Internal Validity Test

Table 4-1:	Comparison between the Ligand-based models when subjected to different input data set	59
------------	---	----

List of Figures

Chapter 1: Introduction

Figure 1-1:	General structure of a G-Quadruplex	10
Figure 1-2:	Schematic diagram of the three main categories of strand directionality	11
Figure 1-3:	Schematic illustration of intra and intermolecular G-Quadruplex	12
Figure 1-4:	Schematic depiction of the three types of loop arrangements	13
Figure 1-5:	Glycosidic bond angles adopted by guanine units	13
Figure 1-6:	Effects of G-Quadruplex on telomerase	15
Figure 1-7:	Schematic figures of the various modes of binding interactions	16

Chapter 2: Methodology

Figure 2-1:	Depiction of topological invariants	20
Figure 2-2:	Simplified example of a filtration process	22
Figure 2-3:	Example of MTF analysis of C ₇₀ fullerene	23
Figure 2-4:	Molecular structure of amidoanthracene-9,10dione	27
Figure 2-5:	ESPH persistent barcodes generated using C atoms	28
Figure 2-6:	ESPH persistent barcodes generated using N atoms	28
Figure 2-7:	ESPH persistent barcodes generated using O atoms	29
Figure 2-8:	Selected G-Quadruplex structures as visualized by AutoDock Vina	31
Figure 2-9:	Screenshot of a completed AutoDock Vina docking simulation	32
Figure 2-10:	π -stacking interaction between G-Quadruplex and stabilizers.	33

Chapter 3: Comparison with Existing Methodologies

Figure 3-1: Overview and comparison between new and existing methods	
.....	55

Abbreviations

AI	Artificial Intelligence
C	Carbon
Cl	Chlorine
CORAL	CORrelations And Logic
CV	Coefficient of Variation
DNA	Deoxyribonucleic Acid
EC	Extended Connectivity
ESPH	Element Specific Persistent Homology
EU	European Union
F	Fluorine
FRI	Flexibility-Rigidity Index
G4LDB	G-Quadruplex Ligands Database
GBT	Gradient Boosting Tree
G-tetrad	Guanine Tetrad
HALO	Halogens (Used in CORAL)
HSG	Hydrogen Suppressed Graph
I	Iodine
IC ₅₀	Half Maximal Inhibitory Concentration
IPH	Interactive Persistent Homology
LPF	Low-Pass Filter
ML	Machine Learning
MTF	Molecular Topological Fingerprint
N	Nitrogen

NOSP	Nitrogen, Oxygen, Sulfur, Phosphorus (Used in CORAL)
NMR	Nuclear Magnetic Resonance
O	Oxygen
P	Phosphorus
P(IC ₅₀)	-log (IC ₅₀)
PDB	Protein Data Bank
QSAR	Quantitative Structure-Activity Relationship
RCSB	Research Collaboratory for Structural Bioinformatics
RF	Random Forest
RI-Score	Rigidity Index-Score
RMSE	Root Mean Square Deviation
S	Sulfur
SMILES	Simplified Molecular-Input Line-Entry System
SD	Standard Deviation

Chapter 1: Introduction

1.1 G-Quadruplex

G-Quadruplex is a type of nucleic acid secondary structure that may exist in deoxyribonucleic acid (DNA) containing guanine rich sequences.¹⁻² The formation of a G-Quadruplex structure may proceed through either an intra or intermolecular process. This means that the structure can be formed from a range of one to four different guanine rich DNA strands.³⁻⁴ The main structure of a G-Quadruplex is, however, consistent regardless of the number of strands involved in its formation. A G-Quadruplex structure consists of multiple layers of planes, usually between two to four layers, stacked on top of each other in a parallel fashion.^{1,3} Within each layer, four guanine bases are held in a square planar conformation by hydrogen bonds through Hoogsteen base pairings to form a guanine tetrad (G-tetrad). In addition, G-Quadruplex structure is also stabilized by the π -stacking interactions between the layers, as well as monovalent cations commonly found in the human biological system, such as sodium Na^+ and potassium K^+ ions. These ions interact with the structure by intercalating between the layers of the G-tetrad and coordinating to the oxygen atoms in the guanine units to minimize the electrostatic repulsion.^{2,5-6}

Figure 1-1: General structure of a G-Quadruplex. (Left) Alignment within a G-tetrad layer, which consist of four guanine units held together in a planar conformation by hydrogen bonds. The metal ion is located at the central cavity to stabilize the structure. (Right) Schematic representation of multiple G-tetrad layers stacking to form a G-Quadruplex structure, supported by intervening loops and π -stacking interactions. Figure is adapted from Reference 7.

1.1.1 Structure and Topology

Using laboratory techniques such as X-ray crystallography and Nuclear Magnetic Resonance (NMR), G-Quadruplex structures were observed to be polymorphic biomolecular structures, which means that they can appear in various forms. This means that a wide range of G-Quadruplex topologies may exist, and they can be discussed through numerous parameters, such as the strand orientations, the number of strands involved, the loop arrangement and the glycosidic bond angle.^{2,7}

Strand orientations

One distinct feature most commonly used to identify and classify G-Quadruplex structures is the strand orientations, which refers to the directions of the DNA strands in the loop surrounding the G-tetrad layers.⁶ Three categories exist, namely parallel (4+0), anti-parallel (2+2) and hybrid (3+1) G-Quadruplex structures.⁵ In a parallel G-Quadruplex, all of the four DNA strands are directed in the same way, whereas in an anti-parallel G-Quadruplex, two of the DNA strands would be directed in an opposite manner with respect to the other two strands. The third category is the hybrid G-Quadruplex, in which structures are classified as such only when a single DNA strand differs in its direction from the other three strands.^{2,3,8}

Figure 1-2: Schematic diagram of the three main categories of strand directionality. (Left) Parallel (4+0) G-Quadruplex structure. (Middle) Hybrid (3+1) G-Quadruplex structure. (Right) Anti-Parallel (2+2) G-Quadruplex structure. Figure is adapted from Reference 9.

Number of DNA strands

G-Quadruplex structures may also be identified based on the number of DNA strands that they are comprised of. G-Quadruplex structures can be formed

intramolecularly by a lone continuous DNA strand and hence known as monomeric. These structures can also be formed through an intermolecular process between two or four strands, thus forming dimeric and tetrameric G-Quadruplex structures.

Figure 1-3: Schematic illustration of intramolecular G-Quadruplex (Monomer) and intermolecular G-Quadruplex structures (Dimer and Tetramer). Figure is adapted from Reference 4.

Loop arrangement

G-Quadruplex structures may also differ with the intervening loops which hold the backbone of the G-tetrad layers. These loops may vary based on their length or order, but they are generally small nucleotide chains, consisting of only one to seven units. In fact, the size of the loops affects the stability of the G-Quadruplex, the shorter the loops, the more stable the structure is.⁵ Three categories of loops are known, namely strand-reversal loop, lateral loop and diagonal loop. Strand-reversal loop, also known as propeller loop, connects two adjacent strands that are aligned in a parallel manner relative to each other. Lateral loop, also known as edgewise loop, connects two adjacent strands that are aligned in an anti-parallel manner relative to each other. Diagonal loop, on the other hand, connects two strands that are located opposite of each other and aligned in an anti-parallel manner relative to each other.²

Figure 1-4: Schematic depiction of the three types of loop arrangements. (Left) Diagonal loop. (Middle) Edgewise loop. (Right) Strand-reversal loop. Figure is adapted from Reference 9.

Glycosidic bond angles

G-Quadruplex structures may also be differentiated based on the guanine arrangements in the G-tetrad layers. Guanine residues may adopt two conformations, syn-glycosidic or anti-glycosidic, relative to the sugar unit in the nucleotide. Syn-glycosidic conformation refers to the conformation in which the six-membered ring of the guanine residue is aligned with the sugar unit in the nucleotide. Anti-glycosidic conformation refers to the conformation in which the six-membered ring of the guanine residue is aligned away from the sugar unit, as seen in Figure 1-5.^{2,7}

Figure 1-5: Glycosidic bond angles adopted by guanine units. (Left) Anti-glycosidic conformer. (Right) Syn-glycosidic conformer. Figure is adapted from Reference 9.

1.1.2 Human Telomeric G-Quadruplex

In the human biological system, it was estimated computationally that there exist over 300 000 DNA sequences which have the potential to form G-Quadruplex structures.^{1,3,7} Among the sequences, the bulk of them were discovered to be at the human telomeric regions. The human telomeric G-

Quadruplex also happened to be the first biologically applicable G-Quadruplex formation to be identified.³⁻⁴ Human telomere refers to the overhang found at the tail end of each human chromosome which consists of usually 100-200 repeated nucleotide sequences, known as TTAGGG repeats, that are guanine rich.⁴ The main purpose of the telomeres is to safeguard the integrity of the chromosomes by preventing them from self-decaying or binding to neighboring chromosomes.¹⁰⁻¹¹ Therefore, telomeres are of high importance in a biological system.

During every DNA replication cycle, telomeres suffer from a shortening effect due to the mechanism in which the process occurs through. Since human DNAs are double stranded and that DNA polymerases only proceed in the direction of 3' to 5' end, these result in the phenomenon of Okazaki fragments along the lagging strand during DNA replication. Ultimately, a portion at the 5' tail of the telomere will be lost after each replication process. The shortening of telomeres may lead to serious consequences such as cell death.¹⁰⁻¹¹ In 1984, Nobel Prize winners Blackburn, E. H. and Greider, C. W., discovered that an enzyme, known as telomerase, not only had the ability to counteract the shortening effect of telomeres, but also able to lengthen them.¹² This means that the presence of telomerase would be able to immortalize cells.¹³ Telomerase is present in most cells, but the level of activity differs drastically across various types of cells. It was discovered that in most somatic cells, telomerase activity was almost negligible. On the other hand, telomerase is highly active in over 90% of cancer cells, thereby being a huge contributor to the immortality of cancer cells, thus making cancer a deadly and fatal disease.¹⁴

However, human telomeres, being guanine rich, are able to form intramolecular G-Quadruplex secondary structure. More importantly, the formation of these monomeric structures was observed to have inhibitory effects on telomerase activity.⁶ In some cases, cancer cells were noticed to undergo apoptosis because of G-Quadruplex stabilization.¹⁵ Hence, the stability and spontaneity with regards to G-Quadruplex formation may hold the key towards overcoming cancer in the future.

Figure 1-6: Effects of G-Quadruplex on telomerase. In normal somatic cell, DNA suffer from telomere shortening with every replication cycle, eventually leading to the death of the cell. In a cancer tumorous cell, the presence of telomerase activity ensures constant elongation, leading to cell immortality. However, the ability of human telomeres to fold into G-Quadruplex structures was discovered to inhibit telomerase activity within cancer cells. Figure is adapted from Reference 16.

In recent years, the ability to selectively stabilize human telomeric G-Quadruplex has increasingly become an area of interest for research.^{2,4} Researchers have particularly invested into the use of small molecules, also known as G-Quadruplex stabilizers/ligands, to stabilize G-Quadruplex structures through binding interactions, in hopes of producing therapeutic effects towards cancer ultimately.^{4,17} In fact, *in vitro* experiments have identified over hundreds of ligands capable of interacting with G-Quadruplex structures to varying degrees.¹⁸ These ligands interact with G-Quadruplex through π -stacking interaction on either the top or bottom most G-tetrad layer. Such a mode of interaction arises due to the planar aromatic nature of the G-tetrad layers, which also meant that strong G-Quadruplex stabilizers tend to have an extended π conjugated system.⁷ While π -stacking interaction is the most established binding mode for G-Quadruplex stabilizers, these stabilizers may theoretically interact with G-Quadruplex structures through other means, albeit

rare and unlikely, such as binding along the grooves or intercalation, in which the stabilizer inserts itself between the G-tetrad layers.^{7,17}

Figure 1-7: Schematic figures of the various modes of binding interactions. Figure is adapted from Reference 19.

1.2 Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) refers to intelligence displayed by machines, which are mainly employed to solve problems through various algorithms.²⁰ In the modern society, AI is largely involved in many applications across various industries, ranging from medical diagnosis in healthcare, to electronic trading in finance.²¹⁻²³ Recently, AI has also begun to be integrated into research works, particularly in the form of machine learning (ML). ML is a branch of AI which utilizes algorithms and mathematical models to train machines to enhance their performances in their assigned taskings.²⁴ ML techniques are increasingly being favored due to their ability to handle problems which involve large-scale probabilities and complex concepts, making them ideal for research involving biomolecular systems.²¹ In addition, ML has the capability to identify patterns within the data set and obtain higher order information such as spatial elements. In fact, ML techniques have become such an invaluable asset to biomolecular analysis that in recent years, nearly all modern computational data projection techniques consist some form of ML components.^{22,25}

An important element in ML techniques with respect to biomolecular structures is featurization, which is the process of converting the three-dimensional

structures of the molecules into features.²² Features refer to discrete and quantifiable properties that can be extracted from a data set and they can range from typical ones such as geometric and electrostatic, to the less conventional topological features. Geometric properties include atom coordinates and bond angles, while electrostatic properties involve charges and potentials of the atoms. Intuitively, the performance of a ML technique is dependent on the type and process of features extraction. While extracting geometric attributes is the most common approach towards studying the relationship between the structure and function of a molecule, the process often suffers from the overwhelming amount of details and complexity, making geometric features almost impossible to handle even through computational means. On the other hand, topological features, despite being a newer and unfamiliar approach, appear to outperform all common features in both results and speed.²⁶ The better performances could be attributed to the fact that topological features are neither restrained by coordinates nor affected by noise.²³ Topological features also provide a significantly simplified data for computation, which contributed to the performance.²⁵ Observations were also made that in many biomolecular structures, there was an apparent correlation between their topological features and functions, both physical and biological.²⁷ As such, many biomolecular researches have been carried out computationally using methodologies that capture topological features, yielding highly positive results and proving that these techniques present a good representation for the studied systems.^{22,25-28}

However, majority of these studies have only been performed on the protein-ligand system.^{22,25-27,29} Other biomolecular systems have either been rarely explored or entirely unexplored using such methodologies, including the interactions between G-Quadruplex structures and ligands. In fact, the overall research done on G-Quadruplex remains relatively sparse compared to other biomolecules, such as protein. A search of ‘G-Quadruplex’ on Google Scholar yielded a mere thirty-four thousand journal results as compared to close to six million journal results of ‘protein’, less than 1% of the total conducted for proteins. Given the importance of G-Quadruplex and its possible potential to curb cancer cells, this represents a glaring gap.

Therefore, our current research is focused on adapting various successful ML techniques that utilizes topological features to study the structure-function relationship in a G-Quadruplex-ligand biomolecular system. This thesis will describe the first stage of the research, which involves the development of a robust and accurate model to represent a G-Quadruplex-ligand system. This was achieved through the adaptation of two main methodologies, namely Element Specific Persistent Homology (ESPH) and Rigidity Index-Score (RI-Score).^{22,28} Several model variations based on the concepts were built and tested, to identify the most ideal model for the G-Quadruplex-ligand system. Assuming that there is indeed a topological-function relationship in all biomolecular systems, a strong correlation score would be obtained and therefore providing a potentially strong representation for the system of interest. Ultimately, a strong positive correlation would pave the path in the future to design a new methodology for the design of new G-Quadruplex stabilizers, which can in turn be used as a target drug for cancer treatment. This may come in the form of a set of guiding principles for the design of innovative ligands, which could be derived from the relations generated through the aforementioned techniques. The development of such a guideline would significantly reduce the financial and time cost associated with laboratory experiments while being an improved alternative to the current drug design processes.

Chapter 2: Methodology

Two topological techniques will be explored in this research. The first technique is known as ESPH, which is derived from the principles of classical topology and persistent homology. The second technique is RI-Score, which is essentially a rigidity-based ML approach that focuses on rigidity strengthening in the binding complex. This arises from the concept that a receptor becomes more rigid upon binding.

2.1 Element Specific Persistent Homology (ESPH)

As mentioned, ESPH was developed upon previous techniques, which include classical topology and persistent homology. Topology is the classical method while persistent homology is a comparatively newer technique in algebraic topology.^{22,25,27}

Mathematically, topology studies the spatial properties of elements which are maintained during ongoing deformation and distinguishes these independent elements in space. The main objective of topology is the extraction of qualities known as topological invariants, which refer to the inherent features in the fundamental topological spaces. These invariants include components, holes as well as voids, which are also labelled as Betti numbers, specifically Betti-0, Betti-1 and Betti-2 respectively.^{25,29-30} An illustration of Betti numbers can be seen in Figure 2-1. Simplicial complex can be contrived and utilized for the extraction of these topological invariants. Simplicial complex refers to the discretization of the topological space through geometric polytope such as simplex, whereby an N-simplex denotes a polytope consisting of $N + 1$ number of vertices. N-simplex denotes a point, line, triangle and tetrahedron for values of $N = 0, 1, 2, 3$ respectively.²⁵

	Point	Circle	Sphere	Torus

Betti-0:	1	1	1	1
Betti-1:	0	1	0	2
Betti-2:	0	0	1	1

Figure 2-1: Depiction of topological invariants. Figure is adapted from Reference 30.

Even though both topology and geometry are concerned with the shapes and sizes of figures, topological studies trump geometric methods mainly due to their capacity to simplify data, as topology can extract important information without being burdened by the extra structural information. However, this also leads to oversimplification, as insufficient geometric data are retained in the process to allow for applications towards biomolecules.²⁵ In classical topology, the topological space is described using the notion of proximity and connectivity as the basis. Thus, a space that undergoes continuous deformation is regarded as having the same topological features as its previously undeformed state. In other words, the topological invariants employed in classical topology oversimplifies the space, and thus inhibits the effectiveness of topology to distinguish biomolecular structures accurately, which tend to fold or warp into different forms spontaneously under different biological conditions.²²

Instead, persistent homology is the preferred topological approach to analyze biomolecular systems as opposed to classical topology. This is because persistent homology incorporates multi-spatial scales into the topological invariants employed, thus achieving a reciprocity between both topology and geometry. In addition, persistent homology is also able to study topological spaces across a variety of multiple scales, thereby obtaining a multiscale representation of the space. This is unlike that of classical topology which can only study a single scale.^{22,25}

The basic idea of persistent homology is to consider each given atom position as a point cloud that turns into a sphere which continuously increases its radius at a steady rate. As neighboring spheres overlap with each other over time, the resulting simplices generated are filled up. For example, any two points would form a one-dimensional edge while any three points would form a two-dimensional face.³¹ Since these topological invariants are measured across an interval, they would therefore appear at a particular value d_x and disappear at another d_y .²⁹ In other words, each simplex can be visualized and labelled as a bar (d_x, d_y) , whereby the start of the bar d_x is labelled as the ‘birth’ and the end at d_y is known as the ‘death’. In addition, the difference between the death and the birth value, labelled as $d_y - d_x$ is known as the persistence of the individual invariant.^{25,27,31} This entire sequence of generating simplicial complex over a spatial scale is unique to persistent homology and referred to as filtration. The produced output which characterizes the topological features in the system can disclose the underlying properties of the system.²⁷

A simple example of the process in forming simplices during persistent homology is given in Figure 2-2. At specific radius values, various simplices were generated. For example, at radius = d_2 , no simplices were formed as none of the spheres overlapped with each other. However, at radius = d_3 , neighboring spheres began to overlap and as a result, lower dimensional simplices started to form. Due to the differences in bond lengths between some of the atoms, 1-simplices were formed only between atoms that are closer in distance. This shows how slight differences in molecular structures can be detected through the process of filtration.³² In addition, 2-simplices were also formed in positions where three spheres overlapped with one another. As the radius increased further to d_4 , more overlapping between spheres occur, thus higher dimensional simplices such as 3-simplex began to form.

Sphere's radius	d_1	d_2	d_3	d_4
Spherical model
Simplicial complex

Figure 2-2: Simplified example of a filtration process.

Biomolecules are typically large and complicated, therefore yielding numerous persistence bars to account for all of the simplices present. Hence, a collection of the individual persistence bars of all topological invariants in a topological space is usually organized together and studied in a persistence barcode.^{27,31} The compiled persistence barcode is also labelled as the molecular topological fingerprint (MTF), which is unique to every biomolecule. Contrary to popular belief, each bar, regardless of its lifetime, represents an important individual geometric attribute within the molecule.²⁵ Figure 2-3 presents a simple example of the MTF for a highly symmetrical molecule, C_{70} fullerene.

Figure 2-3: Example of MTF analysis of C_{70} fullerene. (Left) Three-dimensional model of C_{70} fullerene. (Right) MTF of C_{70} fullerene. The MTF contains three sections, namely Betti-0, Betti-1 and Betti-2 for the top, middle and bottom panels respectively. 70 bars could be observed in Betti-0 panel. This corresponds to the 70 C atoms which make up the C_{70} fullerene structure. As filtration continues, Betti-1 bars appear due to the emergence of rings simplices. C_{70} fullerene contains a mixture of pentagonal and hexagonal rings in the ratio of 12:25. This coincides with the number of Betti-1 bars that emerged, whereby 36 bars were accounted for. (The bar for the unaccounted ring was linearly combined with other rings) Eventually, Betti-2 bars begin to materialize as well, comprising a total of 25 bars. These bars represent the 25 hexagonal rings which persist and alter into two dimensional holes. Among these bars exist a persisting bar, which coincides to the central void in the C_{70} fullerene structure. Figure is adapted from Reference 27.

In recent years, persistent homology has been implemented across various fields as the method for data analysis dealing with large size and high dimensions.^{27,33} However, while persistent homology had its successes in analyzing biomolecular systems, there was still limitations associated with the method. Biomolecules are complicated both biologically and geometrically, and both components play an important role to its functionality. The biological function and structure of biomolecules are strongly dependent on its biological composition. Hence, the fact that persistent homology was formulated in a way which only simplifies the geometric aspect of the molecules prevents accurate characterization of the subject.^{26,30} In addition, when persistent homology is applied to a biomolecule, all of its atoms undergoes the process at the same time, which only generates a general geometric view of the biomolecule at various scale.²²

Therefore, in 2017, a variation was introduced in the form of ESPH.²⁵ ESPH was designed to recognize and simplify both the geometric and biological aspects of the biomolecular system. More importantly, even though the biological aspects are being simplified, important biological and chemical data are preserved.³⁰ Unlike persistent homology, ESPH was not only able to obtain a general view, but also able to uncover specific chemical statistics including hydrophobic and hydrophilic interactions. This was achieved by predetermining specific selection of heavy elements present in the biomolecular system and subjecting these atoms to its algorithm. For example, specifying an all C atoms choice may reveal the hydrophobicity information within, while an all nitrogen or oxygen choice may disclose hydrophilicity characteristics of the biomolecule. ESPH integrates the specific selection of element types along with an all atoms approach to achieve a comprehensive outlook on the function and structure relationship in the biomolecular system.²² In fact, in 2017 and 2018, ESPH was implemented to observe the correlation between the topological features of a biomolecular system and its functions, specifically in a protein-ligand system. These experiments were hugely successful as ESPH was discovered as having the capability to detect unseen topological-function relationships and being a strong representation for the protein-ligand system.^{25,30} ESPH also proved superior over other existing techniques, including conventional topology and persistent homology.

In this experiment, two separate ESPH models were developed, namely Ligand-based and Complex-based models for G-Quadruplex-ligand binding systems. The main purpose of these two models are to observe the representability of G-Quadruplex-ligand binding systems using a topological approach. Firstly, a Ligand-based model is built to study how well ESPH can predict the binding strength of the ligand to G-Quadruplex structures based only on the topological information extracted from the ligands alone. Following which, a Complex-based model will be developed to explore the binding affinity prediction of the topological methodology when the system is input with complex structure topological fingerprints.

2.1.1 Ligand-based model

In this segment, a model was developed to provide a topological representation of the G-Quadruplex ligands using ESPH.

To build an effective model, there must be sufficient good quality data set and information. In fact, deep learning ML techniques thrive more with increasing data scale.²¹ This proposed ESPH topological model was therefore built upon data obtained from G-Quadruplex Ligands Database (G4LDB), which is essentially an extensive database that collates all of the published G-Quadruplex ligands by various research groups. At the point of this report, G4LDB consist more than 900 different G-Quadruplex ligands and is the largest G-Quadruplex database known.¹⁸ Upon further examination of each ligand listed in G4LDB, only 139 ligands were reported with an experimental telomerase half maximal inhibitory concentration (IC_{50}) value, the concentration of the ligand at which telomerase activity was inhibited by half. The other ligands were not reported with an experimental IC_{50} value as they were conducted for various other purposes, and hence other experimental values such as melting temperature were attached instead, which is irrelevant to our experiment. Therefore, the data set for the Ligand-based model consist only these 139 reported G-Quadruplex inhibiting ligands with their respective published experimental values.³⁴⁻⁵² An assumption taken for all models developed in this research is that the decrease in telomerase activity is a result of a binding interaction between the ligand and the telomeric G-Quadruplex. The entire list of ligands utilized in the data set is attached to the report as Appendix A.

To perform ESPH on the data set, topological features must first be extracted. The first step involved the extraction of element types and the xyz coordinates of all heavy atoms present in the ligands. Only the heavy element atom types present in the ligands were considered. Specifically, the elements were of seven different atom types, Carbon (C), Nitrogen (N), Oxygen (O), Sulfur (S), Fluorine (F), Chlorine (Cl) and Iodine (I). By differentiating the different element types, important biological materials can be retained without extensive biomolecular convolution.³⁰ A variation was also conducted by grouping the

seven element types into four subgroups based on their similarities. C and N formed one subgroup each, Group VI elements O and S formed a third subgroup while the Group VII halogens made up the final subgroup. This distinction was made to observe if there were any stark differences in simplifying the model. The data set was then subjected to filtration to generate their respective topological invariants as barcodes. The software package employed in this process is Dionysus.

For effective visualization and representation of these barcodes by means of ML models, these barcodes were reorganized into feature arrays. The python package utilized in this process is NumPy. This reorganization technique employed is known as count in bins. For a particular set of atoms in a ligand, the barcode generated is labelled as M and each individual bar in M can be represented by a lifetime interval where $I_x = [b_x, d_x]$, in which b_x and d_x refers to the birth and death of the particular topological invariant. As such, the persistence p of an individual bar may be represented by the difference between b_x and d_x , or $p = (d_x - b_x)$. All of the barcodes can be separated into a pre-established assembly of bins, and the features of the ligands can be calculated based on the quantity of b , d and p within each bin. This give rise to three feature vectors, F_b^C , F_d^C and F_p^C for the count of birth, death and persistence respectively. The above process was repeated for all three topological dimensions, namely the 0th, 1st and 2nd. Overall, nine two-dimensional topological representations were generated for a Ligand-based model and considered as the nine channels of a topological image. The nine representations were:

$$\{F_{b,x}^C(M^0), F_{d,x}^C(M^0), F_{p,x}^C(M^0), \\ F_{b,x}^C(M^1), F_{d,x}^C(M^1), F_{p,x}^C(M^1), \\ F_{b,x}^C(M^2), F_{d,x}^C(M^2), F_{p,x}^C(M^2)\}$$

For the Ligand-based model, an interval of $[0, 4]\text{\AA}$ was considered and divided into bins of equal length using a standard bin size of $0.1\text{\AA}$. Hence, a $7 \times 9 \times 40$ and a $4 \times 9 \times 40$ tensor were used as the sample input feature for the models of seven and four subgroups respectively. Upon the computation of these topological

images, these features will then act as the input for ML algorithms as training and prediction.

The above explanation is also further illustrated in the example below, in which an amidoanthracene-9,10-dione $C_{34}H_{46}N_4O_6$ was subjected to ESPH using the Ligand-based model. The atoms in the ligand were extracted based on their element type, namely C, N and O, and subjected to persistent homology to generate their persistence barcodes. Following which, the barcodes were reorganized into feature arrays as mentioned.

Figure 2-4: Molecular structure of amidoanthracene-9,10-dione $C_{34}H_{46}N_4O_6$

Figure 2-5: ESPH persistent barcodes generated using C atoms. The barcodes are sort according to the d_x of each bar. The x-axis represents the square of the b_x and dx value, while the y-axis represents the quantity of bars. (Top left) Betti-0 persistent barcode. (Top Right) Betti-1 persistent barcode. (Bottom) Betti-3 persistent barcode.

Figure 2-6: ESPH persistent barcode generated using N atoms. The barcode is sort according to the b_x of each bar. The x-axis represents the square of the b_x and dx value, while the y-axis represents the quantity of bars. Betti-0 persistent barcode.

The barcodes were then reorganized into a 7x9x40 tensor as shown below.

For this experiment, two different ML algorithms, namely gradient boosting tree (GBT) regression and random forest (RF) regression were implemented for the Ligand-based model. Both GBT and RF are ensemble learning methodologies but differ in the manner that their decision trees are built. GBT produces its model stage by stage, decision tree by decision tree, thereby allowing for optimization in the subsequent decision tree based upon the previous. RF on the other hand, develops a large number of decision tree at a time, and trained each tree independently. Both GBT and RF are among the best performing algorithms available and they have been employed in many different fields.⁵³⁻⁵⁴ Their performance typically depends on the type of data set involved.⁵⁵ Both GBT and RF were imported from the software program Scikit-learn. The parameters for GBT was set as {'n_estimators': 20000, 'max_depth': 8,

'min_samples_split': 2, 'learning_rate': 0.005, 'loss': 'ls', 'max_features': 'sqrt', 'subsample': 0.7} while the parameters for RF was set as {max_features=32, n_estimators = 500, random_state=0}.

The features obtained from the data set were randomly split into train and test sets in a 1:9 ratio, or 10-fold cross validation in order to provide a more accurate result. However, computational and ML algorithms are typically random and tends to fluctuate even if the same calculations were performed using the exact same parameters. Thus, to overcome this randomness and to minimize the biasness in the results, 10-fold cross validation was performed ten times for each model. The Pearson correlation coefficients, along with its associated root mean square deviation (RMSE), between the topological features and the experimental values of the ligands were computed by the Ligand-based model. The average correlation scores were reported along with their respective RMSE for all of the various modifications of the Ligand-based model in Table 2-1. The complete results for all of the computational runs in the Ligand-based model are attached in Appendix B.

2.1.2 Complex-based model

In this segment, a model was developed to provide a topological representation of the complexes formed between G-Quadruplex and its binding ligands.

Similar to the Ligand-based model, to build a Complex-based model, a good quality data set must exist. Unlike the Ligand-based model which uses the ligands structural information, the Complex-based model requires the complex structural information as its input data. While the ligands structural information is readily available in databases, the complex information on the other hand is not. Thus, the first step to the Complex-based model is the identification of the G-Quadruplex structures required for the generation of the complex conformation between the G-Quadruplex and the ligands.

As mentioned in the introduction of G-Quadruplex structures, G-Quadruplex structures are highly polymorphic and thus may exist in many different forms

and variations. For the scope of this experiment, G-Quadruplex structures were selected based on their strand orientations. Hence, three different examples of human telomeric G-Quadruplex were chosen to represent each of the categories namely 2MB3, 2MCC and 4G0F, which are hybrid, anti-parallel and parallel G-Quadruplex structures respectively. Structural files of these G-Quadruplex were obtained from solution NMR and X-Ray diffraction which had been deposited in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB).⁵⁶⁻⁵⁹ The need to select multiple G-Quadruplex structures arises from the fact that human cells may contain different types of telomeric G-Quadruplex structures, and that telomeric G-Quadruplex structures may adopt different forms under various conditions.⁶⁰⁻⁶¹ Furthermore, some ligands may show selectivity towards a particular form of G-Quadruplex structure over another, while other ligands may bind to all forms.

Figure 2-8: Selected G-Quadruplex structures as visualized by AutoDock Vina. (Top left) hybrid 2MB3. (Top right) anti-parallel 2MCC. (Bottom) parallel 4G0F.

All 139 ligands were subjected to molecular docking simulations using AutoDock Vina with each of the G-Quadruplex structures selected to generate a complex conformation for every scenario. This meant that a total of 417 simulations run were conducted. Molecular docking refers to the computational technique which projects the most stable conformation when two molecules interacts with each other, usually between a receptor and a ligand. Docking programs are known for their ability to accurately identifying the appropriate binding conformations.⁶²⁻⁶³ AutoDock Vina, developed by The Scripps Research Institute, is an example of a docking software and ranks among one of the most accurate and most cited option available.⁶⁴⁻⁶⁵ AutoDock Vina typically produce two output, a binding affinity between the two molecules based on the scoring function adopted by the software, and a set of binding conformations of the ligand with respect to the receptor, as seen in Figure 2-5.

```
WARNING: The search space volume > 27000 Angstrom^3 (See FAQ)
Output will be G4L0021_out.pdbqt
Detected 8 CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: -1008241712
Performing search ...
0% 10 20 30 40 50 60 70 80 90 100%
|----|----|----|----|----|----|----|----|----|----|
*****
done.
Refining results ... done.

mode |  affinity | dist from best mode
    | (kcal/mol) | rmsd l.b. | rmsd u.b.
-----+-----+-----+-----
  1 |   -10.0 |    0.000 |    0.000
  2 |    -9.3 |    5.279 |   13.303
  3 |    -9.1 |    4.907 |   13.466
  4 |    -9.0 |    4.541 |   13.713
  5 |    -8.9 |    6.296 |   13.622
  6 |    -8.8 |    6.526 |   13.688
  7 |    -8.7 |    4.460 |   13.237
  8 |    -8.7 |    5.044 |    7.893
  9 |    -8.7 |    5.947 |   13.778
Writing output ... done.
```

Figure 2-9: Screenshot of a completed AutoDock Vina docking simulation. Typically, a total of nine binding affinities with their associated conformations are generated each time.

Therefore, the Complex-based model was developed using conformations obtained from AutoDock Vina. Each docking simulation run on AutoDock Vina generally produces a list of nine best conformations ranked according to their stability as determined by the program. However, since only one conformation was required for each G-Quadruplex ligand pair in the development of the Complex-based model, two selection criteria were established as follow:

1. The conformation with the strongest binding affinity was selected if it stacks in a parallel fashion on top of the G-tetrad layers.

Otherwise;

2. The next most stable conformation in the ranking list that interacts in a parallel manner was selected.

The above guiding rules were implemented based on the assumption that ligands bind to G-Quadruplex structures through a π -stacking interaction. This is a fair assumption given that all 139 ligands were observed to contain a π -conjugated system, and a π -stacking interaction between the ligand and the G-tetrad layers would theoretically yield the most stable conformation. Furthermore, *in vitro* experiments have shown that G-Quadruplex stabilizers largely interact through π -stacking interaction, while other modes of binding are rarely observed experimentally, as mentioned in Chapter 1.1.2. Figure 2-6 illustrates two published examples of ligands, a strong G-Quadruplex stabilizer (Figure 2-6i) and a weak G-Quadruplex stabilizer (Figure 2-6ii), having π -stacking interactions on top of G-tetrad layers.

Figure 2-10: π -stacking interaction between G-Quadruplex and stabilizers. (Left) Top view of bis-dimethylaminoethyl quindoline stacking on G-Quadruplex. (Right) Top view of 1,4 isomer of 1-piperidinyl anthracene-9,10-dione stacking on G-Quadruplex. Figure is adapted from Reference 34 and 39 respectively.

To perform ESPH on the data set, topological features must first be extracted. The first step involved the extraction of element types and the xyz coordinates of all heavy atoms present in both the G-Quadruplex structures and the ligands. Non-heavy atoms such as hydrogen atoms are excluded from ESPH, as the process of attaching missing hydrogen atoms to the receptor is dependent on the force field employed by the program and doing so would affect the accuracy of the methodology. Persistent homology was executed twice for the complex-model, once on all heavy atoms in both receptor and ligand, and once more on all C atoms in both receptor and ligand only. Generating the topological fingerprint of all C atoms only is important in representing the hydrophobic interactions between G-Quadruplex and the ligand.

In the selected G-Quadruplex structures, the heavy element types included C, N, O and Phosphorus (P) while in the ligands, the element types included C, N, O, S, F, Cl and I. In addition, to describe the interactions between G-Quadruplex and ligands; and to study the changes in topological invariants during the interactions, Interactive Persistent Homology (IPH) was introduced. IPH involves a process which pairs up two element types, with one element coming from the receptor and the ligand each, within a predefined cut off distance. To achieve this, the following distance matrix, $\hat{M}$, was utilized.

$$\hat{M} = \begin{cases} d_{ij}, & \text{if } a_i \in A_1, a_j \in A_2 \text{ or } a_i \in A_2, a_j \in A_1, \\ d_{\infty}, & \text{otherwise,} \end{cases}$$

Where d_{ij} refers to the Euclidean distance two atoms. A_1 and A_2 refer to the set of atoms from the G-Quadruplex and ligand respectively while a_i and a_j refers to the atoms i and j . d_{∞} refers to an infinitely large number that is set, when the atoms a_i and a_j are from the same set of atoms, A_1 or A_2 .

IPH is only interested in studying the interactions between atoms from the receptor and the ligand. As such, d_{∞} would ensure that the value goes beyond the predefined filtration limit. The features generated from IPH account for a total of 28 atom pairs ($\{C,N,O,P\}$ from G-Quadruplex and $\{C,N,O,S,F,Cl,I\}$ from the ligands), over a filtration interval of $[0, 50]\text{\AA}$ which was divided into

equal bin sizes of 0.25Å in length. Hence this feature will have a 200x28 tensor size.

Following which, the barcodes generated from filtration were subjected to a similar bin counting technique in Ligand-based model to reorganize the results into feature arrays. However, for the Complex-based model, two-dimensional topological features must be computed for the barcodes of both the complex structure as well as the differences between the barcodes of the complex and the G-Quadruplex structure itself. As a result, for both all heavy atoms and all C atoms, a total of eighteen two-dimensional topological representations were generated each and considered as the eighteen channels of a topological image which have a 200x18 tensor size. These features were then fed into the ML algorithms employed for both training and prediction purposes.

For the Complex-based model, the features were subjected to two ML algorithms, namely GBT and RF. Both GBT and RF were imported from the software program Scikit-learn and the parameters for GBT was set as {'n_estimators': 20000, 'max_depth': 8, 'min_samples_split': 2, 'learning_rate': 0.005, 'loss': 'ls', 'max_features': 'sqrt', 'subsample': 0.7} while the parameters for RF was set as {max_features=32, n_estimators = 500, random_state=0}. The features were then randomly split using a 10-fold cross validation into training and testing sets to minimize the risk of overfitting. Each model was computed ten times and the average Pearson correlation coefficients with their respective associated RMSE were reported in Table 2-3. The complete results for all of the computational runs in the Complex-based model are attached in Appendix C.

2.1.3 Results

Table 2-1: Comparison between various modifications of the Ligand-based model.

Model	Average score	Average RMSE	Average time/run
Gradient Boosting Tree (GBT)			
7-subgroups	0.7871	0.4621	390s
4-subgroups	0.7738	0.4648	285s
Random Forest (RF)			
7-subgroups	0.7662	0.4851	15s
4-subgroups	0.7614	0.4866	14s

Table 2-1 illustrates a comparison between the average Pearson correlation coefficient across four variations of Ligand-based model developed. Across all four variations, a highly positive linear correlation score could be observed, with the average score in the range of 0.76 to 0.79. The high correlation signifies a strong relationship between the topological invariants of the ligands and the observed experimental IC_{50} values. The RMSE associated with all of the correlation scores hover between 0.46 to 0.49, which is less than 10% of the median of the data set, therefore showing a good concentration of the data around the best fit line.

In addition, it was observed that the use of GBT as the ML algorithm choice yielded better performances than its counterpart. However, the computational time taken to complete a single run was significantly different between the two algorithms. On average, models employed with GBT algorithm required approximately 20 to 26 times more time than models employed with RF algorithm. Given that both algorithms gave reasonably comparable scores, such a huge time difference could possibly swing the algorithm preference towards RF, especially when the scale of the data set increases in the future.

Interestingly, the variation in the Ligand-based model based on the number of subgroups configured only showed minor differences between the scores. Intuitively, the more simplified the model or the lesser number of subgroups configured, the less element specific the model would be and thus the poorer

the representability of the model. However, the eventual minor differences in the results could be attributed to the small contrast between the two number of subgroups selected for the experiment. A larger data set size that contains a more extensive variety of different atom types should give a more distinctive result. It was also observed that 7-subgroups Ligand-based models were generally more expensive, due to the larger set of features to be generated. This difference was much more significant when GBT was employed and the difference is likely increase further with increasing specificity of the elements in the models.

Table 2-2: Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the Ligand-based model.

Model	Within a single 10-fold cross validation		Across all cycles of 10-fold cross validation	
	Average SD	Average CV	Average SD	Average CV
Gradient Boosting Tree (GBT)				
7-subgroups	0.1172	14.9%	0.0223	2.8%
4-subgroups	0.1694	21.9%	0.0363	4.7%
Random Forest (RF)				
7-subgroups	0.1646	21.5%	0.0379	4.9%
4-subgroups	0.1689	22.2%	0.0328	4.3%

Table 2-2 presents the average standard deviations (SD) and coefficient of variation (CV) associated to the four variations of the Ligand-based models. Two types of SD and CV were provided: The average values associated to the scores within a single 10-fold cross validation, and the average values associated to the scores across all ten cycles of 10-fold cross validation.

In general, within a single 10-fold cross validation conducted, the CV is relatively high across most models, with the exception of the 7-subgroups Ligand-based model employed with GBT. A high CV represents a wide spread of the results in the single computational run, which may also signal a possibility of the mean correlation score, as presented in Table 2-1, of being unfairly bumped up to a higher value by a small number of high scoring data, thereby

risking the integrity of the results. Upon a deeper analysis into the full set of results as presented in Appendix B and Appendix C, it was observed that within each 10-fold cross validation performed, there tend to be a range of 0-2 folds that performed poorer than the majority, which had scores in the region of 0.8000. This meant that the results presented in the above tables remained valid, since the high CV appeared to be contributed by a few low scoring data instead. This is theorized to be associated with the quality of the data set. The data set of 139 ligands were conducted by many different research groups. As such, any dissimilar experimental procedures or conditions undertaken by the various research groups may slightly affect the consistency of the experimental results across the ligands. These results, which are the input data in creating the ML models, may therefore in turn cause some inconsistency in the eventual computational models generated.

However, there was an anomaly in which the 7-subgroups Ligand-based model employed with GBT displayed significantly lower SD and CV within a single run. Since the models utilized the same input data set, this difference could likely be due to the combination effect of a larger sub-groups along with the robust GBT algorithm.

Across all ten cycles of 10-fold cross validation runs conducted, the average SD and CV calculated were low for all variations of the Ligand-based model. In general, all four models displayed an average CV of below 5.0%. This small spread of the results meant that the models were consistent in their predictions for the structure-function relationship of the biomolecular system.

Table 2-3: Comparison between various modifications of the Complex-based model.

Model	Average score	Average RMSE	Average time/run
Gradient Boosting Tree (GBT)			
2MB3	0.7420	0.5248	137s
2MCC	0.7611	0.5041	145s
4G0F	0.7115	0.5463	146s
Overall	0.7382	0.5251	143s
Random Forest (RF)			
2MB3	0.6719	0.5755	14s
2MCC	0.6914	0.5700	14s
4G0F	0.6676	0.5979	14s
Overall	0.6770	0.5811	14s

Table 2-3 illustrates a comparison between the average Pearson correlation coefficient scores of the individual Complex-based models as well as the overall Complex-based models developed. In general, all variations of the Complex-based model yielded a relatively strong structure-function correlation score. However, between the two algorithms employed, models that adopted GBT consistently outperformed their counterparts which adopted RF. This was opposed to the results in the Ligand-based models, whereby both algorithms yielded more comparable results, though it is not an uncommon sight that different ML algorithm performed to varied successes on different models. An overall Complex-based model score was computed based on the average of the three individual models, comprising of parallel, anti-parallel and hybrid G-Quadruplex structures, for each algorithm. Naturally, since models that adopted GBT yielded a higher score than the RF models, the overall Complex-based model employed with GBT generated a stronger positive correlation coefficient of 0.7382 and an associated RMSE of 0.5251 while the overall Complex-based model which employed RF fared poorer, with a positive correlation coefficient of 0.6770 and RMSE of 0.5811. This difference in performance came at the expense of computational cost, as the employment of GBT increased the time required by approximately 10 folds.

Between the individual G-Quadruplex models, models involving the anti-parallel 2MCC showed the strongest correlation in its structure-function relationship, while models involving parallel 4G0F represented the weakest correlation, regardless of the ML algorithm adopted. The difference in the scores could mean that more ligands in the data set have a higher propensity in binding towards anti-parallel G-Quadruplex structures, and comparatively lower propensity in binding towards the parallel G-Quadruplex structures. The individual models also consumed very similar computational cost within their ML algorithm categories.

Table 2-4: Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the Complex-based model.

Model	Within a single 10-fold cross validation		Across all cycles of 10-fold cross validation	
	Average SD	Average CV	Average SD	Average CV
Gradient Boosting Tree (GBT)				
2MB3	0.1392	18.8%	0.0226	3.1%
2MCC	0.1631	21.4%	0.0247	3.3%
4G0F	0.1625	22.8%	0.0312	4.4%
Overall	0.1549	21.0%	0.0262	3.5%
Random Forest (RF)				
2MB3	0.2276	33.9%	0.0323	4.8%
2MCC	0.2051	29.7%	0.0375	5.4%
4G0F	0.2061	30.9%	0.0200	3.0%
Overall	0.2129	31.5%	0.0299	4.4%

Table 2-4 presents the average standard deviations (SD) and coefficient of variation (CV) associated to all variations of the Complex-based model. Two types of SD and CV were provided: The average values associated to the scores within a single 10-fold cross validation, and the average values associated to the scores across all ten cycles of 10-fold cross validation.

Within a single 10-fold cross validation conducted, the Complex-based models displayed high CV. This was consistent to the trend displayed by the Ligand-

based models, in which the high SD and CV values were a result of a small number of poor scoring folds within each single computational run. The high CV trend would likely to reduce if a consistent experimental data set could be obtained as the input data. In addition, the use of RF algorithm in the Complex-based models appeared to worsen the spread of results. This could mean that GBT is a more robust algorithm than RF.

Across the ten iterations of computational runs performed, the Complex-based models displayed precision with low average SD and CV. The low average CV, at less than 5.0%, meant that the models were consistent in their predictions for the structure-function relationship of the biomolecular system.

Table 2-5: Comparison between Ligand-based and Complex-based model.

Model	Average score	Average RMSE	Average time/run
Gradient Boosting Tree (GBT)			
Ligand-based 7-subgroups	0.7871	0.4621	390s
Ligand-based 4-subgroups	0.7738	0.4648	285s
Complex-based overall	0.7382	0.5251	143s
Random Forest (RF)			
Ligand-based 7-subgroups	0.7662	0.4851	15s
Ligand-based 4-subgroups	0.7614	0.4866	14s
Complex-based overall	0.6770	0.5811	14s

Table 2-5 compares the performances from both Ligand-based and Complex-based models. In general, Ligand-based models outperformed Complex-based models regardless of the ML algorithm adopted. In fact, when RF was employed, the difference between Ligand-based and Complex-based correlation score was significant at 0.0844 to 0.0892. Complex-based models expectedly performed poorer due to the additional steps taken in the preparation of the data set. Specifically, all 139 ligands and 3 G-Quadruplex structures in the data set underwent molecular simulation using AutoDock Vina to generate the complex structural files. Despite the program being among one of the most accurate

docking software, the use of AutoDock Vina along with the assumptions made in the selection of conformations would no doubt still introduce some additional errors, thus plaguing the building of the Complex-based model.⁶⁵ As such, the poorer performances of the Complex-based models are likely to be associated with the errors arising from the docking stage. Nevertheless, the Complex-based models still attained a respectable performance score in being a representation for the structure-function relationship of the biomolecular system.

Between the models, the amount of processing time required to complete each run differed only with the use of GBT algorithm. It was observed that Ligand-based models generally consumed about two to three folds the time required to complete a single Complex-based computation.

2.2 Rigidity Index-Score (RI-Score)

RI-Score is a ML technique that was developed to characterize a protein-ligand binding system. The basis of RI-Score lies in the concept that there is an increase in the rigidity (or decrease in the flexibility) of the protein when a ligand binds to it. Conventionally, it was believed that the flexibility of the receptor and ligands favors and enhances binding. In fact, the more flexible the molecules involved, the more favorable the binding process.

However, in 2017, Wei. G.W and team discovered that enhanced rigidity in protein which arises during the binding process not only plays a crucial role in a protein-ligand system but is also a mechanism for the binding process.²⁸ As such, a ML technique that was dedicated entirely to rigidity strengthening alone was developed. This meant that no other interaction forces, such as van der Waals forces or hydrogen bonding were included in RI-Score.

To achieve this model, RI-Score was performed using the Flexibility-Rigidity Index (FRI). FRI is an algorithm invented to examine the flexibility and rigidity of molecules at an atomic scale. FRI was constructed with the assumption that the functions of protein was completely influenced by the structure and environment of the molecule itself only. Therefore, this meant that FRI, just like

ESPH, also studies and compute the topological connectivity between atoms to obtain a representation of the structural geometric compactness.⁶⁶

In mathematical terms, atomic rigidity index of a molecule with N number of atoms whose coordinates are represented by $\{r_i | r_i \in \mathbb{R}^3\}$ in which $i = 1, 2, \dots, N$, can be expressed by the following equation:

$$\mu_i = \sum_{j=1}^N \Phi_{\tau}(\|r_i - r_j\|)$$

Where $\|r_i - r_j\|$ refers to the Euclidean distance between the coordinates of two atoms, i and j , and Φ is a real valued monotonic function while τ is a variable that takes value above 0.

$$\begin{aligned}\Phi(\|r_i - r_j\|) &= 1 \text{ as } \|r_i - r_j\| \rightarrow 0 \\ \Phi(\|r_i - r_j\|) &= 0 \text{ as } \|r_i - r_j\| \rightarrow \infty\end{aligned}$$

The correlation monotonic function typically utilizes two types of functions, namely exponential and Lorentz function, represented by E and L respectively. Both functions behave close to an ideal low-pass filter (LPF) when parameters κ and ν takes on large values close to infinity.

$$\begin{aligned}\Phi_{\kappa, \tau}^E(\|r_i - r_j\|) &= e^{-\left(\frac{\|r_i - r_j\|}{\eta_{ij}}\right)^{\kappa}}, \kappa > 0 \\ \Phi_{\nu, \tau}^L(\|r_i - r_j\|) &= \frac{1}{1 + \left(\frac{\|r_i - r_j\|}{\eta_{ij}}\right)^{\nu}}, \nu > 0\end{aligned}$$

FRI details the bound receptor-ligand complex in a high degree of abstraction, which favors the extraction of the inherent receptor-ligand interactions from the input data set. As such, RI-Score was especially successful as a representation for the protein-ligand biomolecular system. In fact, it was observed that RI-Score performed better than other scoring functions in comparison. An accurate and high performance also meant that rigidity strengthening in the receptor during binding is an important mechanism of the protein-ligand binding system.

Therefore, in this experiment, the idea is to adapt RI-Score towards the biomolecular system between G-Quadruplex and ligands. The use of RI-Score will be able to firstly determine whether rigidity strengthening is also a mechanism of the G-Quadruplex-ligand biomolecular system, and secondly, whether RI-Score would similarly be a strong representation of the aforementioned system. Since RI-Score was developed with all modes of interactions being removed and leaving behind solely the rigidity changes, a strong positive Pearson correlation score would satisfy the above two goals.

Hence, based on the above atomic rigidity index equation, an element specific G-Quadruplex-ligand rigidity index can be computed through the cross correlations given by the equation as follow:

$$RI_{\beta,\tau,c}^{\alpha}(X - Y) = \sum_{k \in X \in G4} \sum_{l \in Y \in Lig} \Phi_{\beta,\tau}^{\alpha}(|r_k - r_l|)$$

Where α refers to a kernel index that can take values of E or L, for exponential kernel and Lorentz kernel respectively.

β refers to a kernel order index that can take values:

$$\beta = \kappa \text{ when } \alpha = E$$

$$\beta = \nu \text{ when } \alpha = L$$

c refers to a predefined cut off distance for the reduction computer demands and complexity. X and Y refers to the atom type in the G-Quadruplex structure and the ligand respectively.

Exponential kernel is noted for its rapid decay, as opposed to Lorentz kernel which is better known for its slow decay. Thus, exponential kernel is often utilized to identify the effective length for interactions of various range while Lorentz kernel is employed to study interactions of different distances. For this experiment, κ takes on a large value of 15.0, in order to replicate the ideal LPF

behaviour and eliminate interactions exceeding the kernel dimensions. ν , on the other hand, was set at a value of 5.0. τ takes on three values of 1.5, 3 and 5.5.

Similar to Complex-based ESPH models, the data set of 139 complex structural files generated from AutoDock Vina was utilized. Only the heavy atom types present in the receptor and ligands were considered in RI-Score. Specifically, in the selected G-Quadruplex structures, the heavy element types include C, N, O and P while in the ligands, the element types include C, N, O, S, F, Cl and I.

Three sets of rigidity index feature vectors were generated through the model by each kernel, with each set constructed on a different τ value. In total, 84 descriptors made up this model, with each set contributing a total of 28 descriptors generated from the cross correlations of the four atom types in G-Quadruplex and the seven atom types in the ligands. This is summarized as follow by $RJ_{\beta_1, \tau_1, c_1; \beta_2, \tau_2, c_2; \beta_3, \tau_3, c_3}^{\alpha_1 \alpha_2 \alpha_3}$.

Following the construction of the features, they were subjected to the same ML algorithms, GBT and RF. Both GBT and RF were imported from the software program Scikit-learn and the parameters for GBT was set as {'n_estimators': 20000, 'max_depth': 8, 'min_samples_split': 2, 'learning_rate': 0.005, 'loss': 'ls', 'max_features': 'sqrt', 'subsample': 0.7} while the parameters for RF was set as {max_features=32, n_estimators = 500, random_state=0}. The features were randomly split using a 10-fold cross validation into training and testing sets to minimize the risk of overfitting. Each model was computed ten times and the average Pearson correlation coefficients with their respective associated RMSE were reported in Table 2-6 and Table 2-8. The complete results for all of the computational runs in the exponential kernel model and the Lorentz kernel model are attached in Appendix D and Appendix E respectively.

2.2.1 Results

Table 2-6: Comparison between various modifications of the exponential kernel model.

Model	Average score	Average RMSE	Average time/run
Gradient Boosting Tree (GBT)			
2MB3	0.7765	0.4659	94s
2MCC	0.7667	0.4717	86s
4G0F	0.6596	0.5737	147s
Overall	0.7343	0.5038	109s
Random Forest (RF)			
2MB3	0.7538	0.4840	15s
2MCC	0.7553	0.4881	15s
4G0F	0.6727	0.5391	14s
Overall	0.7273	0.5037	15s

Table 2-6 illustrates a comparison between the average Pearson correlation coefficient scores of the various models when an exponential kernel was employed. The table included the three individual G-Quadruplex models trained with GBT and RF as well as the overall model obtained by the average of the three individual models. In general, all variations of the exponential kernel model yielded a relatively strong structure-function correlation score. As a result, the overall exponential kernel models displayed strong positive correlation scores of 0.7343 (0.5038) and 0.7273 (0.5037) with the use of GBT and RF respectively. Similar to the ESPH models introduced earlier, the model that utilized GBT outperformed the model that utilized RF, at the expense of greater computational cost. Models took at least 6 times longer to develop a relation in the data set when employed with GBT. Since both algorithms generated very comparable correlation scores, the difference in the processing time may be a factor to consider for computational researchers, particularly as the scale of data set increases.

Between the individual G-Quadruplex models, models involving the hybrid 2MB3 and anti-parallel 2MCC produced strong correlation results in the range

of 0.75 to 0.78 with a low RMSE, regardless of the ML algorithm adopted. However, models involving the parallel 4G0F performed significantly poorer as compared to their counterparts. This difference could once again point to the same reasoning that the data set deployed contains more ligands that have a lower propensity towards binding with parallel G-Quadruplex structures. Comparisons made across corresponding models between the two ML algorithms showed minor differences, much like the results in the Ligand-based ESPH model. This was however not surprising since the use of RI-Score in protein-ligand biomolecular systems likewise produced comparable results with either GBT or RF.²⁸

With the exception of 4G0F model when GBT was employed, the individual models within each ML algorithm category typically took a similar average time to complete each run.

Table 2-7: Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the exponential kernel model.

Model	Within a single 10-fold cross validation		Across all cycles of 10-fold cross validation	
	Average SD	Average CV	Average SD	Average CV
Gradient Boosting Tree (GBT)				
2MB3	0.1471	18.9%	0.0266	3.4%
2MCC	0.1464	19.1%	0.0215	2.8%
4G0F	0.2118	32.1%	0.0291	4.4%
Overall	0.1684	22.9%	0.0257	3.5%
Random Forest (RF)				
2MB3	0.1702	22.6%	0.0359	4.8%
2MCC	0.1599	21.2%	0.0245	3.2%
4G0F	0.2028	30.1%	0.0472	7.0%
Overall	0.1776	24.4%	0.0359	4.9%

Table 2-7 presents the average standard deviations (SD) and coefficient of variation (CV) associated to all variations of the exponential kernel model. Two types of SD and CV were provided: The average values associated to the scores

within a single 10-fold cross validation, and the average values associated to the scores across all ten cycles of 10-fold cross validation.

Within a single 10-fold cross validation conducted, all exponential kernel models exhibited high CV, similar to the ESPH models developed. These high average CV values were likewise caused by a small number of low scoring folds within every individual iteration.

Across the ten repeats of computational runs performed, the exponential kernel models were consistent with their predictions. Hence, low average SD and CV values were obtained.

Table 2-8: Comparison between various modifications of the Lorentz kernel model.

Model	Average score	Average RMSE	Average time/run
Gradient Boosting Tree (GBT)			
2MB3	0.7530	0.4678	89s
2MCC	0.7703	0.4658	88s
4G0F	0.7546	0.4702	89s
Overall	0.7593	0.4679	89s
Random Forest (RF)			
2MB3	0.7390	0.4686	15s
2MCC	0.7739	0.4691	15s
4G0F	0.7675	0.4704	15s
Overall	0.7601	0.4694	15s

Table 2-8 illustrates a comparison between the average Pearson correlation coefficient scores of the various models when a Lorentz kernel was employed. The table included the three individual G-Quadruplex models trained with GBT and RF, along with the overall model derived from the average of the three individual models. In general, all variations of the Lorentz kernel model produced a strong structure-function correlation score. Therefore, this also translated to the high correlation scores displayed by both overall Lorentz

models, achieving an average Pearson correlation of 0.7593 (0.4679) and 0.7601 (0.4694) with the use of GBT and RF algorithms respectively.

Consistent to all of the models that had been developed prior to these, Lorentz kernel models employed with GBT consumed greater computational cost than their RF counterparts. On average, models utilizing GBT required approximately six times the average processing time than models utilizing RF. However, unlike the previously introduced models, RF algorithm appeared to outperform GBT algorithm despite the smaller cost consumed, even though the difference across corresponding models in the two algorithms were insignificant. The negligible differences in the scores between corresponding models from the two ML algorithms meant that the use of Lorentz kernel with RF is the more efficient and effective approach.

Across the individual G-Quadruplex models developed, models involving the anti-parallel 2MCC once again displayed the strongest correlation coefficient in characterizing the structure-function relationship, regardless of the ML algorithm adopted. This observation could mean that a larger percentage of the ligands in the input data set had a higher propensity towards interacting with anti-parallel G-Quadruplex structures which was consistent with the observations in the Complex-based models.

In addition, all of the individual models also consumed very similar computational cost within their ML algorithm categories.

Table 2-9: Average standard deviations (SD) and coefficient of variation (CV) within and between various modifications of the Lorentz kernel model.

Model	Within a single 10-fold cross validation		Across all cycles of 10-fold cross validation	
	Average SD	Average CV	Average SD	Average CV
Gradient Boosting Tree (GBT)				
2MB3	0.1771	23.5%	0.0411	5.5%
2MCC	0.1614	21.0%	0.0365	4.7%
4G0F	0.1931	25.6%	0.0266	3.5%
Overall	0.1772	23.3%	0.0348	4.6%
Random Forest (RF)				
2MB3	0.1918	26.0%	0.0520	7.0%
2MCC	0.1668	21.5%	0.0141	1.8%
4G0F	0.1658	21.6%	0.0305	4.0%
Overall	0.1748	23.0%	0.0322	4.2%

Table 2-9 presents the average standard deviations (SD) and coefficient of variation (CV) associated to all variations of the Lorentz kernel models. Two types of SD and CV were provided: The average values associated to the scores within a single 10-fold cross validation, and the average values associated to the scores across all ten cycles of 10-fold cross validation.

In a single 10-fold cross validation performed, the Lorentz kernel models exhibited high CV. Across all ten cycles of cross validations performed, the Lorentz kernel however exhibited low CV. These trends were observable in the large majority of the models developed.

In general, the overall high performances as presented in the above Table 2-8 and Table 2-9 showed that the phenomenon of rigidity strengthening occur in a G-Quadruplex-ligand system too and that it is a mechanism in the binding process. As a result, RI-Score presents a possible alternative model to characterize and represent such a biomolecular system.

Table 2-10: Ranking across all developed models.

Rank	Model	Average score (Average RMSE)
1	Ligand-based 7-subgroups (GBT)	0.7871 (0.4621)
2	Ligand-based 4-subgroups (GBT)	0.7738 (0.4648)
3	Ligand-based 7-subgroups (RF)	0.7662 (0.4851)
4	Ligand-based 4-subgroups (RF)	0.7614 (0.4866)
5	Lorentz RI-Score (RF)	0.7601 (0.4694)
6	Lorentz RI-Score (GBT)	0.7593 (0.4679)
7	Complex-based overall (GBT)	0.7382 (0.5251)
8	Exponential RI-Score (GBT)	0.7343 (0.5038)
9	Exponential RI-Score (RF)	0.7273 (0.5037)
10	Complex-based overall (RF)	0.6770 (0.5811)

Table 2-10 presents a ranking of the developed models based on their performances. Earlier, it was established that the Ligand-based models presented a stronger representation as compared to the Complex-based models. From the ranking above, it could be seen that RI-Score models yielded very comparable results to the ESPH models, with the RI-Score models generally a stronger representation than the Complex-based models but a weaker representation than the Ligand-based models. This observation could be attributed to the methodology used to set up the models. Both Complex-based and RI-Score models, unlike the Ligand-based models, utilized complex structures from AutoDock Vina as input files, which introduced additional errors as elaborated in Chapter 2.1.3.

Therefore, at present, among the ESPH and RI-Score models introduced, it appeared that the Ligand-based model presents the best mode of representation for a G-Quadruplex-ligand binding system. However, if a more robust technique for generating complex structures could be employed and a more standardized set of input data could be obtained, the Lorentz RI-Score models could present a potential alternative.

Furthermore, while GBT and RF algorithms are known to perform to varying success depending on the factors involved, the above table may suggest that GBT is the more superior algorithm for the models introduced. Most of the models employed with GBT algorithm consistently yielded better results than their RF counterparts, with the Lorentz kernel RI-Score models being the only exceptions. However, even then, the difference within the Lorentz kernel models was negligible at a mere 0.0008. This showed that GBT is the more robust ML algorithm between the two choices with respect to the data set and methodologies utilized. However, RF algorithm remained as an important and useful ML algorithm for the models. The better performances of GBT models came at the high expense of computational time. With a small data set of 139 ligands, GBT models were observed to consume between a range of 6 to 26 folds the time required by RF models. As research and interests in G-Quadruplex continue to progress in the near future, the available data for G-Quadruplex stabilizers would increase. This would lead to the increase in the input data set for the ML models and as such, the time taken for each model would likewise increase. The increase in computational cost could follow an exponential curve with the increase in the size of data set, and such a situation would present a huge dilemma between the usage of the two algorithms. Therefore, GBT and RF algorithms present two useful tools of different purposes for computational researchers, depending on their priority for greater accuracy or faster speed.

Chapter 3: Comparison with Existing Methodologies

In Chapter 2, several models based on ESPH and RI-Score were introduced to respectable performances on the given data set of 139 G-Quadruplex stabilizers. However, the high generated scores alone would not be meaningful in proving the potential of the methodologies and models introduced. Hence, to further validate the reliability of these models and to provide a benchmark for comparison, the same data set was subjected to an existing technique that has been widely used in modern day drug design processes. This technique is known as Quantitative Structure-Activity Relationship (QSAR), and the software employed in this experiment is known as CORrelations And Logic (CORAL).

QSAR is a computational technique that develops a mathematical model to obtain a quantitative structure-activity relationship for the dataset of input compounds. QSAR is commonly utilized for both regression and classification problems.⁶⁷ Similar to the basis of ESPH and RI-SCORE models introduced in the research, a QSAR model can be trained up by a data set containing of structural files and experimental values of the input compounds. QSAR is also a versatile technique, allowing for the use of various types of descriptors in the development of the model. This includes geometrical, constitutional and electrostatic amongst others.⁶⁸

An example of QSAR software is known as CORAL developed by the Mario Negri Institute based upon previous research projects known as CHEMPREDICT in 2008.⁶⁹ CORAL is a free standalone software that adopts the use of both Monte Carlo technique and Simplified Molecular-Input Line-Entry System (SMILES) to construct its models. The program is not only capable of handling classical chemical structures but is also adept at dealing with nanomaterials.⁷⁰ As such, CORAL is a well-known choice of QSAR software and has been tasked to resolve problems on various European Union (EU) projects.⁷¹⁻⁷³

Hence, CORAL was employed on the same data set of 139 ligands to produce a QSAR model for comparison. The descriptors for CORAL may be obtained

from both graph and SMILES notation. Therefore, the parameters configured for the extraction of molecular features were as follow: {Hydrogen suppressed graph (HSG), Extended Connectivity 0 (EC0), EC1, EC2} based on graph and {s, BOND, NOSP, HALO} based on SMILES notation. HSG represents vertices using chemical elements instead of electronic structures, and only considers non-hydrogen atoms for calculations. EC0, EC1 and EC2 represents the extended connectivity for the vertices in the graph. s is a local representation of the molecular fragment based on the SMILES notation while BOND, NOSP and HALO are global representations. BOND describes the presence or absence of multiple bonds, such as double, triple or even stereochemical bonds. NOSP refers to the presence or absences of four specific element types: N, O, S and P. HALO indicates the presence or absence of halogen elements, such as C and I.⁶⁹ The complete list of ligands in their SMILES notation is attached in Appendix F.

The features obtained from the data set were randomly split into train and test sets in a 1:9 ratio, or 10-fold cross validation in order to provide a more accurate result. 10-fold cross validation was performed for ten cycles in order to account for the randomness of CORAL. The Pearson correlation value was generated between the features and the experimental values of the ligands and compared to all of the other models in Figure 3-1 below. The complete results for the use of CORAL are attached in Appendix G.

Figure 3-1: Overview and comparison between new and existing methods.

The above bar graph plots the various models from the current research in comparison to the existing QSAR technique. CORAL performed reasonably well on the given data set, generating a Pearson correlation score of 0.7557, ranking 7th among the eleven models. In fact, CORAL outperformed both Complex-based models and both exponential kernel RI-Score models, while only differed from both Lorentz kernel RI-Score models within a difference of 0.0050. However, all four Ligand-based models notably remained as the top four ranking models and outshining the well-established CORAL.

The strong performance of CORAL did not come as a surprise given that QSAR techniques are widely used in drug design companies. However, the poorer performance of the Complex-based and RI-Score models could also be due to the use of complex structural files as generated by AutoDock Vina, thereby subjecting the models to increase errors. This was in opposed to CORAL, which made use of the SMILES notation of the G-Quadruplex stabilizers alone, and thus experienced no such errors. This also meant that the Ligand-based models would therefore be the closest form of comparisons to the CORAL QSAR

model, since both of them made use of only information extracted from the stabilizers. Hence, a more detailed comparison was made in the table below.

Table 3-1: Average standard deviations (SD) and coefficient of variation (CV) within and between Ligand-based ESPH and CORAL QSAR models.

Model	Within a single 10-fold cross validation		Across all cycles of 10-fold cross validation		Average time/run
	Average SD	Average CV	Average SD	Average CV	
7-subgroups (GBT)	0.1172	14.9%	0.0223	2.8%	390s
4-subgroups (GBT)	0.1694	21.9%	0.0363	4.7%	285s
7-subgroups (RF)	0.1646	21.5%	0.0379	4.9%	15s
4-subgroups (RF)	0.1689	22.2%	0.0328	4.3%	14s
CORAL QSAR	0.1115	14.8%	0.0197	2.6%	870s

From the above table, it can be observed that while CORAL QSAR yielded a lower Pearson correlation score than all of the Ligand-based models, it however exhibited a more cohesive result than most of them. In fact, a comparison across all factors would only place the Ligand-based 7-subgroups model employed with GBT firmly above the CORAL model. The Ligand-based 7-subgroups model utilizing GBT algorithm not only displayed a stronger Pearson correlation coefficient by a margin of 0.0314, but also achieved this at less than half the time required by CORAL. Both models also obtained similar associated RMSE and CV values. This is an interesting finding that showcased the strength of topological features as a strong descriptor for representing the G-Quadruplex-ligand biomolecular system. These comparisons also asserted the potential of the Ligand-based 7-subgroups (GBT) model, whose performance is likely to improve further with increasing data set made available as research in G-Quadruplex increases. Other models introduced here, despite not performing as

well as expected, displayed sufficient potential to be worth the time investment for further developments.

Chapter 4: Internal Validity Test

In Chapter 2, the models developed in the thesis have been shown to provide a strong representation score of the structure-function relationship of a G-Quadruplex-ligand system. In Chapter 3, several of the models were shown to have outperformed existing state of the art technique using the same input data set. In order to further prove that the high performing scores were not a result of coincidence and biasness, but instead a result of the underlying relationship between the topological features of the stabilizers and their activities, an internal test was conducted on the methodology with the highest potential, the Ligand-based model.

This internal test involved the addition of new ligand molecules into the data set but configured with wrong associated experimental values. A total of seventeen ligands were added, which represented approximately 10% the size of the original data set of 139 stabilizers.^{34-36,38,42} All of the seventeen ligands had been tested experimentally and proven to have high telomerase IC₅₀ values of over 50μM, even though the exact inhibitory values were not published. In other words, they are considered to be poor stabilizers of G-Quadruplex structures. For this internal test, two new data set each consisting of 156 ligand molecules were created. In the first data set, which would be labelled as Test set 1 thereafter, all of the seventeen ligands were attached with an experimental value of 0.018μM, which signifies that these ligands are strong stabilizers of G-Quadruplex. In the second data set, which would be labelled as Test set 2 thereafter, all of the seventeen ligands were attached with an experimental value of 50μM, the minimum value that these ligands were proven to showcase experimentally. The purpose of this internal test would be able to test for the ability of the topological models in identifying wrong input data. The list of the additional ligands utilized in the data set is attached to the report as Appendix H.

These two data set were then subjected as input data to the Ligand-based models in accordance to the methodology as presented in Chapter 2.1.1 and comparisons were made between the results exhibited by the Ligand-based models. The performances of the Ligand-based models were reported in Table 4-1. The complete results for all of the computational runs using Test set 1 and Test set 2 are attached in Appendix I and Appendix J respectively.

Table 4-1: Comparison between the Ligand-based models when subjected to different input data set.

Input data	Average score	% Change	Average RMSE	% Change
7-subgroups GBT				
Test set 1	0.7131	- 9.4%	0.7033	+52.2%
Test set 2	0.7701	-2.2%	0.4944	+7.0%
Original set*	0.7871	-	0.4621	-
4-subgroups GBT				
Test set 1	0.7021	-9.3%	0.7033	+51.3%
Test set 2	0.7659	-1.0%	0.4988	+7.3%
Original set*	0.7738	-	0.4648	-
7-subgroups RF				
Test set 1	0.6739	-12.0%	0.7565	+55.9%
Test set 2	0.7473	-2.5%	0.5377	+10.8%
Original set*	0.7662	-	0.4851	-
4-subgroups RF				
Test set 1	0.6889	-9.5%	0.7297	+50.0%
Test set 2	0.7350	-3.5%	0.5414	+11.3%
Original set*	0.7614	-	0.4866	-

*Original set refers to the data set of 139 ligands without any error input on purpose, as described in Chapter 2.1.1. The results act as a benchmark.

Table 4-1 illustrates a comparison between the average Pearson correlation coefficient of the Ligand-based model when input with varying data set. Across all of the four variations of the Ligand-based model, all of them displayed the same trend, in which the lowest Pearson correlation and highest associated

RMSE was observed when subjected with Test set 1 while the highest Pearson correlation and lowest associated RMSE was observed when subjected with the Original set. Test set 1 contained the largest amount of errors, in which all seventeen new ligands were configured to be extremely good stabilizers with low telomerase IC₅₀ values, a far stretch from their actual activity. As such, the Ligand-based models suffered a major drop in the correlation performances as compared to the benchmark, since 11% of the data set was laden with massive errors. This was because the Ligand-based models were unable to find the correlation between the topological features and the activity of these wrongly input data. The Ligand-based models suffered on average 10.1% decrease in the Pearson correlation coefficient, while experiencing a massive increase of 52.4% on average in their RMSE. As the wrong data accounted for only 11% of the entire data set, the Ligand-based model still achieved a reasonable performance despite the major decrease.

Test set 2 contained much lower amount of errors than Test set 1, as all seventeen new ligands were configured to be poor stabilizers at high telomerase IC₅₀ values of 50 μ M. The actual activities of these ligands were not known and merely published as “above 50 μ M”, this meant that a configured value of 50 μ M could be either close or far from their actual performances on G-Quadruplex structures. Regardless, Test set 2 would represent a much more accurate input data set than Test set 1, despite not being entirely accurate. As such, the Ligand-based models still suffered a decrease in their performances, albeit not as major as seen when Test set 1 was utilized. The Ligand-based models suffered on average 2.3% decrease in the Pearson correlation coefficient, while experiencing a large increase of 9.1% increase in their associated RMSE. This meant that the Ligand-based models were sensitive enough to identify the presence of an error in the data set.

Hence, the internal test conducted not only proved the existence of a relationship between the topological features of the G-Quadruplex stabilizers and their G-Quadruplex stabilizing functions, but also displayed the sensitivity of the Ligand-based models.

Chapter 5: Limitations

While the models developed in this thesis had their advantages and strengths, they are still far from being the ideal complete model in representing the G-Quadruplex-ligand biomolecular systems. The limitations could be split into two categories: Limitations that affects all models, and limitations that affect the Complex-based and RI-Score models.

5.1 All Models

One major limitation present in all of the models presented is the lack of input data. The field of G-Quadruplex is a much lesser known and studied field, and thus less research has been done on the biomolecular system, as mentioned in Chapter 1.2. This could be seen from the small number of G-Quadruplex stabilizers available in the database. All of the models introduced in the research were performed on a data set of 139 ligands, which is considerably small, given that ML data scale tends to be in the range of thousands. As such, these models may not be viewed as comprehensive nor representative of G-Quadruplex ligands in general. Furthermore, a small data set runs a high risk of generating a model that is overfitted which may threaten the validity of the results. However, measures have been taken to avoid the issue of overfitting, by training the models with 10-fold cross validation and the use of GBT algorithm, which was found to be robust against problems of overfitting.²² Given the limited information available, the results presented were a relative success and these models showed the existence of a relationship between the topological features and functions.

Another limitation associated to the input data is the varying conditions utilized by the research groups to measure the experimental telomerase IC₅₀ values. The different conditions may affect the outcome of the experimental values, and in turn affecting the consistency of the input data for the models, and thereby resulting in the models being a poorer representation of the biomolecular system. However, if a standardized data set could be made available, the ML models developed in this thesis would likely to perform even better.

Finally, the models developed thus far are still in the early stages and are thereby too premature to be utilized in the real world. While the models have been proven to be capable of characterizing the G-Quadruplex-ligand system, outperforming existing technique and being sensitive to input data errors, the models are still too raw for usage. At the moment, the models are unable to provide clear explanations and insights into this correlation. Furthermore, despite the models being able to sense the presence of errors within the data set, they are, however, unable to identify the exact ligands of errors.

5.2 Complex-based and RI-Score Models

In addition to the limitations listed in Chapter 5.1, Complex-based and RI-Score models suffer from further limitations mainly due to the molecular simulation steps incorporated.

Firstly, in the AutoDock Vina simulations conducted, the computational environment was assumed to be an ideal environment. This meant that within each docking simulation, a single G-Quadruplex structure was docked with a single ligand with no presence of any interfering species. In reality, telomeric G-Quadruplex structures exist in cells, which consist of many other substances that may either enhance or inhibit the binding of the ligand to the various G-Quadruplex structures. As such, the simulation results may be skewed towards stronger binding scores. However, given that only the simulated binding conformations were utilized in the development of the models, and that these ligands were proven experimentally to bind with G-Quadruplex structures, the models may therefore only be affected slightly, thereby justifying their relatively good performances.

Secondly, the performances of Complex-based and RI-Score models may also be affected by the AutoDock Vina selection rules in choosing the complex conformational input file. In the selection rules utilized for the experiments conducted in the thesis, conformations were extracted based on two factors: The presence of a π -stacking interaction and the ranking provided by AutoDock

Vina. The former is a fair assumption given the presence of conjugated aromatic rings in all of the ligands involved and that π -stacking interaction appears to be the most established mode of interaction within a G-Quadruplex-ligand system. However, the latter rule is subjected to the scoring function adopted by the docking software chosen, therefore the top ranked conformation determined by the program may not be the ‘true’ conformation.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

Cancer is a class of life-threatening diseases that have plagued the society for centuries, with high number of deaths occurring annually. According to the World Health Organization, in the year 2018 alone, there were more than 18.1 million new cases of cancer patients and over 9.6 million deaths as a result of cancer.⁷⁴ The immortality of the cancer cells has played no small part in contributing to this menace, with the high activity of telomerase at the heart of it. As such, the discovery that human chromosomal telomeres could fold into G-Quadruplex structures and in turn inhibit the activity of telomerase was a huge revelation within the field. In some cases, the inhibition was documented to have caused the apoptosis of cancer cells. Therefore, over the years, much research has been performed on the development of G-Quadruplex stabilizers in the effort to work towards overcoming cancerous cells.

However, progress in the research on G-Quadruplex stabilizers have generally been slow, with only over 900 different ligands being tested and recorded in G4LDB, the largest G-Quadruplex database known. This could be due to various reasons such as relatively new research field, or the lack of an efficient ligand design methodology.

In the past few years, much success has been found from the use of ML techniques incorporated with the extraction of topological features in the protein-ligand biomolecular system. The methodologies have been observed to outperform all other commonly utilized and well-established features. Hence, this research focuses mainly on the adaptation of similar methodologies for the G-Quadruplex-ligand biomolecular system and subsequently develop ML models that provide a strong representation and characterization of the system. Achieving these would pave the path towards the long-term goal for a more efficient and effective G-Quadruplex stabilizer design technique. Two main ML methodologies were explored in this research, namely ESPH and RI-Score.

Firstly, the concept of ESPH was utilized to develop two model types, namely Ligand-based and Complex-based. Both models function on the basis of extracting topological features from the input files and establishing a structure-activity relationship for the G-Quadruplex-ligand system. Each model was trained with a database of 139 ligands, using both GBT and RF algorithms. 10-fold cross validation was performed a total of ten cycles for each model to account for overfitting as well as the randomness of the algorithms. In total, four Ligand-based and six Complex-based models were developed. It was found that all of the models yielded a high positive average Pearson correlation in the range of 0.65 to 0.79, signifying the existence of a structure-function relationship between the topological features and the experimental values within the G-Quadruplex-ligand biomolecular system. Among the ESPH models, all Ligand-based models outperformed the Complex-based models. In addition, GBT algorithm was observed to be more superior than the RF algorithm.

Next, RI-Score was employed to develop two model types based on the exponential and Lorentz kernels. RI-Score function on the concept of enhanced rigidity mechanism within the biomolecular system. Similar to the ESPH models, RI-Score models were built using the same database and ML algorithms as well as tested by ten cycles of 10-fold cross validation. In total, six models each were developed using the exponential and Lorentz kernel. It was found that all of the models similarly yielded a high positive average Pearson correlation in the range of 0.65 to 0.78, signifying that rigidity strengthening is a mechanism of the G-Quadruplex-ligand system and therefore present a strong representation for the system. Within the RI-Score models, Lorentz kernel models typically generated stronger performances than their exponential counterparts.

Following which, comparisons were made between the developed models and an existing methodology, CORAL QSAR. The same database was employed for CORAL to ensure a fair comparison. The CORAL model developed performed strongly and proved superior to several of the new models developed within the research, in particular the Complex-based and RI-Score model. However, all four Ligand-based models outshone the QSAR method. A deeper

analysis and comparison showed that the Ligand-based 7-subgroups (GBT) model developed in the thesis was not only the most consistent model, but also the strongest performer, and therefore the best representation for the G-Quadruplex-ligand biomolecular system.

Finally, the Ligand-based models were subjected to internal testing using two sets of error laden input data. This was implemented to prove that the performances of the models were in fact due to the existence of the correlations between the topological features and the functions, rather than coincidence and luck. The Ligand-based models were shown to be sensitive to the presence of errors within the input data.

6.2 Future Work

The current progress of the research has shown the strength and effectiveness of topological features in representing G-Quadruplex-ligand biomolecular systems. While several models have been developed to varying degrees of success, there remained much room for further development before they can be considered for implementation in drug design processes. The penultimate goal would be to extract and utilize the topological features generated to produce a form of guiding rules that would assist in the design of new drug structures. Designing using these rules would, in principle, be very accurate given their origins from experimental database. However, at this stage, the goal is still a distance away, particularly given that ML algorithms results are sometimes inexplicable despite its apparent strong predictions and results.

Nevertheless, small steps of progress can still be made. In this thesis, the models have only been shown to provide a strong and robust representation of the G-Quadruplex-ligand system based on the set of 139 ligand database. The models developed are however, not yet configured to predict the performances of new untested ligands. Therefore, in the near future, the aim would be on the development of the predicting capability. Following which, experimental tests could be conducted to observe the accuracy of these predictions. This would come in the form of a collaboration with an experimental research group in

which several new ligands beyond the database could be selected for testing through both computational and experimental means. In other words, the selected ligands would be subjected to the topological models developed in this research for a prediction of their performances and compared to the experimental results as performed by the experimental groups. A high accuracy and precision rate would further endorse the credibility of these developed models.

References

1. Qiu, J.; Wang, M.; Zhang, Y.; Zeng, P.; Ou, T.; Tan, J.; Huang, S.; An, L.; Wang, H.; Gu, L.; Huang, Z.; Li, D. Biological Function and Medicinal Research Significance of G-Quadruplex Interactive Proteins. *Curr. Top. Med. Chem.* **2015**, *15*, 1971-1987.
2. Onel, B.; Lin, C.; Yang, D. DNA G-Quadruplex and its potential as anticancer drug target. *Sci China Chem.* **2014**, *57*(12), 1605-1614.
3. Rhodes, D.; Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acid Res.* **2015**, *43*(18), 8627-8637.
4. Lin, C.; Yang, D. Human Telomeric G-Quadruplex Structures and G-Quadruplex-Interactive Compounds. *Methods Mol Biol.* **2017**, *1587*, 171-196.
5. Bochman, M. L.; Paeschke, K.; Zakian, V. A. DNA secondary structures: stability and function of G-Quadruplex structures. *Nat. Rev. Genet.* **2012**, *13*(11), 770-780.
6. Han, H.; Hurley, L. H. G-quadruplex DNA: a potential target for anti-cancer drug design. *Trends Pharmacol Sci.* **2000**, *21*, 136-142.
7. Cimino-Reale, G.; Zaffaroni, N.; Folini, M. Emerging Role of G-Quadruplex DNA as Target in Anticancer Therapy. *Curr. Pharm. Des.* **2016**, *22*, 6612-6624.
8. Kwok, C. K.; Merrick, C. J. G-Quadruplexes: Prediction, Characterization, and Biological Application. *Trends Biotechnol.* **2017**, *35*(10), 997-1013.
9. Phan, A. T.; Human telomeric G-Quadruplex: structures of DNA and RNA sequences. *FEBS J.* **2010**, *277*, 1107-1117.
10. Jafri, M. A.; Ansari, S. A.; Alqahtani, M. H.; Shay, J. W. Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome Med.* **2016**, *8*(69).
11. Shammass, M. A. Telomeres, lifestyle, cancer, and aging. *Curr. Opin. Clin. Nutr. Metab. Care.* **2011**, *14*(1), 28-34.
12. Greider, C. W.; Blackburn, E. H. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. *Nature.* **1989**, *337*, 331-337.
13. Bodna, A. G.; Ouellette, M.; Frolkis, M.; Holt, S. E.; Chiu, C.; Morin, G. B.; Harley, C. B.; Shay, J. W.; Lichtsteiner, S.; Wright, W. E. Extension of Life-Span by Introduction of Telomerase into Normal Human Cells. *Science.* **1998**, *279*, 349-352.
14. Kim, N. W.; Piatyszek, M. A.; Prowse, K. R.; Harley, C. B.; West, M. D.; Ho, P. L.; Coviello, G. M.; Wright, W. E.; Weinrich, S. L.; Shay, J. W. Specific association of human telomerase activity with immortal cells and cancer. *Science.* **1994**, *266*(5193), 2011-2015.
15. Neidle, S. Quadruplex Nucleic Acids as Novel Therapeutic Targets. *J. Med. Chem.* **2016**, *59*(13), 5987-6011.
16. Yaku, H.; Fujimoto, T.; Murashima, T.; Miyoshi, D.; Sugimoto, N. Phthalocyanines: a new class of G-Quadruplex-ligands with many potential applications. *Chem. Commun.* **2012**, *48*, 6203-6216.
17. Largy, E.; Granzhan, A.; Hamon, F.; Verga, D.; Teulade-Fichou, M. Visualizing the Quadruplex: From Fluorescent Ligands to Light-Up Probes. *Top. Curr. Chem.* **2013**, *330*, 111-77.

18. Li, Q.; Xiang, J.; Yang, Q.; Sun, H.; Guan, A.; Tang, Y. G4LDB: a database for discovering and studying G-quadruplex ligands. *Nucleic Acids Res.* **2013**, *41*, 1115-1123.
19. Chen, B.; Wu, Y.; Tanaka, Y.; Zhang, W. Small Molecules Targeting c-Myc Oncogene: Promising Anti-Cancer Therapeutics. *Int. J. Biol. Sci.* **2014**, *10*, 10841096.
20. Cartwright, H. Using Artificial Intelligence in Chemistry and Biology: A Practical Guide, 1st ed.; CRC Press: Florida, 2008.
21. Panteleev, J.; Gao, H.; Jia, L. Recent applications of machine learning in medicinal chemistry. *Bioorganic Med. Chem. Lett.* **2018**, *28*, 2807-2815.
22. Cang, Z.; Mu, L.; Wei, G. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **2018**, *14*(1).
23. Sizemore, A. E.; Phillips-Cremins, J. E.; Ghrist, R.; Bassett, D. S. The importance of the whole: topological data analysis for the network neuroscientist. *arXiv*. [online] **2018**, 1806.05167. <https://arxiv.org/pdf/1806.05167.pdf>
24. Gerbhart, T.; Schrater, P. Adversary Detection in Neural Networks via Persistent Homology. *arXiv*. [online] **2017**, 1711.10056. <https://arxiv.org/pdf/1711.10056.pdf>
25. Cang, Z.; Wei, G. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **2017**, *13*(7).
26. Cang, Z.; Wei, G. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics.* **2017**, *33*(22), 3549-3557.
27. Xia, K.; Wei, G. Persistent homology analysis of protein structure, flexibility, and folding. *Int. J. Numer. Meth. Biomed. Engng.* **2014**, *30*, 814-844.
28. Nguyen, D. D.; Xiao, T.; Wang M.; Wei, G. Rigidity Strengthening: A Mechanism for Protein-Ligand Binding. *J. Chem. Inf. Model.* **2017**, *57*, 1715-1721.
29. Kovacev-Nikolic, V.; Bubenik, P.; Nikolic, D.; Heo, G. Using persistent homology and dynamical distances to analyse protein binding. *arXiv*. [online] **2015**, 1412.1394. <https://arxiv.org/pdf/1412.1394.pdf>
30. Cang, Z.; Wei, G. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int J Numer Meth Biomed Engng.* **2018**, *34*.
31. Ghrist R. Barcodes: The Persistent Topology of Data. *Bull. Am. Math. Soc.* **2008**, *45*(1), 61-75.
32. Kimura, Y.; Imai, K. Quantification of LSS using the persistent homology in the SDSS fields. *Adv. Space. Res.* **2017**, *60*, 722-736.
33. Xia, K.; Feng, X.; Tong, Y.; Wei, G. W. Persistent Homology for the Quantitative Prediction Fullerene Stability. *J. Comput. Chem.* **2015**, *36*, 408-422.
34. Perry, P. J.; Reszka, A. P.; Wood, A. A.; Read, M. A.; Gowan, S. M.; Dosanjh, H. S.; Trent, J. O.; Jenkins, T. C.; Kelland, L. R.; Neidle, S. Human Telomerase Inhibition by Regioisomeric Disubstituted Amidoanthracene-9,10-diones. *J. Med. Chem.* **1998**, *41*, 4873-4884.

35. Perry, P. J.; Read, M. A.; Davies, R. T.; Gowan, S. M.; Reszka, A. P.; Wood, A. A.; Kelland, L. R.; Neidle, S. 2,7-Disubstituted Amidofluorenone Derivatives as Inhibitors of Human Telomerase. *J. Med. Chem.* **1999**, *42*, 2679-2684.
36. Harrison, R. J.; Gowan, S. M.; Kelland, L. R.; Neidle, S. Human Telomerase Inhibition by Substituted Acridine Derivatives. *Bioorganic Med. Chem. Lett.* **1999**, *9*, 2463-2468.
37. Gowan, S. M.; Heald, R.; Stevens, M. F. G.; Kelland, L. R. Potent Inhibition of Telomerase by Small-Molecule Pentacyclic Acridines Capable of Interacting with G-Quadruplexes. *Mol. Pharmacol.* **2001**, *60*(5), 981-988.
38. Perry, P. J.; Gowan, S. M.; Reszka, A. P.; Polucci, P.; Jenkins, T. C.; Kelland, L. R.; Neidle, S. 1,4- and 2,6-Disubstituted Amidoanthracene-9,10-dione Derivatives as Inhibitors of Human Telomerase. *J. Med. Chem.* **1998**, *41*, 3253-3260.
39. Caprio, V.; Guyen, B.; Opoku-Boahen, Y.; Mann, J.; Gowan, S. M.; Kelland, L. M.; Read, M. A.; Neidle, S. A Novel Inhibitor of Human Telomerase Derived from 10*H*-Indolo[3,2-*b*]quinoline. *Bioorganic Med. Chem. Lett.* **2000**, *10*, 2063-2066.
40. Mergny, J.; Lacroix, L.; Teulade-Fichou, M.; Hounsou, C.; Guittat, L.; Hoarau, M.; Arimondo, P. B.; Vigneron, J.; Lehn, J.; Riou, J.; Garestier, T.; Helene, C. Telomerase inhibitors based on quadruplex ligands selected by a fluorescence assay. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*(6), 3062-3067.
41. Koepfel, F.; Riou, J.; Laoui, A.; Mailliet, P.; Arimondo, P. B.; Labit, D.; Petitgenet, O.; Helene, C.; Mergny, J. Ethidium derivatives bind to G-quartets, inhibit telomerase and act as fluorescent probes for quadruplexes. *Nucleic Acids Res.* **2001**, *29*(5), 1087-1096.
42. Read, M.; Harrison, R. J.; Romagnoli, B.; Tanious, F. A.; Gowan, S. H.; Reszka, A. P.; Wilson, W. D.; Kelland, L. R.; Neidle, S. Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*(9), 4844-4849.
43. Duan, W.; Rangan, A.; Vankayalapati, H.; Kim, M.; Zeng, Q.; Sun, D.; Han, H.; Fedoroff, O. Y.; Nishioka, D.; Rha, S. Y.; Izbicka, E.; Hoff, D. D. V.; Hurley, L. H. Design and Synthesis of Fluoroquinophenoxazines That Interact with Human Telomeric G-Quadruplexes and Their Biological Effects. *Mol Cancer Ther.* **2011**, *1*, 103-120.
44. Riou, J. F.; Guittat, L.; Mailliet, P.; Laoui, A.; Renou, E.; Petitgenet, O.; Megnin-Chanet, F.; Helene, C.; Mergny, J. L. Cell senescence and telomere shortening induced by a new series of specific G-quadruplex DNA ligands. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*(5), 2672-2677.
45. Alberti, P.; Schmitt, P.; Nguyen, C.; Rivalle, C.; Hoarau, M.; Grierson, D. S.; Mergny, J. Benzoindoloquinolines Interact with DNA Tetraplexes and Inhibit Telomerase. *Bioorganic Med. Chem. Lett.* **2002**, *12*, 1071-1074.
46. Gowan, S. M.; Harrison, J. R.; Patterson, L.; Valenti, M.; Read, M. A.; Neidle, S.; Kelland, L. R. A G-Quadruplex-Interactive Potent Small-Molecule Inhibitor of Telomerase Exhibiting in Vitro and in Vivo Antitumor Activity. *Mol. Pharmacol.* **2002**, *61*(5), 1154-1162.
47. Heald, R. A.; Stevens, M. F. G. Antitumour polycyclic acridines. Palladium(0) mediated syntheses of quino[4,3,2-*kl*]acridines bearing peripheral substituents as potential telomere maintenance inhibitors. *Org. Biomol. Chem.* **2003**, *1*, 3377-3389.

48. Guittat, L.; Alberti, P.; Rosu, F.; Miert, S. V.; Thetiot, E.; Pieters, L.; Gabelica, V.; Pauw, E. D.; Ottaviani, A.; Riou, J.; Mergny, J. Interactions of cryptolepine and neocryptolepine with unusual DNA structures. *Biochimie*. **2003**, *85*, 535-547.
49. Sissi, C.; Lucatello, L.; Krapcho, A. P.; Maloney, D. J.; Boxer, M. B.; Camarasa, M. V.; Pezzoni, G.; Menta, E.; Palumbo, M. Tri-, tetra- and heptacyclic perylene analogues as new potential antineoplastic agents based on DNA telomerase inhibition. *Bioorg. Med. Chem.* **2007**, *15*, 555-562.
50. Ma, Y.; Ou, T.; Hou, J.; Lu, Y.; Tan, J.; Gu, L.; Huang, Z. 9-*N*-Substituted berberine derivatives: Stabilization of G-quadruplex DNA and down-regulation of oncogene *c-myc*. *Bioorg. Med. Chem.* **2008**, *16*, 7582-7591.
51. Ma, Y.; Ou, T.; Tan, J.; Hou, J.; Huang, S.; Gu, L.; Huang, Z. Synthesis and evaluation of 9-*O*-substituted berberine derivatives containing aza-aromatic terminal group as highly selective telomeric G-quadruplex stabilizing ligands. *Bioorganic Med. Chem. Lett.* **2009**, *19*, 3414-3417.
52. Laronze-Cochard, M.; Kim, Y.; Brassart, B.; Riou, J.; Laronze, J.; Sapi, J. Synthesis and biological evaluation of novel 4,5-bis(9dialkylaminoalkyl)-substituted acridines as potent telomeric G-quadruplex ligands. *Eur. J. Med. Chem.* **2009**, *44*, 3880-3888.
53. Brillante, L.; Gaiotti, F.; Lovat, L.; Vincenzi, S.; Giacosa, S.; Torchio, F.; Segade, S. R.; Rolle, L.; Tomasi, D. Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical-mechanical characteristics in wine grapes. *Comput Electron Agr.* **2015**, *117*, 186-193.
54. Ogutu, J. O.; Piepho, H.; Schulz-Streeck, T. A comparison of random forests, boosting and support vector machines for genomic selection. *Proceedings*. **2011**, *5*.
55. Nawar, S.; Mouazen, A. M. Comparison between Random Forests, Artificial Neural Networks and Gradient Boosted Machines Methods of On-Line Vis-NIR Spectroscopy Measurements of Soil Total Nitrogen and Total Carbon. *Sensors*. **2017**, *17*.
56. Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Costanzo, L. D.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Kalro, T.; Liang, Y.; Lowe, R.; Namkoong, H.; Peisach, E.; Periskova, I.; Prlic, A.; Randle, C.; Rose, A.; Rose, P.; Sala, R.; Sekharan, M.; Shao, C.; Tan, L.; Tao, Y.; Valasatava, Y.; Voigt, M.; Westbrook, J.; Woo, J.; Yang, H.; Young, J.; Zhuravleva, M.; Zardecki, C. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **2019**, *47*, 464-474.
57. Wilson, T.; Costa, P. J.; Felix, V.; Williamson, M. P.; Thomas, J. A. Structural Studies on Dinuclear Ruthenium(II) Complexes That Bind Diastereoselectively to an Antiparallel Folded Human Telomere Sequence. *J. Med. Chem.* **2013**, *56*(21), 8674-8683.
58. Chung, W. J.; Heddi, B.; Tera, M.; Iida, K.; Nagasawa, K.; Phan, A. T. Solution Structure of an Intramolecular (3 + 1) Human Telomeric G-Quadruplex Bound to a Telomestatin Derivative. *J. Am. Chem. Soc.* **2013**, *135*(36), 13495-13501.

59. Nicoludis, J. M.; Miller, S. T.; Jeffrey, P. D.; Barrett, S. P.; Rablen, P. R.; Lawton, T. J.; Yatsunyk, L. A. Optimized End-Stacking Provides Specificity of *N*-Methyl Mesoporphyrin IX for Human Telomeric G-Quadruplex DNA. *J. Am. Chem. Soc.* **2012**, *134*(50), 20446-20456.
60. Chen M.; Song, G.; Wang, C.; Hu, D.; Ren, J.; Qu, X. Small-Molecule Selectively Recognizes Human Telomeric G-Quadruplex DNA and Regulates its Conformational Switch. *Biophys J.* **2009**, *97*, 2014-2023.
61. Lin, Y.; Chuang, S.; Wu, P.; Chen, C.; Jeyachandran, S.; Lo, S.; Huang, H.; Hou, M. Selective recognition and stabilization of new ligands targeting the potassium form of the human telomeric G-quadruplex DNA. *Sci. Rep.* **2016**, *6*(31019).
62. Ignjatovic, M. M.; Caldararu, O.; Dong, G.; Munoz-Gutierrez, C.; Adasme-Carreno, F.; Ryde, U. Binding-affinity predictions of HSP90 in the D3R Grand Challenge 2015 with docking, MM/GBSA, QM/MM, and free-energy simulations. *J Comput Aided Mol Des.* **2016**, *30*, 707-730.
63. Usha, S.; Selvaraj, S. Prediction of kinase-inhibitor binding affinity using energetic parameters. *J. Biomed. Inform.* **2016**, *12*(3), 172-181.
64. Forli, S.; Huey, R.; Pique, M. E.; Sanner, M. F.;Goodsell, D. S.; Olson, A. J. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* **2016**, *11*(5), 905-919.
65. Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for molecular docking: a review. *Biophys Rev.* **2017**, *9*, 91-102.
66. Xia, K.; Opron, K.; Wei, G. Multiscale Multiphysics and multidomain models – Flexibility and rigidity. *J. Chem. Phys.* **2013**, *139*.
67. Kubinyi, H. QSAR and 3D QSAR in drug design Part 1: methodology. *Drug Discov. Today.* **1997**, *2*(11), 457-467.
68. Young, D. C. Computational Drug Design: A Guide for Computational and Medicinal Chemists. Wiley: New York, 2009.
69. Toropov, A. A.; Toropova, A. P.; Benfenati, E. QSPR modelling for enthalpies of formation of organometallic compounds by means of SMILES-based optimal descriptors. *Chem. Phys. Lett.* **2008**, *461*, 343-347.
70. Manganelli, S.; Benfenati, E.; Manganaro A.; Kulkarni, S.; Barton-Maclaren, T. S.; Honma, M. New Quantitative Structure-Activity Relationship Models Improve Predictability of Ames Mutagenicity for Aromatic Azo Compounds. *Toxicol. Sci.* **2016**, *153*(2), 316-326.
71. Toropova, A. P.; Toropov, A. A. Use of the Index of Ideality of Correlation to improve models of Eco-toxicity. *Environ Sci Pollut Res.* **2018**, *25*, 31771-31775.
72. Toropova, A. P.; Toropov, A. A. Hybrid Optimal Descriptors as a Tool to Predict Skin Sensitization in accordance to OECD principles. *Toxicol. Lett.* **2017**, *275*, 57-66.
73. Toropova, A. P.; Toropov, A. A. The index of ideality of correlation: A criterion of predictability of QSAR models for skin permeability?. *Sci. Total. Environ.* **2017**, *586*, 466-472.
74. World Health Organization. International Agency for Research on Cancer. Latest global cancer data. [online] **2018**, <https://www.who.int/cancer/PRGlobocanFinal.pdf>

Appendix A

Complete list of ligands

<u>Index</u>	<u>Formula</u>	<u>IC50</u> (μ M)	<u>P(IC50)</u>	<u>Structure</u>
G4L0021 ³⁹	C ₃₄ H ₄₆ N ₄ O ₆	4.1	5.39
G4L0025 ³⁴	C ₂₈ H ₃₂ N ₄ O ₆	33.5	4.47
G4L0027 ³⁴	C ₂₈ H ₃₆ N ₄ O ₄	1.8	5.74
G4L0028 ³⁴	C ₃₀ H ₃₆ N ₄ O ₄	2.3	5.64

G4L0031 ³⁴	C ₂₄ H ₂₈ N ₄ O ₄	1.3	5.89
G4L0032 ³⁴	C ₂₈ H ₃₆ N ₄ O ₄	2.7	5.57
G4L0033 ³⁴	C ₃₀ H ₃₆ N ₄ O ₄	3.7	5.43
G4L0036 ³⁴	C ₂₄ H ₂₈ N ₄ O ₄	6.4	5.19
G4L0037 ³⁴	C ₂₈ H ₃₆ N ₄ O ₄	4.2	5.38

G4L0038 ³⁴	C ₃₀ H ₃₆ N ₄ O ₄	4.5	5.35
G4L0039 ³⁴	C ₂₈ H ₃₂ N ₄ O ₄	1.8	5.74
G4L0041 ³⁴	C ₂₄ H ₂₈ N ₄ O ₄	4.1	5.39
G4L0042 ³⁴	C ₂₈ H ₃₆ N ₄ O ₄	3.5	5.46
G4L0043 ³⁴	C ₃₀ H ₃₆ N ₄ O ₄	3.1	5.51

G4L0044 ³⁴	C ₂₈ H ₃₂ N ₄ O ₄	2.	5.70
G4L0046 ³⁴	C ₂₄ H ₂₈ N ₄ O ₄	4.7	5.33
G4L0047 ³⁴	C ₂₈ H ₃₆ N ₄ O ₄	4.3	5.37
G4L0048 ³⁴	C ₃₂ H ₄₂ N ₄ O ₄	11.1	4.95
G4L0049 ³⁴	C ₃₀ H ₃₈ N ₄ O ₄	5.0	5.30

G4L0050 ³⁴	C ₃₀ H ₃₈ N ₄ O ₆	34.5	4.46
G4L0051 ³⁴	C ₂₆ H ₃₄ N ₄ O ₄	7.0	5.15
G4L0052 ³⁴	C ₃₀ H ₄₂ N ₄ O ₄	3.1	5.51
G4L0053 ³⁴	C ₃₂ H ₄₂ N ₄ O ₄	8.6	5.07
G4L0054 ³⁴	C ₃₀ H ₃₈ N ₄ O ₄	8.8	5.06

G4L0055 ³⁴	C ₃₀ H ₃₈ N ₄ O ₆	14.0	4.85
G4L0056 ³⁴	C ₂₆ H ₃₄ N ₄ O ₄	13.2	4.88
G4L0057 ³⁴	C ₃₀ H ₄₂ N ₄ O ₄	16.8	4.77
G4L0058 ³⁴	C ₃₂ H ₄₂ N ₄ O ₄	7.8	5.11
G4L0059 ³⁴	C ₃₀ H ₃₈ N ₄ O ₄	8.2	5.09

G4L0060 ³⁴	C ₃₀ H ₃₈ N ₄ O ₆	10.0	5.00
G4L0061 ³⁴	C ₂₆ H ₃₄ N ₄ O ₄	4.4	5.36
G4L0062 ³⁴	C ₃₀ H ₄₂ N ₄ O ₄	7.5	5.12
G4L0066 ³⁴	C ₂₆ H ₃₄ N ₄ O ₄	17.3	4.76
G4L0068 ³⁴	C ₃₂ H ₄₂ N ₄ O ₄	7.8	5.11

G4L0069 ³⁴	C ₃₀ H ₃₈ N ₄ O ₄	16.0	4.80
G4L0070 ³⁴	C ₃₀ H ₃₈ N ₄ O ₆	16.5	4.78
G4L0071 ³⁴	C ₂₆ H ₃₄ N ₄ O ₄	14.5	4.84
G4L0080 ³⁵	C ₂₃ H ₂₈ N ₄ O ₃	16.2	4.79
G4L0081 ³⁵	C ₂₇ H ₃₆ N ₄ O ₃	15.5	4.81

G4L0083 ³⁵	C ₂₉ H ₃₆ N ₄ O ₃	9.0	5.05
G4L0084 ³⁵	C ₃₁ H ₄₀ N ₄ O ₃	12.0	4.92
G4L0086 ³⁵	C ₃₃ H ₄₄ N ₄ O ₃	14.0	4.85
G4L0087 ³⁵	C ₃₁ H ₄₀ N ₄ O ₅	8.0	5.10
G4L0088 ³⁵	C ₃₃ H ₄₄ N ₄ O ₅	16.5	4.78

G4L0089 ³⁵	C ₂₉ H ₃₆ N ₄ O ₅	18.5	4.73
G4L0090 ³⁵	C ₃₁ H ₄₀ N ₄ O ₃	44.0	4.36
G4L0091 ³⁵	C ₃₃ H ₄₄ N ₄ O ₃	17.0	4.77
G4L0096 ³⁵	C ₂₅ H ₃₄ N ₄ O ₃	18.5	4.73
G4L0097 ³⁵	C ₂₉ H ₄₂ N ₄ O ₃	25.4	4.60

G4L0098 ³⁵	C ₂₉ H ₃₈ N ₄ O ₅	27.3	4.56
G4L0099 ³⁵	C ₃₁ H ₄₂ N ₄ O ₃	21.0	4.68
G4L0103 ³⁵	C ₃₃ H ₄₆ N ₄ O ₅	33.0	4.48
G4L0108 ³⁶	C ₂₇ H ₄₀ N ₅ O ₂	5.8	5.24
G4L0109 ³⁶	C ₂₃ H ₃₂ N ₅ O ₂	8.2	5.09

G4L0110 ³⁶	C ₂₉ H ₄₀ N ₅ O ₂	2.8	5.55
G4L0112 ³⁶	C ₂₇ H ₃₆ N ₅ O ₂	5.2	5.28
G4L0113 ³⁶	C ₃₃ H ₄₈ N ₅ O ₂	2.7	5.57
G4L0114 ³⁶	C ₃₁ H ₄₄ N ₅ O ₂	2.6	5.59
G4L0115 ³⁶	C ₃₁ H ₄₄ N ₅ O ₂	1.35	5.87

G4L0116 ³⁶	C ₃₁ H ₄₄ N ₅ O ₂	4.4	5.36
G4L0117 ³⁶	C ₃₁ H ₄₄ N ₅ O ₄	5.4	5.27
G4L0118 ³⁶	C ₃₃ H ₄₈ N ₅ O ₄	4.1	5.39
G4L0119 ³⁶	C ₂₉ H ₄₀ N ₅ O ₄	8.0	5.10
G4L0120 ³⁶	C ₃₁ H ₄₄ N ₅ O ₂	3.1	5.51

G4L0124 ³⁷	C ₂₂ H ₁₇ F ₂ N ₂	0.33	6.48
G4L0133 ³⁸	C ₃₂ H ₄₂ N ₄ O ₆	9.4	5.03
G4L0142 ³⁸	C ₃₄ H ₄₄ N ₄ O ₆	16.5	4.78
G4L0154 ³⁹	C ₂₃ H ₂₈ N ₄	16.0	4.80
G4L0155 ³⁹	C ₂₀ H ₁₇ NO ₄	7.0	5.15

G4L0158 ⁴⁰	C ₂₈ H ₃₀ N ₄	0.3	6.52
G4L0159 ⁴⁰	C ₂₄ H ₂₁ N ₃	1.45	5.84
G4L0160 ⁴⁰	C ₂₆ H ₂₆ N ₄	0.75	6.12
G4L0161 ⁴⁰	C ₂₈ H ₃₀ N ₄	1.0	6.00
G4L0162 ⁴⁰	C ₂₈ H ₃₀ N ₄	0.5	6.30

G4L0163 ⁴⁰	C ₃₂ H ₄₀ N ₆	0.9	6.05
G4L0164 ⁴⁰	C ₂₈ H ₂₈ N ₄	2.0	5.70
G4L0166 ⁴⁰	C ₃₂ H ₄₀ N ₆	0.028	7.55
G4L0167 ⁴¹	C ₂₁ H ₂₀ N ₃	0.183	6.74
G4L0168 ⁴¹	C ₂₉ H ₂₇ N ₆	0.018	7.74

G4L0169 ⁴¹	C ₂₈ H ₂₆ N ₇	0.019	7.72
G4L0170 ⁴¹	C ₂₈ H ₂₆ N ₇	0.097	7.01
G4L0171 ⁴¹	C ₂₇ H ₂₈ N ₆	0.047	7.33
G4L0172 ⁴²	C ₂₇ H ₃₄ N ₅ O ₂	5.2	5.28
G4L0174 ⁴²	C ₃₅ H ₄₄ N ₇ O ₂	0.095	7.02

G4L0192 ⁴⁴	C ₂₈ H ₃₁ N ₉ S	0.041	7.39
G4L0193 ⁴⁵	C ₂₄ H ₂₃ N ₅ O ₂	0.5	6.30
G4L0194 ⁴⁵	C ₂₅ H ₂₆ N ₄ O	1.1	5.96
G4L0195 ⁴⁵	C ₂₄ H ₂₄ N ₄ O	9.8	5.01
G4L0196 ⁴⁵	C ₂₅ H ₂₆ N ₄ O	1.0	6.00

G4L0197 ⁴⁶	C ₃₅ H ₄₃ N ₇ O ₂	0.09	7.05
G4L0213 ⁴⁷	C ₂₄ H ₂₃ N ₂	2.0	5.70
G4L0214 ⁴⁷	C ₂₈ H ₂₆ N ₃ O ₂	0.37	6.43
G4L0238 ⁴⁸	C ₁₆ H ₁₃ N ₂	9.4	5.03
G4L1025 ⁴⁹	C ₂₄ H ₂₈ N ₄ O ₄	3.5	5.46

G4L1027 ⁴⁹	C ₂₄ H ₃₀ N ₄ O ₄	2.5	5.60
G4L1028 ⁴⁹	C ₂₆ H ₃₄ N ₄ O ₄	2.5	5.60
G4L1029 ⁴⁹	C ₂₅ H ₃₀ N ₄ O ₄	6.8	5.17
G4L1032 ⁴⁹	C ₃₄ H ₃₀ N ₂ O ₄	0.16	6.80
G4L1033 ⁴⁹	C ₄₀ H ₃₈ N ₂ O ₄	0.2	6.70

G4L1034 ⁴⁹	C ₃₄ H ₃₀ N ₂ O ₆	0.06	7.22
G4L1035 ⁴⁹	C ₃₆ H ₃₆ N ₂ O ₄	0.03	7.52
G4L1036 ⁴⁹	C ₄₀ H ₄₆ N ₂	0.4	6.40
G4L1178 ⁵⁰	C ₂₃ H ₂₆ ClN ₃ O ₃	3.9	5.41
G4L1179 ⁵⁰	C ₂₅ H ₂₆ ClN ₃ O ₄	7.1	5.15

G4L1180 ⁵⁰	C ₂₆ H ₃₀ ClN ₃ O ₃	4.0	5.40
G4L1181 ⁵⁰	C ₂₁ H ₂₁ ClN ₂ O ₄	5.5	5.26
G4L1182 ⁵⁰	C ₂₇ H ₂₅ ClN ₂ O ₃	3.6	5.44
G4L1183 ⁵⁰	C ₂₄ H ₂₈ ClN ₃ O ₃	3.3	5.48
G4L1184 ⁵⁰	C ₂₆ H ₃₀ ClN ₃ O ₄	4.8	5.32

G4L1185 ⁵⁰	C ₂₇ H ₃₂ ClN ₃ O ₃	3.2	5.49
G4L1186 ⁵⁰	C ₂₆ H ₃₀ ClN ₃ O ₃	2.3	5.64
G4L1187 ⁵⁰	C ₂₂ H ₂₄ ClN ₃ O ₃	2.0	5.70
G4L1311 ⁵¹	C ₂₇ H ₃₁ ClN ₂ O ₄	16.0	4.80
G4L1312 ⁵¹	C ₂₉ H ₃₁ ClN ₃ O ₄	15.0	4.82

G4L1313 ⁵¹	C ₂₇ H ₂₆ ClN ₂ O ₄	17.0	4.77
G4L1314 ⁵¹	C ₂₉ H ₃₀ ClN ₂ O ₄	15.0	4.82
G4L1315 ⁵¹	C ₃₀ H ₃₂ ClN ₂ O ₄	15.0	4.82
G4L1316 ⁵¹	C ₂₈ H ₂₈ ClN ₂ O ₄	14.0	4.85
G4L1317 ⁵¹	C ₂₈ H ₂₇ ClFN 2O ₄	16.0	4.80

G4L1318 ⁵¹	C ₂₉ H ₃₁ ClN ₃ O ₄	14.0	4.85
G4L1319 ⁵¹	C ₂₈ H ₂₈ ClN ₂ O ₅	28.0	4.55
G4L1320 ⁵¹	C ₂₆ H ₂₇ ClN ₃ O ₄	16.0	4.80
G4L1321 ⁵¹	C ₃₁ H ₂₈ ClN ₂ O ₄	15.0	4.82
G4L1322 ⁵¹	C ₂₆ H ₂₅ ClN ₄ O ₅	22.0	4.66

G4L1377 ⁵²	C ₂₄ H ₃₃ N ₅ O ₂	30.0	4.52	<p>CH₄</p>
G4L1378 ⁵²	C ₂₆ H ₃₇ N ₅ O ₂	30.0	4.52	<p>CH₄</p>
G4L1381 ⁵²	C ₂₉ H ₃₅ N ₃ O ₂	30.0	4.52
G4L1382 ⁵²	C ₃₁ H ₃₉ N ₃ O ₂	6.0	5.22
G4L1383 ⁵²	C ₃₃ H ₄₃ N ₃ O ₂	13.0	4.89

G4L1385 ⁵²	C ₃₃ H ₄₃ N ₃ O ₂	30.0	4.52
G4L1386 ⁵²	C ₃₁ H ₃₉ N ₃ O ₂	12.0	4.92
G4L1387 ⁵²	C ₃₃ H ₄₅ N ₅ O ₂	30.0	4.52
G4L1388 ⁵²	C ₂₃ H ₂₉ I ₂ N ₅ O ₂	7.3	5.14
G4L1389 ⁵²	C ₂₅ H ₃₃ I ₂ N ₅ O ₂	3.0	5.52

Appendix B

Complete results of ligand-based ESPH model

Ligand-based 4-subgroups ESPH model

Gradient Boosting Tree		
	Correlation	RMSE
1	0.9212	0.3339
2	0.7346	0.3343
3	0.7905	0.4874
4	0.7711	0.7768
5	0.8003	0.6476
6	0.6990	0.3320
7	0.9912	0.3525
8	0.8089	0.3572
9	0.8604	0.3493
10	0.7073	0.5602
MEAN	0.8084	0.4531
SD	0.0933	

Random Forest		
	Correlation	RMSE
1	0.8815	0.2653
2	0.6098	0.6101
3	0.7879	0.4039
4	0.9020	0.3554
5	0.8308	0.5195
6	0.8540	0.4659
7	0.8339	0.4679
8	0.9565	0.3130
9	0.6975	0.6822
10	0.7074	0.6457
MEAN	0.8061	0.4729
SD	0.1062	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.3889	0.8235
2	0.8943	0.4699
3	0.9246	0.3981
4	0.8176	0.3884
5	0.8782	0.4583
6	0.8471	0.6302
7	0.7938	0.5227
8	0.8420	0.5806
9	0.9631	0.1605
10	0.3342	0.4092
MEAN	0.7684	0.4841
SD	0.2205	

Random Forest		
	Correlation	RMSE
1	0.5793	0.3551
2	0.5502	0.5708
3	0.7045	0.6085
4	0.8261	0.3969
5	0.8513	0.4810
6	0.7848	0.3993
7	0.8873	0.3504
8	0.7216	0.3580
9	0.8392	0.8414
10	0.9005	0.5140
MEAN	0.7645	0.4875
SD	0.1232	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8218	0.5759
2	0.6913	0.4995
3	0.9023	0.4338
4	0.8606	0.5044
5	0.8446	0.4335
6	0.9473	0.3142
7	0.9229	0.3677

Random Forest		
	Correlation	RMSE
1	0.9218	0.6728
2	0.5014	0.4133
3	0.7838	0.3870
4	0.0902	0.3701
5	0.8711	0.3249
6	0.9074	0.5666
7	0.6293	0.6145

8	0.6506	0.5096
9	0.6269	0.3851
10	0.6118	0.6029
MEAN	0.7880	0.4627
SD	0.1297	

8	0.6800	0.6035
9	0.8229	0.4199
10	0.7925	0.4843
MEAN	0.7000	0.4857
SD	0.2514	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.0790	0.5617
2	0.7467	0.5993
3	0.3829	0.6613
4	0.9244	0.5247
5	0.5440	0.4176
6	0.9008	0.3054
7	0.9323	0.1881
8	0.9033	0.3944
9	0.8843	0.4276
10	0.7043	0.8016
MEAN	0.7002	0.4882
SD	0.2853	

Random Forest		
	Correlation	RMSE
1	0.6445	0.3289
2	0.8568	0.4119
3	0.9287	0.3767
4	0.8675	0.4571
5	0.7870	0.7106
6	0.2377	0.6926
7	0.9252	0.5585
8	0.7247	0.5418
9	0.7802	0.5059
10	0.9288	0.2904
MEAN	0.7681	0.4874
SD	0.2090	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8901	0.4005
2	0.8111	0.4311
3	0.8442	0.5356
4	0.9130	0.5112
5	0.5994	0.6301
6	0.4015	0.4929
7	0.8558	0.5233
8	0.7722	0.3998
9	0.7286	0.4017
10	0.7864	0.3394
MEAN	0.7602	0.4666
SD	0.1547	

Random Forest		
	Correlation	RMSE
1	0.9282	0.3325
2	0.8773	0.6249
3	0.7325	0.6616
4	0.4951	0.7161
5	0.5029	0.3538
6	0.6111	0.3172
7	0.7804	0.5795
8	0.4231	0.4600
9	0.9461	0.4545
10	0.8742	0.4237
MEAN	0.7171	0.4924
SD	0.1956	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6978	0.6210
2	0.8583	0.5392
3	0.8888	0.3338
4	0.7181	0.5156
5	0.8108	0.2833

Random Forest		
	Correlation	RMSE
1	0.8598	0.4745
2	0.8504	0.4740
3	0.9054	0.2525
4	0.5587	0.5142
5	0.8047	0.3975

6	0.8501	0.4868
7	0.6967	0.5367
8	0.7062	0.5036
9	0.7379	0.4405
10	0.9645	0.3552
MEAN	0.7929	0.4616
SD	0.0948	

6	0.9343	0.2302
7	0.5139	0.5698
8	0.6955	0.8788
9	0.8467	0.5813
10	0.8115	0.5427
MEAN	0.7781	0.4916
SD	0.1429	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.9103	0.4896
2	0.6749	0.4657
3	0.5049	0.5626
4	0.7513	0.4894
5	0.8578	0.2679
6	0.8943	0.3324
7	0.9515	0.2804
8	0.1425	0.5839
9	0.7874	0.4970
10	0.7894	0.6846
MEAN	0.7264	0.4654
SD	0.2429	

Random Forest		
	Correlation	RMSE
1	0.5271	0.6439
2	0.5735	0.4688
3	0.8069	0.5160
4	0.8342	0.4597
5	0.9204	0.4572
6	0.7261	0.3254
7	0.8445	0.4168
8	0.5114	0.5981
9	0.9226	0.3820
10	0.9488	0.5913
MEAN	0.7615	0.4859
SD	0.1682	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8422	0.4808
2	0.9486	0.4350
3	0.8739	0.4071
4	0.7704	0.5173
5	0.9318	0.3086
6	0.5146	0.6744
7	0.9291	0.2812
8	0.8195	0.3848
9	0.6580	0.5515
10	0.8727	0.4397
MEAN	0.8161	0.4480
SD	0.1370	

Random Forest		
	Correlation	RMSE
1	0.5818	0.5518
2	0.6623	0.6208
3	0.6127	0.6863
4	0.8786	0.6012
5	0.1475	0.4389
6	0.9239	0.3397
7	0.9465	0.4722
8	0.8773	0.3136
9	0.8593	0.4707
10	0.9537	0.3956
MEAN	0.7444	0.4891
SD	0.2522	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.1683	0.5612
2	0.8767	0.5377
3	0.8503	0.4191

Random Forest		
	Correlation	RMSE
1	0.8237	0.2466
2	0.6946	0.5222
3	0.8391	0.5541

4	0.8560	0.4744
5	0.8004	0.5307
6	0.8871	0.4660
7	0.8236	0.4618
8	0.8871	0.3621
9	0.8596	0.4316
10	0.8721	0.4117
MEAN	0.7881	0.4656
SD	0.2195	

4	0.7313	0.6500
5	0.6060	0.4072
6	0.8909	0.3432
7	0.8957	0.4819
8	0.6949	0.6146
9	0.7759	0.4818
10	0.9620	0.6344
MEAN	0.7914	0.4936
SD	0.1106	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7196	0.5517
2	0.5955	0.4869
3	0.8872	0.4691
4	0.7013	0.4807
5	0.6526	0.3503
6	0.9129	0.3124
7	0.8742	0.2930
8	0.8814	0.5864
9	0.7650	0.5669
10	0.9009	0.4254
MEAN	0.7891	0.4523
SD	0.1166	

Random Forest		
	Correlation	RMSE
1	0.8199	0.4871
2	0.8842	0.4988
3	0.8016	0.5138
4	0.8621	0.4600
5	0.4910	0.8439
6	0.8588	0.2758
7	0.6245	0.4065
8	0.9099	0.5464
9	0.7826	0.3125
10	0.7927	0.4500
MEAN	0.7827	0.4795
SD	0.1293	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7738	0.4648	0.0363	0.7614	0.4866	0.0328

Ligand-based 7-subgroups ESPH model

Gradient Boosting Tree		
	Correlation	RMSE
1	0.9220	0.3924
2	0.7240	0.6923
3	0.5809	0.6308
4	0.9198	0.2876
5	0.7069	0.4566
6	0.8762	0.5310
7	0.6486	0.3971
8	0.8295	0.3270
9	0.7225	0.4394
10	0.8835	0.3124
MEAN	0.7814	0.4467
SD	0.1206	

Random Forest		
	Correlation	RMSE
1	0.7796	0.4284
2	0.8794	0.3538
3	0.6110	0.3971
4	0.8155	0.5162
5	0.8814	0.5578
6	0.7569	0.4955
7	0.7059	0.4802
8	0.7766	0.4792
9	0.9361	0.4532
10	0.7737	0.6219
MEAN	0.7916	0.4783
SD	0.0937	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8061	0.5754
2	0.6941	0.5601
3	0.6532	0.6085
4	0.4888	0.3344
5	0.7929	0.6644
6	0.6005	0.5279
7	0.8424	0.5000
8	0.8941	0.3602
9	0.9281	0.2850
10	0.8191	0.4266
MEAN	0.7519	0.4843
SD	0.1390	

Random Forest		
	Correlation	RMSE
1	0.8399	0.5110
2	0.8872	0.5428
3	0.8066	0.4500
4	0.8151	0.5500
5	0.3807	0.4832
6	0.4405	0.5561
7	0.8385	0.5913
8	0.9408	0.2883
9	0.8753	0.4741
10	0.7993	0.2307
MEAN	0.7624	0.4678
SD	0.1907	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8398	0.5051
2	0.9084	0.3441
3	0.7399	0.7301
4	0.6633	0.4773
5	0.7049	0.6325
6	0.9242	0.3214
7	0.9581	0.3516
8	0.8265	0.4720
9	0.8508	0.3419
10	0.4752	0.3570

Random Forest		
	Correlation	RMSE
1	0.8220	0.6977
2	0.7672	0.4651
3	0.8796	0.3195
4	0.7169	0.7665
5	0.8559	0.4082
6	0.6396	0.5861
7	0.6182	0.7048
8	0.9353	0.3633
9	0.6920	0.3800
10	0.8074	0.3481

MEAN	0.7891	0.4533
SD	0.1463	

MEAN	0.7734	0.5039
SD	0.1052	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7280	0.5386
2	0.7679	0.4550
3	0.8670	0.4845
4	0.8659	0.2968
5	0.9122	0.3226
6	0.7282	0.4983
7	0.8148	0.7736
8	0.7418	0.6080
9	0.6922	0.3869
10	0.6295	0.3724
MEAN	0.7747	0.4737
SD	0.0886	

Random Forest		
	Correlation	RMSE
1	0.8969	0.4743
2	0.8948	0.4874
3	0.5155	0.3713
4	0.4389	0.8044
5	0.7485	0.3807
6	0.7995	0.5445
7	0.7980	0.4740
8	0.7995	0.3753
9	0.8719	0.5045
10	0.9078	0.5125
MEAN	0.7671	0.4929
SD	0.1627	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7103	0.5022
2	0.6772	0.6218
3	0.8084	0.5335
4	0.6627	0.5288
5	0.9463	0.3179
6	0.8544	0.4448
7	0.8922	0.2910
8	0.8500	0.5045
9	0.8934	0.4276
10	0.7505	0.4928
MEAN	0.8045	0.4665
SD	0.0991	

Random Forest		
	Correlation	RMSE
1	0.9616	0.4666
2	0.6389	0.5440
3	0.7914	0.4923
4	0.7243	0.3839
5	0.8510	0.4730
6	0.9114	0.6603
7	0.7771	0.3567
8	0.7429	0.5317
9	0.4317	0.5476
10	0.9067	0.2596
MEAN	0.7737	0.4716
SD	0.1552	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7390	0.5423
2	0.8627	0.5305
3	0.7604	0.4796
4	0.9438	0.4744
5	0.9077	0.2748
6	0.8158	0.3449
7	0.7730	0.6857
8	0.7244	0.3449

Random Forest		
	Correlation	RMSE
1	0.7484	0.4586
2	0.8955	0.4652
3	0.8914	0.3638
4	0.8929	0.4625
5	0.8920	0.6057
6	0.5921	0.7358
7	0.9209	0.3725
8	0.7006	0.3610

9	0.8557	0.4677
10	0.7414	0.4550
MEAN	0.8124	0.4600
SD	0.0769	

9	0.4383	0.5975
10	0.7464	0.3920
MEAN	0.7719	0.4815
SD	0.1601	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8399	0.6218
2	0.7616	0.3603
3	0.8842	0.4277
4	0.6993	0.5660
5	0.7484	0.9152
6	0.7575	0.2659
7	0.8236	0.3995
8	0.7956	0.4892
9	0.8752	0.3336
10	0.4946	0.5230
MEAN	0.7680	0.4902
SD	0.1125	

Random Forest		
	Correlation	RMSE
1	0.8563	0.4058
2	0.8469	0.3012
3	0.9448	0.5306
4	0.9476	0.5261
5	0.8789	0.2975
6	0.4240	0.7841
7	0.8264	0.2885
8	0.8265	0.5866
9	0.8801	0.4728
10	0.6068	0.5318
MEAN	0.8038	0.4725
SD	0.1635	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7760	0.3633
2	0.8272	0.5660
3	0.8011	0.6381
4	0.7225	0.5702
5	0.7498	0.4457
6	0.9800	0.2188
7	0.9315	0.2896
8	0.7412	0.4774
9	0.8831	0.4568
10	0.7595	0.4091
MEAN	0.8172	0.4435
SD	0.0873	

Random Forest		
	Correlation	RMSE
1	0.7997	0.5130
2	0.7750	0.5947
3	0.6764	0.3832
4	0.8209	0.6576
5	0.6930	0.5448
6	0.8953	0.5208
7	0.8529	0.3739
8	0.9220	0.3500
9	0.9365	0.3318
10	0.5915	0.5840
MEAN	0.7963	0.4854
SD	0.1138	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8191	0.3521
2	0.7842	0.5145
3	0.8482	0.3275
4	0.9442	0.3358
5	0.7727	0.5931
6	0.8028	0.5930

Random Forest		
	Correlation	RMSE
1	0.8082	0.4729
2	0.2577	0.8324
3	0.8834	0.5054
4	0.8454	0.4652
5	0.1646	0.3547
6	0.7914	0.7900

7	0.6138	0.6145
8	0.6692	0.5073
9	0.9597	0.3013
10	0.8486	0.3300
MEAN	0.8062	0.4469
SD	0.1074	

7	0.7730	0.5117
8	0.8798	0.3433
9	0.4069	0.3254
10	0.8706	0.4728
MEAN	0.6681	0.5074
SD	0.2788	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8985	0.5386
2	0.9648	0.2044
3	0.4811	0.7465
4	0.9248	0.3270
5	0.8838	0.2942
6	0.8149	0.4232
7	0.9057	0.5904
8	0.7751	0.3098
9	0.4847	0.6147
10	0.5168	0.5101
MEAN	0.7650	0.4559
SD	0.1945	

Random Forest		
	Correlation	RMSE
1	0.9456	0.6290
2	0.7389	0.5723
3	0.1902	0.4630
4	0.8191	0.4400
5	0.8448	0.5909
6	0.6555	0.3443
7	0.8992	0.5183
8	0.6603	0.5176
9	0.8770	0.3748
10	0.9057	0.4473
MEAN	0.7536	0.4898
SD	0.2224	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7871	0.4621	0.0223	0.7662	0.4851	0.0379

Appendix C

Complete results of complex-based ESPH model

2MB3 model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8541	0.3270	0.7283	0.6334
2	0.8438	0.5904	0.7420	0.7392
3	0.7167	0.4638	0.6457	0.6105
4	0.5879	0.8351	0.6812	0.5645
5	0.5507	0.7132	0.6779	0.4053
6	0.5759	0.5491	0.6062	0.6259
7	0.9007	0.5052	0.9202	0.7193
8	0.7785	0.5793	0.6341	0.4846
9	0.8402	0.3729	0.6979	0.5400
10	0.9313	0.3323	0.4086	0.4486
MEAN	0.7580	0.5268	0.6742	0.5771
SD	0.1418		0.1275	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7639	0.7508	0.2143	0.6837
2	0.8085	0.6814	0.9269	0.6170
3	0.9346	0.4460	0.7434	0.4683
4	0.8176	0.4669	0.7288	0.6576
5	0.6537	0.4443	0.9184	0.3844
6	0.7523	0.4771	0.0603	0.6651
7	0.7491	0.5355	0.8271	0.5361
8	0.6453	0.5302	0.6273	0.3848
9	0.8045	0.4715	0.6148	0.7120
10	0.6046	0.4742	0.8519	0.6049
MEAN	0.7534	0.5278	0.6513	0.5714
SD	0.0980		0.2935	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8011	0.6112	0.1305	0.5148
2	0.9014	0.5969	0.8496	0.6529
3	0.7569	0.6265	0.7944	0.5823
4	0.6385	0.5143	0.2462	0.7178
5	0.8059	0.5333	0.7595	0.5206
6	0.7189	0.4399	0.5364	0.7437

7	0.8268	0.3924	0.4767	0.6958
8	0.8977	0.5406	0.9213	0.4819
9	0.7262	0.4220	0.9417	0.4575
10	0.7715	0.4726	0.8164	0.3943
MEAN	0.7845	0.5150	0.6473	0.5762
SD	0.0809		0.2858	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8140	0.5564	0.7651	0.6024
2	0.6778	0.5105	0.8825	0.5047
3	0.4680	0.7316	0.2608	0.8414
4	0.7139	0.5208	0.8557	0.6397
5	0.8630	0.5776	0.8357	0.6171
6	0.8703	0.3676	0.9076	0.4582
7	0.9420	0.5783	0.7534	0.4068
8	0.6258	0.5783	0.8299	0.3793
9	0.8749	0.4574	0.7404	0.5273
10	0.6532	0.4203	0.2107	0.8122
MEAN	0.7503	0.5299	0.7042	0.5789
SD	0.1471		0.2532	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8254	0.3510	0.8652	0.3770
2	0.8487	0.5284	0.3985	0.7001
3	0.8944	0.4640	0.7369	0.5417
4	0.8757	0.4981	0.5118	0.7594
5	0.7150	0.7034	0.5601	0.5083
6	0.7583	0.5813	0.9075	0.5991
7	0.6126	0.4475	0.6516	0.4939
8	0.8971	0.4500	0.7418	0.6829
9	0.4675	0.6658	0.7452	0.5458
10	0.2237	0.4415	0.8819	0.5556
MEAN	0.7118	0.5131	0.7000	0.5764
SD	0.2201		0.1689	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8603	0.5852	0.8066	0.5112
2	0.6937	0.3584	0.4844	0.4913
3	0.5574	0.5998	0.5018	0.8020

4	0.7620	0.6221	0.7538	0.4634
5	0.7012	0.6479	0.9191	0.5728
6	0.9261	0.3607	0.0009	0.4319
7	0.4973	0.3625	0.8139	0.4549
8	0.4603	0.7676	0.5231	0.6462
9	0.8528	0.4294	0.8790	0.6879
10	0.8663	0.4633	0.5543	0.6308
MEAN	0.7177	0.5197	0.6237	0.5692
SD	0.1658		0.2741	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8749	0.4871	0.4189	0.4510
2	0.8450	0.5680	0.5611	0.8159
3	0.6469	0.8615	0.6636	0.3243
4	0.6651	0.5530	0.9297	0.4931
5	0.7888	0.4780	0.4270	0.4301
6	0.7974	0.4922	0.3483	0.5569
7	0.7489	0.6478	0.8065	0.5033
8	0.7734	0.4496	0.8622	0.7656
9	0.5545	0.3722	0.4853	0.7218
10	0.4872	0.4936	0.7480	0.7020
MEAN	0.7182	0.5403	0.6251	0.5764
SD	0.1265		0.2056	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7716	0.5408	0.8217	0.3609
2	0.7287	0.5766	0.8181	0.7806
3	0.8310	0.5769	0.7225	0.6557
4	0.9330	0.4345	0.7636	0.7193
5	0.7844	0.4091	0.6272	0.4611
6	0.6244	0.5274	0.8229	0.4339
7	0.4484	0.6692	0.1621	0.5851
8	0.8567	0.6847	0.4943	0.6259
9	0.7959	0.4771	0.9074	0.5670
10	0.7536	0.5060	0.8712	0.4398
MEAN	0.7528	0.5402	0.7011	0.5629
SD	0.1343		0.2255	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8313	0.6264	0.6519	0.3476
2	0.6932	0.5383	0.1345	0.8208
3	0.2808	0.6675	0.8590	0.4503
4	0.8115	0.4885	0.6484	0.4501
5	0.7777	0.4337	0.8338	0.8403
6	0.8250	0.6971	0.8591	0.5219
7	0.8992	0.6552	0.5783	0.8207
8	0.8347	0.4519	0.8864	0.4954
9	0.5480	0.3507	0.6837	0.5062
10	0.7897	0.4165	0.8381	0.6375
MEAN	0.7291	0.5326	0.6973	0.5891
SD	0.1849		0.2266	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7521	0.5673	0.8540	0.5025
2	0.6235	0.3332	0.5986	0.5129
3	0.8749	0.5159	0.7657	0.5599
4	0.8524	0.4688	0.7747	0.4996
5	0.7936	0.6028	0.5849	0.6647
6	0.6854	0.4045	0.8861	0.5384
7	0.6919	0.5990	0.7782	0.4559
8	0.8489	0.2590	0.8329	0.4415
9	0.6548	0.7524	0.1510	1.0480
10	0.6679	0.5199	0.7190	0.5542
MEAN	0.7445	0.5023	0.6945	0.5778
SD	0.0924		0.2153	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7420	0.5248	0.0226	0.6719	0.5755	0.0323

2MCC model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7721	0.6221	0.3576	0.5490
2	0.8468	0.4881	0.8824	0.5396
3	0.8390	0.4964	0.9030	0.6789
4	0.7234	0.4519	0.5805	0.7454
5	0.8643	0.6287	0.8655	0.5432
6	0.9290	0.4535	0.8141	0.6479
7	0.8506	0.5721	0.6219	0.4930
8	0.8051	0.4914	0.4518	0.6235
9	0.6892	0.3965	0.8457	0.5406
10	0.6361	0.2061	0.8578	0.4007
MEAN	0.7956	0.4807	0.7180	0.5762
SD	0.0898		0.1992	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8779	0.5790	0.4182	0.5710
2	0.8359	0.3365	0.8290	0.5309
3	0.7568	0.6355	0.9239	0.7201
4	0.5778	0.4464	0.1829	0.7118
5	0.6905	0.4373	0.6720	0.5028
6	0.8368	0.5529	0.8160	0.6882
7	0.5901	0.2972	0.7403	0.5376
8	0.8380	0.6499	0.6063	0.3678
9	0.8030	0.5522	0.7193	0.8858
10	0.8248	0.4301	0.3680	0.4006
MEAN	0.7631	0.4917	0.6276	0.5917
SD	0.1077		0.2349	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7829	0.7116	0.6754	0.5189
2	0.8515	0.5622	0.4592	0.4895
3	0.8441	0.3302	0.8146	0.7969
4	0.7578	0.3956	0.6261	0.4776
5	0.7010	0.4050	0.7758	0.5964
6	0.5243	0.5492	0.7526	0.4754
7	0.8302	0.6393	0.9382	0.4030
8	0.7858	0.4374	0.6558	0.5738
9	0.9028	0.4135	0.5765	0.5331
10	0.7989	0.5447	0.6888	0.8527

MEAN	0.7779	0.4989	0.6963	0.5717
SD	0.1050		0.1332	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6974	0.5955	0.5852	0.5083
2	0.8985	0.5145	0.7219	0.5239
3	0.6828	0.4694	0.8824	0.5839
4	0.9130	0.5832	0.8307	0.5936
5	0.9532	0.3740	0.6199	0.6494
6	0.8789	0.3208	0.4433	0.5990
7	0.0149	0.6824	0.5828	0.4518
8	0.8958	0.3860	0.9092	0.7919
9	0.5616	0.6325	0.7878	0.5947
10	0.8636	0.4624	0.8159	0.4365
MEAN	0.7360	0.5021	0.7179	0.5733
SD	0.2832		0.1534	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.2305	0.7899	0.8801	0.2873
2	0.7713	0.4219	0.8210	0.7100
3	0.8505	0.4842	0.7437	0.6430
4	0.7006	0.3828	0.5198	0.4978
5	0.8610	0.5230	0.2806	0.6161
6	0.8485	0.6913	0.7253	0.4352
7	0.8075	0.4045	0.6803	0.7010
8	0.9159	0.3029	0.8134	0.5483
9	0.8078	0.5124	0.9415	0.5807
10	0.8680	0.6497	0.6723	0.6012
MEAN	0.7662	0.5163	0.7078	0.5621
SD	0.1973		0.1914	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7608	0.5117	0.9326	0.6760
2	0.7746	0.5989	0.8673	0.6106
3	0.6876	0.5324	0.5057	0.4936
4	0.8218	0.3903	0.8288	0.5462
5	0.0073	0.3800	0.0604	0.4435
6	0.8809	0.4842	0.8749	0.3658
7	0.9609	0.5193	0.2312	0.5327
8	0.8876	0.5299	0.8032	0.7253

9	0.6625	0.5332	0.6599	0.6564
10	0.7014	0.5513	0.8322	0.5895
MEAN	0.7145	0.5031	0.6596	0.5640
SD	0.2665		0.2999	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9047	0.6168	0.9323	0.6935
2	0.7115	0.4658	0.9011	0.6495
3	0.7024	0.4457	0.6835	0.6244
4	0.6681	0.5336	0.8914	0.5133
5	0.9095	0.5339	0.4489	0.5417
6	0.6568	0.3513	0.7067	0.4497
7	0.8886	0.5891	0.4907	0.6247
8	0.9092	0.3254	0.6615	0.6179
9	0.6858	0.7213	0.9070	0.5638
10	0.5871	0.5180	0.3505	0.4088
MEAN	0.7624	0.5101	0.6973	0.5687
SD	0.1257		0.2120	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7991	0.4605	0.7838	0.3113
2	0.8177	0.6024	0.7525	0.4385
3	0.8172	0.5242	0.7928	0.4323
4	0.8583	0.3255	0.6482	0.6441
5	0.6445	0.6106	0.8681	0.7195
6	0.8298	0.4403	0.8342	0.4991
7	0.8315	0.4375	0.7441	0.5009
8	0.7944	0.4903	0.8077	0.5547
9	0.6721	0.6776	0.7258	0.8241
10	0.8614	0.5039	0.3905	0.7258
MEAN	0.7926	0.5073	0.7348	0.5650
SD	0.0743		0.1355	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8745	0.5679	0.6901	0.6201
2	0.6848	0.5545	0.9240	0.5778
3	0.6736	0.4714	0.1100	0.4350
4	0.7215	0.6143	0.4648	0.5946
5	0.1868	0.5104	0.6582	0.3949
6	0.9072	0.8136	0.8358	0.6509

7	0.9360	0.4075	0.3203	0.7037
8	0.8799	0.4188	0.8541	0.4398
9	0.8518	0.4654	0.5955	0.5892
10	0.8330	0.4252	0.9000	0.6485
MEAN	0.7549	0.5249	0.6353	0.5655
SD	0.2203		0.2689	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7225	0.4388	0.7820	0.5955
2	0.9281	0.5111	0.7801	0.6623
3	0.4647	0.4283	0.3993	0.6373
4	0.6696	0.4437	0.7794	0.6904
5	0.6862	0.4807	0.8200	0.5149
6	0.8330	0.5093	0.7513	0.3259
7	0.7925	0.5360	0.8725	0.4197
8	0.9463	0.7004	0.2316	0.5006
9	0.8890	0.4958	0.8848	0.8257
10	0.5460	0.5146	0.8980	0.4516
MEAN	0.7478	0.5059	0.7199	0.5624
SD	0.1607		0.2225	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7611	0.5041	0.0247	0.6914	0.5700	0.0375

4G0F Model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7339	0.5081	0.5761	0.5649
2	0.6633	0.5037	0.8474	0.4312
3	0.6629	0.4825	0.4695	0.7189
4	0.8710	0.6242	0.5471	0.6037
5	0.8578	0.6047	0.5373	0.6747
6	0.7213	0.5331	0.7460	0.6829
7	0.3551	0.1640	0.7690	0.4032
8	0.8305	0.6184	0.7723	0.6785
9	0.7691	0.5280	0.7421	0.5714
10	0.5334	0.6963	0.6062	0.7266
MEAN	0.6998	0.5263	0.6613	0.6056
SD	0.1588		0.1281	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8140	0.5089	0.8466	0.6673
2	0.6415	0.7178	0.3783	0.5693
3	0.8910	0.6136	0.7046	0.7173
4	0.6669	0.4384	0.1466	0.4358
5	0.7782	0.5320	0.7070	0.5249
6	0.8021	0.5015	0.8117	0.8508
7	0.8671	0.5967	0.3743	0.5548
8	0.4290	0.6960	0.8091	0.5435
9	0.8101	0.3593	0.7172	0.5156
10	0.7020	0.5634	0.8012	0.6106
MEAN	0.7402	0.5528	0.6296	0.5990
SD	0.1366		0.2410	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7473	0.6110	0.7765	0.3697
2	0.6924	0.4228	0.7032	0.5687
3	0.6707	0.5095	0.8403	0.7870
4	0.7906	0.5099	0.3756	0.5602
5	0.7883	0.5921	0.1994	0.7724
6	0.7483	0.3076	0.7308	0.4874
7	0.8496	0.4289	0.8259	0.5755
8	0.5840	0.7571	0.9024	0.4608
9	0.6486	0.4614	0.8545	0.5525
10	0.6679	0.8100	0.6551	0.7898

MEAN	0.7188	0.5410	0.6864	0.5924
SD	0.0799		0.2269	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.4525	0.5138	0.8678	0.6282
2	0.7737	0.5879	0.8762	0.5231
3	0.8056	0.3542	0.7955	0.5218
4	0.7372	0.5340	0.7977	0.7434
5	0.9032	0.4897	0.4426	0.5328
6	0.8389	0.5046	0.5516	0.7543
7	0.6500	0.8008	0.6735	0.5239
8	0.3155	0.4721	0.6938	0.6033
9	0.7694	0.5836	0.8262	0.4736
10	0.7967	0.6560	0.1307	0.7343
MEAN	0.7043	0.5497	0.6656	0.6039
SD	0.1838		0.2345	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.4688	0.6092	0.7608	0.4864
2	0.8669	0.7290	0.7610	0.7285
3	0.7667	0.3482	0.8500	0.5335
4	0.8519	0.4610	0.5380	0.7377
5	0.4330	0.5916	0.5358	0.8244
6	0.8165	0.4089	0.6409	0.6008
7	0.8460	0.5321	0.1897	0.4902
8	0.9312	0.6510	0.7710	0.5712
9	0.3789	0.6755	0.8498	0.5411
10	0.8707	0.4139	0.7842	0.4399
MEAN	0.7230	0.5420	0.6681	0.5954
SD	0.2096		0.2032	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6198	0.5637	0.6559	0.5364
2	0.8062	0.6397	0.9142	0.4161
3	0.7933	0.4862	0.8608	0.6273
4	0.8076	0.6310	0.7015	0.6503
5	0.5454	0.5172	0.3722	0.7255
6	0.7167	0.5820	0.7163	0.2888
7	0.7520	0.5705	0.5672	0.6398
8	0.6289	0.4011	0.9426	0.5354

9	0.4987	0.4753	0.7765	0.6356
10	0.8783	0.5390	0.0977	1.0227
MEAN	0.7047	0.5406	0.6605	0.6078
SD	0.1256		0.2605	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.5735	0.5387	0.7072	0.7369
2	0.7038	0.3361	0.3911	0.5868
3	0.8485	0.5981	0.5801	0.5329
4	0.7418	0.6987	0.9064	0.6719
5	0.6373	0.6930	0.3318	0.6606
6	0.7126	0.5023	0.8801	0.4053
7	0.9267	0.4719	0.8595	0.5756
8	0.5801	0.4430	0.2804	0.5415
9	0.7451	0.5638	0.9465	0.5759
10	0.7936	0.6820	0.5536	0.7258
MEAN	0.7263	0.5528	0.6437	0.6013
SD	0.1123		0.2522	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.3882	0.5729	0.5311	0.4482
2	0.7683	0.4934	0.8099	0.6282
3	0.9125	0.7371	0.4979	0.6554
4	0.7590	0.7321	0.9122	0.7217
5	0.1921	0.5458	0.8932	0.6508
6	0.8234	0.6566	0.7216	0.6115
7	0.6647	0.6472	0.5072	0.6535
8	0.8110	0.4066	0.6617	0.2500
9	0.1870	0.4791	0.5610	0.6278
10	0.9569	0.4152	0.7921	0.6207
MEAN	0.6463	0.5686	0.6888	0.5868
SD	0.2863		0.1597	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.5943	0.5137	0.4709	0.6305
2	0.9195	0.4084	0.8385	0.8806
3	0.7621	0.6408	0.8680	0.5161
4	0.5414	0.5766	0.7720	0.6773
5	0.3393	0.5227	0.4189	0.5568
6	0.8966	0.7339	0.5527	0.5106

7	0.8216	0.4522	0.8849	0.6252
8	0.6737	0.5484	0.6618	0.4577
9	0.8515	0.5695	0.5106	0.5033
10	0.4961	0.5549	0.8916	0.5920
MEAN	0.6896	0.5521	0.6870	0.5950
SD	0.1932		0.1863	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9250	0.5588	0.3952	0.7392
2	0.8335	0.3616	0.7614	0.5409
3	0.5260	0.6499	0.6834	0.4846
4	0.6168	0.4932	0.8907	0.5989
5	0.6862	0.7345	0.7837	0.5473
6	0.8563	0.6141	0.8204	0.6559
7	0.8078	0.4956	0.4667	0.5770
8	0.8353	0.6151	0.5521	0.6381
9	0.9133	0.3068	0.6438	0.6426
10	0.6191	0.5434	0.8538	0.4893
MEAN	0.7619	0.5373	0.6851	0.5914
SD	0.1390		0.1686	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7115	0.5463	0.0312	0.6676	0.5979	0.0200

Overall Model					
Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7382	0.5251	0.0262	0.6770	0.5811	0.0299

Appendix D

Complete results of exponential kernel RI-Score model

2MB3 model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8610	0.2885	0.8930	0.4136
2	0.8882	0.4563	0.8332	0.5131
3	0.9271	0.3501	0.8179	0.5275
4	0.5559	0.4494	0.8852	0.3998
5	0.8917	0.3690	0.8357	0.3756
6	0.9199	0.5271	0.7150	0.5602
7	0.5818	0.6506	0.8503	0.4736
8	0.4139	0.5376	0.8227	0.4838
9	0.8898	0.3879	0.7550	0.3989
10	0.8084	0.5203	0.1321	0.8974
MEAN	0.7738	0.4537	0.7540	0.5044
SD	0.1850		0.2251	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8337	0.5489	0.8454	0.4624
2	0.9517	0.3125	0.8211	0.4654
3	0.7938	0.5591	0.8804	0.3155
4	0.7808	0.5913	0.7717	0.4637
5	0.0671	0.5686	0.2625	0.4617
6	0.9169	0.3910	0.8068	0.3927
7	0.8108	0.4142	0.9582	0.4735
8	0.7813	0.4725	0.7495	0.6147
9	0.8538	0.3263	0.6251	0.5924
10	0.8442	0.3873	0.6871	0.6156
MEAN	0.7634	0.4572	0.7408	0.4858
SD	0.2511		0.1928	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9384	0.2211	0.8766	0.4424
2	0.8395	0.4170	0.7639	0.3547
3	0.8369	0.4825	0.8883	0.3546
4	0.8059	0.5447	0.8190	0.4683
5	0.8883	0.3417	0.8008	0.4152
6	0.8720	0.4614	0.8052	0.7009

7	0.7887	0.5135	0.7095	0.4561
8	0.6210	0.5432	0.6303	0.5750
9	0.7971	0.5663	0.8523	0.5022
10	0.7555	0.5262	0.8362	0.4963
MEAN	0.8143	0.4618	0.7982	0.4766
SD	0.0864		0.0791	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7120	0.4910	0.8569	0.4694
2	0.8791	0.5110	0.7419	0.4858
3	0.9221	0.4157	0.7953	0.5378
4	0.8087	0.4413	0.7563	0.4837
5	0.7780	0.5450	0.6590	0.5429
6	0.5289	0.8419	0.9323	0.4963
7	0.7230	0.4249	0.9027	0.3930
8	0.6245	0.6823	0.7909	0.5077
9	0.9554	0.2764	0.8524	0.4055
10	0.6027	0.3786	0.5379	0.6045
MEAN	0.7534	0.5008	0.7825	0.4927
SD	0.1418		0.1178	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8869	0.4232	0.8782	0.4934
2	0.8063	0.5435	0.6040	0.5563
3	0.8431	0.3978	0.6430	0.3605
4	0.9200	0.5310	0.8676	0.4585
5	0.8356	0.4174	0.7100	0.4885
6	0.3628	0.5183	0.3488	0.5567
7	0.5079	0.4993	0.5983	0.3270
8	0.9166	0.4274	0.8818	0.5187
9	0.8020	0.3908	0.7623	0.6536
10	0.7444	0.4740	0.7685	0.6205
MEAN	0.7625	0.4623	0.7062	0.5034
SD	0.1839		0.1658	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8417	0.5788	0.6741	0.5660
2	0.6673	0.5345	0.4620	0.4889
3	0.5755	0.3316	0.8894	0.4328

4	0.8049	0.5316	0.4288	0.4370
5	0.8775	0.3995	0.3513	0.4780
6	0.9391	0.4060	0.9443	0.4212
7	0.8060	0.4798	0.8598	0.2599
8	0.9130	0.4141	0.8750	0.4300
9	0.8371	0.3847	0.6549	0.6353
10	0.7907	0.3942	0.9554	0.5897
MEAN	0.8053	0.4455	0.7095	0.4739
SD	0.1101		0.2287	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8589	0.4862	0.7078	0.5399
2	0.7603	0.3306	0.4111	0.4482
3	0.7731	0.5788	0.9056	0.3910
4	0.8212	0.4071	0.9396	0.5600
5	0.7799	0.5434	0.7705	0.5944
6	0.6902	0.4423	0.8120	0.4128
7	0.5069	0.5651	0.6463	0.5937
8	0.6868	0.7900	0.6982	0.5354
9	0.9411	0.3834	0.9123	0.3343
10	0.8880	0.2997	0.7637	0.4355
MEAN	0.7706	0.4827	0.7567	0.4845
SD	0.1231		0.1565	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8973	0.4225	0.5756	0.4644
2	0.9572	0.6437	0.6302	0.3526
3	0.5530	0.3736	0.9574	0.4862
4	0.7308	0.3633	0.7274	0.4682
5	0.4453	0.7556	0.8236	0.5696
6	0.6585	0.4114	0.6824	0.4445
7	0.7579	0.5481	0.7366	0.5556
8	0.5312	0.6727	0.9212	0.4022
9	0.9634	0.3098	0.4164	0.6247
10	0.7891	0.4736	0.7904	0.6300
MEAN	0.7284	0.4974	0.7261	0.4998
SD	0.1812		0.1616	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7753	0.2940	0.9611	0.5089
2	0.7575	0.2708	0.7366	0.6383
3	0.8509	0.4014	0.8349	0.5014
4	0.7320	0.3536	0.8519	0.3626
5	0.8313	0.4846	0.9612	0.3162
6	0.8552	0.6151	0.8565	0.4360
7	0.8533	0.4298	0.1076	0.4916
8	0.7727	0.4678	0.7445	0.4116
9	0.7531	0.5964	0.5866	0.5538
10	0.7606	0.6209	0.8557	0.4973
MEAN	0.7942	0.4534	0.7496	0.4718
SD	0.0479		0.2514	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9076	0.3542	0.8917	0.5024
2	0.7659	0.6615	0.8501	0.4628
3	0.7292	0.5299	0.9136	0.3537
4	0.8565	0.3211	0.7887	0.3475
5	0.9103	0.5052	0.8256	0.4335
6	0.9088	0.4292	0.8951	0.2122
7	0.3888	0.5072	0.8235	0.5388
8	0.7611	0.4779	0.8542	0.3774
9	0.9214	0.3475	0.8153	0.6470
10	0.8448	0.3101	0.4832	0.5981
MEAN	0.7994	0.4444	0.8141	0.4473
SD	0.1607		0.1228	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7765	0.4659	0.0266	0.7538	0.1702	0.0359

2MCC model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8850	0.3957	0.9512	0.3990
2	0.7224	0.5833	0.8849	0.3754
3	0.8255	0.4046	0.3511	0.5616
4	0.2150	0.4228	0.9162	0.5124
5	0.7257	0.5016	0.6965	0.3923
6	0.8534	0.4829	0.6070	0.5562
7	0.8427	0.5703	0.8096	0.5316
8	0.8621	0.3271	0.6756	0.5013
9	0.9718	0.3061	0.8112	0.3975
10	0.6477	0.6730	0.7305	0.5310
MEAN	0.7551	0.4667	0.7434	0.4758
SD	0.2115		0.1768	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6914	0.4114	0.7925	0.4209
2	0.5883	0.5489	0.8769	0.5623
3	0.8801	0.4020	0.8359	0.4588
4	0.8388	0.6725	0.8315	0.5362
5	0.7411	0.2832	0.6312	0.5602
6	0.7505	0.5752	0.6812	0.5328
7	0.7976	0.4852	0.8843	0.3942
8	0.8786	0.5213	0.8903	0.3781
9	0.7339	0.5269	0.7001	0.4911
10	0.9041	0.3261	0.5910	0.4369
MEAN	0.7804	0.4753	0.7715	0.4772
SD	0.0989		0.1114	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7795	0.3936	0.8391	0.4198
2	0.8315	0.4818	0.7581	0.4652
3	0.8089	0.5138	0.6540	0.4601
4	0.8596	0.3773	0.8386	0.4680
5	0.4050	0.4802	0.9156	0.4482
6	0.8036	0.5171	0.9136	0.5874
7	0.8846	0.3799	0.8255	0.5983
8	0.6641	0.5760	0.8309	0.3821
9	0.8472	0.5610	0.5654	0.5536
10	0.8993	0.5033	0.6750	0.4302

MEAN	0.7783	0.4784	0.7816	0.4813
SD	0.1469		0.1161	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8308	0.4744	0.4641	0.3933
2	0.6809	0.5434	0.9157	0.3654
3	0.7040	0.4852	0.6200	0.4132
4	0.9246	0.2482	0.8845	0.4767
5	0.7843	0.6227	0.8160	0.5220
6	0.7403	0.4350	0.7231	0.6557
7	0.8238	0.6323	0.7172	0.5680
8	0.8635	0.4396	0.8526	0.4430
9	0.9212	0.4030	0.9491	0.4058
10	0.4700	0.5253	0.7130	0.5084
MEAN	0.7743	0.4809	0.7655	0.4752
SD	0.1355		0.1489	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8922	0.3948	0.7770	0.4597
2	0.7901	0.5664	0.8010	0.5001
3	0.8613	0.3236	0.7193	0.5614
4	0.9017	0.4469	0.8975	0.4511
5	0.8234	0.5531	0.9247	0.4364
6	0.8288	0.5002	0.6455	0.6038
7	0.6807	0.4098	0.8555	0.4953
8	0.5099	0.3113	0.5352	0.4617
9	0.8126	0.5730	0.8687	0.3824
10	0.7188	0.6403	0.0440	0.4181
MEAN	0.7820	0.4719	0.7068	0.4770
SD	0.1183		0.2622	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7683	0.3989	0.5363	0.5904
2	0.8068	0.3986	0.7805	0.2436
3	0.9226	0.4422	0.5618	0.5574
4	0.6873	0.5669	0.9425	0.3293
5	0.9509	0.3541	0.8147	0.5090
6	0.7752	0.6336	0.4872	0.4780
7	0.5231	0.4452	0.8563	0.3855
8	0.8067	0.4741	0.8436	0.6191

9	0.7845	0.6157	0.8574	0.6790
10	0.7283	0.4908	0.8962	0.4231
MEAN	0.7754	0.4820	0.7576	0.4814
SD	0.1194		0.1649	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7493	0.4076	0.5597	0.6298
2	0.7450	0.6514	0.6747	0.5029
3	0.9020	0.4307	0.7445	0.3919
4	0.6982	0.3961	0.9177	0.6630
5	0.5660	0.4599	0.7397	0.4583
6	0.7155	0.4790	0.5993	0.5306
7	0.8302	0.4898	0.8849	0.3362
8	0.9455	0.3632	0.8969	0.7379
9	0.7257	0.3997	0.6493	0.4311
10	0.9218	0.3554	0.9660	0.3069
MEAN	0.7799	0.4433	0.7633	0.4989
SD	0.1186		0.1444	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8115	0.4359	0.5911	0.3268
2	0.6421	0.4330	0.7977	0.4991
3	0.9561	0.2206	0.5303	0.9386
4	0.9087	0.6255	0.7158	0.5322
5	0.8362	0.3529	0.9077	0.3835
6	0.8775	0.4888	0.9260	0.4376
7	0.9168	0.3179	0.7780	0.5453
8	0.6352	0.4877	0.9678	0.3369
9	0.7076	0.6462	0.8492	0.5606
10	0.4963	0.6046	0.8068	0.4221
MEAN	0.7788	0.4613	0.7870	0.4983
SD	0.1512		0.1416	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.4374	0.4628	0.8891	0.4172
2	0.8395	0.5122	0.7615	0.3228
3	0.8579	0.4828	0.8356	0.2195
4	0.7858	0.5365	0.7431	0.4545
5	0.5965	0.5917	0.9327	0.5099
6	0.8577	0.4932	0.6051	0.5478

7	0.5304	0.4297	0.7876	0.6334
8	0.9430	0.3198	0.8684	0.7228
9	0.7294	0.4906	0.2743	0.9853
10	0.8922	0.4024	0.5922	0.4293
MEAN	0.7470	0.4722	0.7290	0.5243
SD	0.1699		0.1953	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7601	0.6234	0.8045	0.3418
2	0.7207	0.3031	0.9322	0.5417
3	0.1714	0.4398	0.7401	0.5075
4	0.7733	0.3568	0.8035	0.5039
5	0.8222	0.5597	0.7838	0.6520
6	0.7944	0.4842	0.7173	0.4534
7	0.7389	0.4220	0.9277	0.3643
8	0.7819	0.4968	0.5320	0.3872
9	0.8134	0.6494	0.5340	0.7108
10	0.7792	0.5178	0.7009	0.4526
MEAN	0.7155	0.4853	0.7476	0.4915
SD	0.1937		0.1372	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7667	0.4717	0.0215	0.7553	0.4881	0.0245

4G0F Model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.4608	0.6068	0.8230	0.4279
2	0.5717	0.6373	0.3324	0.4962
3	0.6305	0.7823	0.1755	0.6631
4	0.2846	0.6493	0.6652	0.4058
5	0.7965	0.6041	0.7229	0.3059
6	0.7767	0.5119	0.5162	0.8684
7	0.8691	0.4075	0.7189	0.7353
8	0.5760	0.7831	0.9381	0.4776
9	0.9117	0.4166	0.9090	0.3764
10	0.7419	0.6711	0.6056	0.5840
MEAN	0.6619	0.6070	0.6407	0.5341
SD	0.1950		0.2439	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.5283	0.7754	0.5167	0.4699
2	0.1124	0.8442	0.7238	0.5581
3	0.8414	0.4904	0.7878	0.5148
4	0.8770	0.3066	0.6289	0.7246
5	0.6413	0.6548	0.8813	0.4070
6	0.0410	0.5628	0.8449	0.4389
7	0.9235	0.4233	0.7374	0.5733
8	0.6190	0.6688	0.6726	0.7312
9	0.8710	0.4724	0.7802	0.5488
10	0.8732	0.3095	0.9103	0.3451
MEAN	0.6328	0.5508	0.7484	0.5312
SD	0.3222		0.1203	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6909	0.7126	0.8321	0.4982
2	0.6494	0.6967	0.7802	0.6896
3	0.6670	0.4462	0.5773	0.5868
4	0.7092	0.5128	0.6210	0.5810
5	0.5957	0.7068	0.8659	0.4540
6	0.8851	0.2393	0.4247	0.5825
7	0.5089	0.3944	0.7917	0.5456
8	0.5788	0.6831	0.6637	0.4824
9	0.4406	0.8771	0.7989	0.4449
10	0.5533	0.8076	0.9036	0.5286

MEAN	0.6279	0.6077	0.7259	0.5394
SD	0.1234		0.1501	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7351	0.6115	0.9024	0.4200
2	0.7401	0.5001	0.9533	0.2861
3	0.1929	0.4244	0.3794	0.7869
4	0.9030	0.4914	0.9090	0.4385
5	0.4854	0.5112	0.8393	0.4405
6	0.7463	0.7295	0.8383	0.3983
7	0.6825	0.5200	0.5669	0.5568
8	0.2560	0.8450	0.4664	0.5251
9	0.6381	0.5453	0.5634	0.7796
10	0.9003	0.3447	0.7482	0.6003
MEAN	0.6280	0.5523	0.7166	0.5232
SD	0.2447		0.2056	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8885	0.3296	0.7116	0.3548
2	0.4637	0.4910	0.8171	0.5257
3	0.8179	0.6290	0.7135	0.2530
4	0.6405	0.6172	0.6249	0.4982
5	0.1210	0.7105	0.0523	0.6959
6	0.7241	0.6081	0.7534	0.4852
7	0.7746	0.4913	0.7482	0.6882
8	0.6157	0.8232	0.8532	0.5904
9	0.5615	0.7281	0.6269	0.8042
10	0.6639	0.7025	0.6123	0.4681
MEAN	0.6271	0.6131	0.6513	0.5364
SD	0.2172		0.2254	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9311	0.4099	0.7558	0.3084
2	0.8710	0.3820	0.8347	0.5262
3	0.2432	0.8304	0.9664	0.3741
4	0.6924	0.5752	0.7532	0.7244
5	0.6358	0.6772	0.5288	0.6131
6	0.7372	0.5093	0.9247	0.5177
7	0.8320	0.4786	0.4683	0.7050
8	0.5667	0.7762	0.2529	0.5337

9	0.9565	0.2303	0.7842	0.6063
10	0.5953	0.8069	0.7710	0.3451
MEAN	0.7061	0.5676	0.7040	0.5254
SD	0.2130		0.2212	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.4326	0.8318	0.5571	0.5582
2	0.8513	0.4567	0.3538	0.6156
3	0.6537	0.6655	0.7855	0.4424
4	0.1717	0.7074	0.6624	0.8259
5	0.4610	0.6579	0.6921	0.6961
6	0.8200	0.5195	0.8423	0.3108
7	0.9248	0.3758	0.6942	0.6001
8	0.9096	0.3508	0.5873	0.6027
9	0.6962	0.5582	0.8459	0.3955
10	0.8682	0.6235	0.4775	0.5375
MEAN	0.6789	0.5747	0.6498	0.5585
SD	0.2507		0.1590	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7807	0.4522	0.7613	0.4568
2	0.8265	0.4957	0.7015	0.6431
3	0.3962	0.6288	0.2803	0.6756
4	0.6462	0.5188	0.3665	0.7153
5	0.6221	0.7969	0.4153	0.4921
6	0.7299	0.5534	0.8569	0.4096
7	0.7407	0.6823	0.9228	0.5614
8	0.7692	0.5768	0.6972	0.4535
9	0.8478	0.3831	0.5364	0.8492
10	0.5231	0.5482	0.7689	0.4664
MEAN	0.6882	0.5636	0.6307	0.5723
SD	0.1425		0.2186	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8110	0.4535	0.4258	0.8808
2	0.7984	0.4672	0.5001	0.3455
3	0.8575	0.4942	0.8505	0.5907
4	0.5871	0.5827	0.9061	0.3698
5	0.6875	0.5963	0.6962	0.5995
6	0.7318	0.5057	0.4730	0.6368

7	0.7407	0.8141	0.6699	0.7769
8	0.8502	0.4446	0.7517	0.4715
9	0.3475	0.8303	0.3459	0.5349
10	0.2302	0.5644	0.9119	0.3500
MEAN	0.6642	0.5753	0.6531	0.5556
SD	0.2153		0.2065	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8743	0.4992	0.7416	0.5493
2	0.7261	0.5215	0.7946	0.4563
3	0.4978	0.4851	0.7562	0.6367
4	0.8278	0.4717	0.4669	0.6179
5	0.7053	0.6436	0.8041	0.5928
6	0.6507	0.6589	0.8413	0.5716
7	0.8660	0.4997	0.8334	0.4625
8	0.5528	0.7100	0.0471	0.7211
9	0.2729	0.4544	0.5349	0.2266
10	0.8379	0.3048	0.2426	0.3151
MEAN	0.6812	0.5249	0.6063	0.5150
SD	0.1937		0.2772	

Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.6596	0.5737	0.0291	0.6727	0.5391	0.0472

Overall Model					
Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7343	0.5038	0.0257	0.7273	0.5037	0.0359

Appendix E

Complete results of Lorentz kernel RI-Score model

2MB3 model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8169	0.4849	0.7561	0.4428
2	0.6860	0.5725	0.8614	0.5381
3	0.7213	0.5988	0.7541	0.3637
4	0.6953	0.5334	0.9342	0.3052
5	0.6357	0.3006	0.8886	0.3972
6	0.8268	0.5472	0.9456	0.3012
7	0.9245	0.3398	0.7301	0.4402
8	0.9330	0.3970	0.6738	0.6193
9	0.3816	0.3362	0.5658	0.8810
10	0.6561	0.5313	0.3222	0.6443
MEAN	0.7277	0.4642	0.7432	0.4933
SD	0.1615		0.1905	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9259	0.6287	0.7941	0.4936
2	0.9140	0.3736	0.7681	0.6479
3	0.7904	0.5128	0.9571	0.3296
4	0.8850	0.3724	0.6306	0.5014
5	0.1587	0.5560	0.8710	0.4642
6	0.5906	0.4172	0.7962	0.4834
7	0.8302	0.4525	0.3837	0.5563
8	0.5808	0.7216	0.5609	0.4710
9	0.8009	0.3014	0.3714	0.3471
10	0.6543	0.3199	0.7954	0.5370
MEAN	0.7131	0.4656	0.6928	0.4832
SD	0.2324		0.1997	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9107	0.3841	0.9516	0.3283
2	0.6664	0.3825	0.1776	0.6524
3	0.9162	0.2957	0.8660	0.4083
4	0.7405	0.5095	0.7129	0.5003
5	0.8741	0.4498	0.8410	0.5782
6	0.6736	0.5174	0.4663	0.5213

7	0.8227	0.5668	0.4683	0.3892
8	0.8952	0.4097	0.6908	0.4783
9	0.6959	0.6838	0.8429	0.4661
10	0.7665	0.4947	0.8530	0.5680
MEAN	0.7962	0.4694	0.6870	0.4890
SD	0.1000		0.2437	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8187	0.3444	0.7079	0.5803
2	0.9001	0.3314	0.9206	0.3341
3	0.3362	0.6905	0.9517	0.3805
4	0.4590	0.2808	0.8881	0.3045
5	0.3559	0.6104	0.3709	0.6419
6	0.9180	0.5871	0.7976	0.5037
7	0.8231	0.4798	0.8755	0.4811
8	0.8192	0.5170	0.7914	0.5290
9	0.8892	0.3453	0.8846	0.4430
10	0.8385	0.4587	0.7056	0.4215
MEAN	0.7158	0.4645	0.7894	0.4620
SD	0.2339		0.1695	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7654	0.4106	0.9491	0.3681
2	0.6855	0.6260	0.8520	0.4057
3	0.8723	0.4368	0.7054	0.4755
4	0.9130	0.4713	0.5955	0.6970
5	0.8384	0.2614	0.8982	0.3713
6	0.8834	0.4501	0.7762	0.2963
7	0.8263	0.4368	0.8930	0.3539
8	0.2938	0.5844	0.7787	0.4761
9	0.8218	0.5889	0.8661	0.4909
10	0.7997	0.5015	0.5328	0.5277
MEAN	0.7700	0.4768	0.7847	0.4463
SD	0.1792		0.1368	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9248	0.4405	0.6675	0.4366
2	0.9029	0.3961	0.8808	0.3708
3	0.9008	0.3764	0.9720	0.1811

4	0.6180	0.3627	0.7978	0.4499
5	0.9068	0.5654	0.9365	0.4872
6	0.8490	0.4465	0.5679	0.4859
7	0.3830	0.5987	0.8469	0.4861
8	0.2901	0.4038	0.6300	0.4106
9	0.9194	0.3125	0.9559	0.3976
10	0.1333	0.6882	0.6317	0.8160
MEAN	0.6828	0.4591	0.7887	0.4522
SD	0.3053		0.1523	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7705	0.2656	0.7661	0.4275
2	0.6668	0.3150	0.8687	0.4704
3	0.9127	0.3610	0.4341	0.7142
4	0.8451	0.4787	0.9066	0.3970
5	0.9116	0.3778	0.8707	0.3913
6	0.8939	0.4771	0.8562	0.2269
7	0.7534	0.5855	0.8104	0.6321
8	0.8399	0.4485	0.9195	0.2908
9	0.7245	0.6724	0.4481	0.5608
10	0.3430	0.7010	0.9381	0.2741
MEAN	0.7661	0.4683	0.7818	0.4385
SD	0.1704		0.1866	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7619	0.3964	0.9112	0.5349
2	0.8318	0.4265	0.7822	0.5561
3	0.9496	0.3099	0.7202	0.5072
4	0.9016	0.4758	0.5047	0.5110
5	0.7579	0.4784	0.8084	0.4449
6	0.9766	0.2166	0.9401	0.4691
7	0.7114	0.5899	0.7327	0.4168
8	0.7420	0.8156	0.8223	0.4930
9	0.5017	0.6701	0.6912	0.4352
10	0.9221	0.3781	0.7686	0.4694
MEAN	0.8056	0.4757	0.7681	0.4838
SD	0.1426		0.1219	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8576	0.4152	0.9278	0.4447
2	0.7885	0.5645	0.0314	0.5449
3	0.8013	0.5155	0.9593	0.5183
4	0.7137	0.6246	0.5326	0.5743
5	0.9054	0.3667	0.5916	0.3527
6	0.8519	0.4502	0.6880	0.5380
7	0.7820	0.5455	0.9470	0.4062
8	0.5233	0.4142	0.5913	0.4033
9	0.9180	0.3473	0.3580	0.5533
10	0.7798	0.3454	0.8224	0.3664
MEAN	0.7921	0.4589	0.6449	0.4702
SD	0.1132		0.2945	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6473	0.6026	0.9002	0.5359
2	0.8143	0.6735	0.6628	0.7202
3	0.7948	0.4829	0.6557	0.5195
4	0.8301	0.3690	0.8422	0.4269
5	0.7318	0.4625	0.2543	0.4282
6	0.9469	0.2849	0.9437	0.3024
7	0.7819	0.3164	0.7975	0.3408
8	0.8284	0.5378	0.8477	0.4897
9	0.7774	0.5134	0.7839	0.3389
10	0.4523	0.5100	0.4061	0.5741
MEAN	0.7605	0.4753	0.7094	0.4677
SD	0.1323		0.2225	

Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7530	0.4678	0.0411	0.7390	0.4686	0.0520

2MCC model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9260	0.3100	0.3915	0.8301
2	0.7982	0.5634	0.8801	0.3838
3	0.8120	0.3802	0.8291	0.4701
4	0.7263	0.5168	0.9304	0.4423
5	0.7790	0.5054	0.8780	0.4147
6	0.8834	0.4731	0.3558	0.3434
7	0.7209	0.5920	0.6641	0.6133
8	0.8403	0.5615	0.8782	0.4271
9	0.9255	0.4462	0.8488	0.4802
10	0.4841	0.4896	0.9018	0.4439
MEAN	0.7896	0.4838	0.7558	0.4849
SD	0.1296		0.2140	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.5565	0.7177	0.8939	0.3635
2	0.8784	0.4148	0.5644	0.5435
3	0.9202	0.3612	0.8776	0.3025
4	0.8389	0.4774	0.7137	1.0133
5	0.8817	0.3736	0.8600	0.3391
6	0.8604	0.5800	0.8657	0.3957
7	0.6975	0.5414	0.6612	0.6107
8	0.8103	0.7446	0.7042	0.5546
9	0.6441	0.3884	0.8378	0.4288
10	0.6398	0.3515	0.8673	0.3779
MEAN	0.7728	0.4951	0.7846	0.4930
SD	0.1269		0.1144	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8620	0.6387	0.8192	0.4772
2	0.8570	0.4209	0.9044	0.5104
3	0.9107	0.3322	0.6245	0.3904
4	0.4284	0.4551	0.8908	0.4254
5	0.9754	0.4230	0.7904	0.6133
6	0.8055	0.4404	0.8009	0.3896
7	0.7033	0.5084	0.8085	0.5024
8	0.9261	0.4044	0.8957	0.3348
9	0.5518	0.6454	0.1769	0.8060
10	0.7329	0.5413	0.8463	0.4070

MEAN	0.7753	0.4810	0.7558	0.4857
SD	0.1743		0.2187	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9057	0.2036	0.8549	0.5808
2	0.7796	0.5527	0.7741	0.3674
3	0.8970	0.3996	0.8642	0.3935
4	0.7526	0.7452	0.9400	0.3881
5	0.6990	0.6538	0.2803	0.5728
6	0.8988	0.3827	0.5894	0.7426
7	0.6666	0.4537	0.6278	0.3608
8	0.6535	0.4168	0.8754	0.4320
9	0.8377	0.3783	0.8469	0.3384
10	0.8895	0.3255	0.9322	0.4418
MEAN	0.7980	0.4512	0.7585	0.4618
SD	0.1010		0.2054	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6711	0.4658	0.7590	0.4601
2	0.8897	0.3682	0.8551	0.5432
3	0.6068	0.6523	0.8427	0.6415
4	0.8616	0.3967	0.9107	0.4180
5	0.9195	0.3140	0.8081	0.4891
6	0.8898	0.3925	0.6131	0.3769
7	0.9281	0.2778	0.6625	0.3710
8	0.8965	0.3391	0.8140	0.4521
9	0.7827	0.4511	0.4631	0.4729
10	0.7038	0.7198	0.9447	0.3755
MEAN	0.8150	0.4377	0.7673	0.4600
SD	0.1160		0.1478	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7694	0.5527	0.5794	0.4061
2	0.8319	0.3652	0.8313	0.3961
3	0.9649	0.2707	0.6814	0.5259
4	0.6124	0.7797	0.9649	0.3308
5	0.8371	0.3757	0.7761	0.5075
6	0.8678	0.3294	0.6432	0.6234
7	0.7388	0.6469	0.7854	0.5144
8	0.8434	0.6424	0.8287	0.3607

9	0.0588	0.4075	0.8325	0.3479
10	0.8384	0.4581	0.9100	0.4113
MEAN	0.7363	0.4828	0.7833	0.4424
SD	0.2552		0.1190	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.3781	0.7072	0.4163	0.4356
2	0.9555	0.3627	0.8960	0.3317
3	0.8633	0.2164	0.8722	0.4128
4	0.8044	0.5386	0.9362	0.3164
5	0.7079	0.6539	0.9373	0.3432
6	0.8933	0.4135	0.6600	0.5212
7	0.7865	0.3906	0.8048	0.5440
8	0.9096	0.2737	0.9682	0.5492
9	0.7491	0.6672	0.7651	0.5643
10	0.8306	0.5565	0.7146	0.5999
MEAN	0.7878	0.4780	0.7971	0.4618
SD	0.1625		0.1685	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8594	0.4066	0.7606	0.5394
2	0.8953	0.3681	0.7922	0.4543
3	0.6511	0.6175	0.8107	0.3604
4	0.7508	0.3074	0.7704	0.4705
5	0.7848	0.5144	0.7509	0.4717
6	0.8762	0.4580	0.6034	0.4703
7	0.8647	0.4577	0.6694	0.5452
8	0.8276	0.5204	0.8125	0.4582
9	0.8393	0.5144	0.8678	0.6542
10	0.4864	0.3335	0.9759	0.3972
MEAN	0.7836	0.4498	0.7814	0.4821
SD	0.1271		0.1016	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8798	0.5670	0.8474	0.4787
2	0.7921	0.6044	0.9418	0.3677
3	0.6267	0.4332	0.7360	0.6298
4	0.8640	0.3739	0.8678	0.3835
5	0.6086	0.4727	0.4534	0.4774
6	0.8880	0.3447	0.8611	0.4379

7	0.4617	0.3418	0.5916	0.5149
8	0.7956	0.4561	0.7295	0.4810
9	0.8497	0.5025	0.8407	0.3635
10	0.8128	0.4111	0.9403	0.3703
MEAN	0.7579	0.4507	0.7810	0.4505
SD	0.1431		0.1567	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8398	0.4656	0.8302	0.4448
2	0.9371	0.4029	0.1474	0.5056
3	0.8584	0.4285	0.8112	0.3986
4	0.7420	0.5361	0.8722	0.5651
5	0.7026	0.3571	0.8480	0.4946
6	0.8468	0.3113	0.8423	0.4847
7	0.0162	0.5060	0.8363	0.4403
8	0.8949	0.4853	0.8407	0.4105
9	0.5211	0.5562	0.8904	0.4564
10	0.5049	0.4288	0.8253	0.4862
MEAN	0.6864	0.4478	0.7744	0.4687
SD	0.2785		0.2215	

Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7703	0.4658	0.0365	0.7739	0.4691	0.0141

4G0F Model

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.5909	0.5186	0.6714	0.3814
2	0.8089	0.5179	0.8448	0.3318
3	0.8319	0.4818	0.5475	0.4910
4	0.0195	0.5331	0.8436	0.6025
5	0.8684	0.5909	0.9727	0.4010
6	0.9625	0.2270	0.9521	0.5578
7	0.8571	0.3598	0.7375	0.5300
8	0.5883	0.5523	0.5751	0.4772
9	0.8612	0.4422	0.6158	0.6606
10	0.8917	0.3865	0.6374	0.3527
MEAN	0.7280	0.4610	0.7398	0.4786
SD	0.2776		0.1548	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7715	0.5032	0.8409	0.5447
2	0.8256	0.5664	0.8275	0.4786
3	0.8440	0.3834	0.9018	0.3506
4	0.4083	0.5704	0.6688	0.3864
5	0.9479	0.2979	0.8569	0.5424
6	0.8611	0.4622	0.8809	0.3406
7	0.8244	0.4274	0.6343	0.3469
8	0.7610	0.5615	0.9067	0.3345
9	0.6037	0.4297	0.5743	0.5661
10	0.9009	0.2995	0.4827	0.8674
MEAN	0.7748	0.4502	0.7575	0.4758
SD	0.1588		0.1535	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6318	0.4339	0.9491	0.3874
2	0.9391	0.3848	0.4520	0.8931
3	0.6278	0.6280	0.9033	0.2948
4	0.7458	0.4409	0.9155	0.3461
5	0.9186	0.5965	0.9496	0.3633
6	0.9327	0.3405	0.7460	0.4177
7	0.3655	0.4941	0.5574	0.6523
8	0.7783	0.5608	0.9049	0.5727
9	0.5452	0.3386	0.6938	0.5084
10	0.9051	0.4362	0.9203	0.3626

MEAN	0.7390	0.4654	0.7992	0.4798
SD	0.1943		0.1790	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8469	0.4618	0.7812	0.4133
2	0.9175	0.2889	0.7591	0.6903
3	0.3549	0.4979	0.6458	0.4777
4	0.8306	0.4474	0.9192	0.3214
5	0.6313	0.4354	0.7257	0.6158
6	0.9500	0.3906	0.9715	0.3490
7	0.9579	0.4168	0.6678	0.5190
8	0.9236	0.4080	0.9318	0.4779
9	0.6971	0.4732	0.9025	0.3090
10	0.2627	0.6798	0.4411	0.4647
MEAN	0.7372	0.4500	0.7746	0.4638
SD	0.2508		0.1642	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.6911	0.5782	0.7991	0.5173
2	0.9016	0.3848	0.9525	0.5076
3	0.8933	0.3011	0.7972	0.5198
4	0.9205	0.3047	0.8964	0.3796
5	0.7298	0.4860	0.9020	0.4377
6	0.5231	0.4762	0.7187	0.4691
7	0.9790	0.4088	0.3538	0.3619
8	0.8257	0.6121	0.8177	0.5104
9	0.2284	0.6767	0.2101	0.5059
10	0.8109	0.4733	0.5639	0.5670
MEAN	0.7503	0.4702	0.7011	0.4776
SD	0.2265		0.2483	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.9389	0.4289	0.7171	0.2929
2	0.7110	0.6123	0.7461	0.5771
3	0.8602	0.4740	0.5410	0.5932
4	0.6003	0.3397	0.5967	0.6053
5	0.7911	0.3362	0.9211	0.3252
6	0.6332	0.5194	0.7099	0.5637
7	0.4691	0.6677	0.9271	0.4869
8	0.8519	0.4785	0.8983	0.4510

9	0.8983	0.4246	0.9312	0.4090
10	0.8479	0.4440	0.8621	0.2922
MEAN	0.7602	0.4725	0.7851	0.4597
SD	0.1516		0.1436	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.4170	0.8834	0.6478	0.6481
2	0.7326	0.5665	0.9388	0.4857
3	0.7774	0.5842	0.8267	0.3257
4	0.9231	0.3189	0.6350	0.4674
5	0.7634	0.5874	0.6039	0.6620
6	0.8286	0.4949	0.9431	0.4017
7	0.9012	0.4691	0.9087	0.3055
8	0.6623	0.3630	0.7359	0.4174
9	0.4388	0.3301	0.7605	0.4911
10	0.8391	0.4533	0.8637	0.5850
MEAN	0.7283	0.5051	0.7864	0.4790
SD	0.1760		0.1287	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7942	0.5659	0.5866	0.5958
2	0.7840	0.3521	0.8876	0.5347
3	0.9465	0.3231	0.6986	0.5253
4	0.3894	0.6404	0.8793	0.2896
5	0.7135	0.4206	0.8489	0.4792
6	0.8526	0.4533	0.7489	0.3089
7	0.9293	0.4023	0.8815	0.4029
8	0.7851	0.5160	0.5019	0.4401
9	0.9060	0.4072	0.6390	0.6692
10	0.6745	0.8415	0.8639	0.4459
MEAN	0.7775	0.4922	0.7536	0.4692
SD	0.1631		0.1409	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.7901	0.4496	0.8215	0.6407
2	0.8873	0.4180	0.8787	0.3246
3	0.8085	0.3868	0.9524	0.3495
4	0.8680	0.5144	0.5475	0.5183
5	0.8043	0.3736	0.9349	0.3755
6	0.0804	0.5131	0.7735	0.4927

7	0.8175	0.4806	0.6407	0.4289
8	0.8855	0.3200	0.8997	0.3852
9	0.8375	0.6356	0.9148	0.5129
10	0.6138	0.5828	0.4118	0.6404
MEAN	0.7393	0.4675	0.7776	0.4669
SD	0.2443		0.1846	

	Gradient Boosting Tree		Random Forest	
	Correlation	RMSE	Correlation	RMSE
1	0.8546	0.3383	0.8571	0.3154
2	0.9088	0.4366	0.8268	0.5013
3	0.7960	0.5377	0.7981	0.3478
4	0.7205	0.4394	0.8745	0.4596
5	0.7897	0.5834	0.8187	0.4775
6	0.8107	0.5100	0.9295	0.3620
7	0.9069	0.4151	0.8122	0.4292
8	0.8173	0.3973	0.8816	0.5576
9	0.8849	0.3806	0.8473	0.3598
10	0.6234	0.6392	0.3584	0.7301
MEAN	0.8113	0.4678	0.8004	0.4540
SD	0.0883		0.1601	

Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7546	0.4702	0.0266	0.7675	0.4704	0.0305

Overall Model					
Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7593	0.4679	0.0348	0.7601	0.4694	0.0322

Appendix F

Complete list of ligands in their SMILES notation

Index	Formula	SMILES
G4L0021	C ₃₄ H ₄₆ N ₄ O ₆	<chem>C[N@@+]5(CCC(=O)Nc4ccc3c(=O)c2cc(NC(=O)CC[N@@+]1(C)CCCC[C@@H]1CO)ccc2c(=O)c3c4)CCCC[C@@H]5CO</chem>
G4L0025	C ₂₈ H ₃₂ N ₄ O ₆	<chem>O=C(CCN1CCOCC1)Nc4ccc(NC(=O)CCN2CCOCC2)c5c(=O)c3ccccc3c(=O)c45</chem>
G4L0027	C ₂₈ H ₃₆ N ₄ O ₄	<chem>CCN(CC)CCC(=O)Nc2ccc(NC(=O)CCN(CC)CC)c3c(=O)c1cccc1c(=O)c23</chem>
G4L0028	C ₃₀ H ₃₆ N ₄ O ₄	<chem>O=C(CCN1CCCCC1)Nc4cccc5c(=O)c3c(NC(=O)CCN2CCCCC2)cccc3c(=O)c45</chem>
G4L0031	C ₂₄ H ₂₈ N ₄ O ₄	<chem>CN(C)CCC(=O)Nc2cccc3c(=O)c1c(NC(=O)CCN(C)C)cccc1c(=O)c23</chem>
G4L0032	C ₂₈ H ₃₆ N ₄ O ₄	<chem>CCN(CC)CCC(=O)Nc2cccc3c(=O)c1c(NC(=O)CCN(CC)CC)cccc1c(=O)c23</chem>
G4L0033	C ₃₀ H ₃₆ N ₄ O ₄	<chem>O=C(CCN1CCCCC1)Nc4cccc5c(=O)c3cccc(NC(=O)CCN2CCCCC2)c3c(=O)c45</chem>
G4L0036	C ₂₄ H ₂₈ N ₄ O ₄	<chem>CN(C)CCC(=O)Nc2cccc3c(=O)c1cccc(NC(=O)CCN(C)C)c1c(=O)c23</chem>
G4L0037	C ₂₈ H ₃₆ N ₄ O ₄	<chem>CCN(CC)CCC(=O)Nc2cccc3c(=O)c1cccc(NC(=O)CCN(CC)CC)c1c(=O)c23</chem>
G4L0038	C ₃₀ H ₃₆ N ₄ O ₄	<chem>O=C(CCN1CCCCC1)Nc5ccc4c(=O)c3cc(NC(=O)CCN2CCCCC2)ccc3c(=O)c4c5</chem>
G4L0039	C ₂₈ H ₃₂ N ₄ O ₄	<chem>O=C(CCN1CCCCC1)Nc5ccc4c(=O)c3cc(NC(=O)CCN2CCCCC2)ccc3c(=O)c4c5</chem>
G4L0041	C ₂₄ H ₂₈ N ₄ O ₄	<chem>CN(C)CCC(=O)Nc3ccc2c(=O)c1cc(NC(=O)CCN(C)C)ccc1c(=O)c2c3</chem>
G4L0042	C ₂₈ H ₃₆ N ₄ O ₄	<chem>CCN(CC)CCC(=O)Nc3ccc2c(=O)c1cc(NC(=O)CCN(CC)CC)ccc1c(=O)c2c3</chem>
G4L0043	C ₃₀ H ₃₆ N ₄ O ₄	<chem>O=C(CCN1CCCCC1)Nc5ccc4c(=O)c3ccc(NC(=O)CCN2CCCCC2)cc3c(=O)c4c5</chem>
G4L0044	C ₂₈ H ₃₂ N ₄ O ₄	<chem>O=C(CCN1CCCCC1)Nc5ccc4c(=O)c3ccc(NC(=O)CCN2CCCCC2)cc3c(=O)c4c5</chem>
G4L0046	C ₂₄ H ₂₈ N ₄ O ₄	<chem>CN(C)CCC(=O)Nc3ccc2c(=O)c1ccc(NC(=O)CCN(C)C)cc1c(=O)c2c3</chem>
G4L0047	C ₂₈ H ₃₆ N ₄ O ₄	<chem>CCN(CC)CCC(=O)Nc3ccc2c(=O)c1ccc(NC(=O)CCN(CC)CC)cc1c(=O)c2c3</chem>
G4L0048	C ₃₂ H ₄₂ N ₄ O ₄	<chem>C[N+]5(CCC(=O)Nc3ccc(NC(=O)CC[N+]1(C)CCCCC1)c4c(=O)c2ccccc2c(=O)c34)CCCCC5</chem>
G4L0049	C ₃₀ H ₃₈ N ₄ O ₄	<chem>C[N+]5(CCC(=O)Nc3ccc(NC(=O)CC[N+]1(C)CCCCC1)c4c(=O)c2ccccc2c(=O)c34)CCCCC5</chem>

G4L0050	C ₃₀ H ₃₈ N ₄ O ₆	C[N+](C)(C)CCC(=O)Nc3ccc(NC(=O)CC[N+](C)CCOCC1)c4c(=O)c2cccc2c(=O)c34)CCOCC5
G4L0051	C ₂₆ H ₃₄ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc2ccc(NC(=O)CC[N+](C)(C)C)c3c(=O)c1cccc1c(=O)c23
G4L0052	C ₃₀ H ₄₂ N ₄ O ₄	CC[N+](C)(CC)CCC(=O)Nc2ccc(NC(=O)CC[N+](C)(CC)CC)c3c(=O)c1cccc1c(=O)c23
G4L0053	C ₃₂ H ₄₂ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc3cccc4c(=O)c2c(NC(=O)CC[N+](C)CCCCC1)cccc2c(=O)c34)CCCCC5
G4L0054	C ₃₀ H ₃₈ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc3cccc4c(=O)c2c(NC(=O)CC[N+](C)CCCC1)cccc2c(=O)c34)CCCCC5
G4L0055	C ₃₀ H ₃₈ N ₄ O ₆	C[N+](C)(C)CCC(=O)Nc3cccc4c(=O)c2c(NC(=O)CC[N+](C)CCOCC1)cccc2c(=O)c34)CCOCC5
G4L0056	C ₂₆ H ₃₄ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc2cccc3c(=O)c1c(NC(=O)CC[N+](C)(C)C)cccc1c(=O)c23
G4L0057	C ₃₀ H ₄₂ N ₄ O ₄	CC[N+](C)(CC)CCC(=O)Nc2cccc3c(=O)c1c(NC(=O)CC[N+](C)(CC)CC)cccc1c(=O)c23
G4L0058	C ₃₂ H ₄₂ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc3cccc4c(=O)c2cccc(NC(=O)CC[N+](C)CCCCC1)c2c(=O)c34)CCCCC5
G4L0059	C ₃₀ H ₃₈ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc3cccc4c(=O)c2cccc(NC(=O)CC[N+](C)CCCC1)c2c(=O)c34)CCCCC5
G4L0060	C ₃₀ H ₃₈ N ₄ O ₆	C[N+](C)(C)CCC(=O)Nc3cccc4c(=O)c2cccc(NC(=O)CC[N+](C)CCOCC1)c2c(=O)c34)CCOCC5
G4L0061	C ₂₆ H ₃₄ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc2cccc3c(=O)c1cccc(NC(=O)CC[N+](C)(C)C)c1c(=O)c23
G4L0062	C ₃₀ H ₄₂ N ₄ O ₄	CC[N+](C)(CC)CCC(=O)Nc2cccc3c(=O)c1cccc(NC(=O)CC[N+](C)(CC)CC)c1c(=O)c23
G4L0066	C ₂₆ H ₃₄ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc3ccc2c(=O)c1cc(NC(=O)CC[N+](C)(C)C)ccc1c(=O)c2c3
G4L0068	C ₃₂ H ₄₂ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc4ccc3c(=O)c2ccc(NC(=O)CC[N+](C)CCCCC1)cc2c(=O)c3c4)CCCCC5
G4L0069	C ₃₀ H ₃₈ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc4ccc3c(=O)c2ccc(NC(=O)CC[N+](C)CCCC1)cc2c(=O)c3c4)CCCCC5
G4L0070	C ₃₀ H ₃₈ N ₄ O ₆	C[N+](C)(C)CCC(=O)Nc4ccc3c(=O)c2ccc(NC(=O)CC[N+](C)CCOCC1)cc2c(=O)c3c4)CCOCC5
G4L0071	C ₂₆ H ₃₄ N ₄ O ₄	C[N+](C)(C)CCC(=O)Nc3ccc2c(=O)c1ccc(NC(=O)CC[N+](C)(C)C)cc1c(=O)c2c3
G4L0080	C ₂₃ H ₂₈ N ₄ O ₃	CN(C)CCC(=O)Nc1ccc3c(c1)c(=O)c2cc(NC(=O)CN(C)C)ccc23
G4L0081	C ₂₇ H ₃₆ N ₄ O ₃	CCN(CC)CCC(=O)Nc1ccc3c(c1)c(=O)c2cc(NC(=O)CCN(CC)CC)ccc23
G4L0083	C ₂₉ H ₃₆ N ₄ O ₃	O=C(CCN1CCCCC1)Nc2ccc5c(c2)c(=O)c4cc(NC(=O)CCN3CCCCC3)ccc45

G4L0084	C ₃₁ H ₄₀ N ₄ O ₃	C[C@H]1CCCCN1CCC(=O)Nc2ccc5c(c2)c(=O)c4cc(NC(=O)CCN3CCCC[C@H]3C)ccc45
G4L0086	C ₃₃ H ₄₄ N ₄ O ₃	CC[C@H]1CCCCN1CCC(=O)Nc2ccc5c(c2)c(=O)c4cc(NC(=O)CCN3CCCC[C@H]3CC)ccc45
G4L0087	C ₃₁ H ₄₀ N ₄ O ₅	O=C(CCN1CCCC[C@H]1CO)Nc2ccc5c(c2)c(=O)c4cc(NC(=O)CCN3CCCC[C@@H]3CO)ccc45
G4L0088	C ₃₃ H ₄₄ N ₄ O ₅	O=C(CCN1CCCC[C@H]1CCO)Nc2ccc5c(c2)c(=O)c4cc(NC(=O)CCN3CCCC[C@H]3CCO)ccc45
G4L0089	C ₂₉ H ₃₆ N ₄ O ₅	O=C(CCN1CCC(O)CC1)Nc2ccc5c(c2)c(=O)c4cc(NC(=O)CCN3CCC(O)CC3)ccc45
G4L0090	C ₃₁ H ₄₀ N ₄ O ₃	CC5CCN(CCC(=O)Nc1ccc4c(c1)c(=O)c3cc(NC(=O)CCN2CCC(C)CC2)ccc34)CC5
G4L0091	C ₃₃ H ₄₄ N ₄ O ₃	CCC5CCN(CCC(=O)Nc1ccc4c(c1)c(=O)c3cc(NC(=O)CCN2CCC(CC)CC2)ccc34)CC5
G4L0096	C ₂₅ H ₃₄ N ₄ O ₃	C[N+](C)(C)CCC(=O)Nc1ccc3c(c1)c(=O)c2cc(NC(=O)CC[N+](C)(C)C)ccc23
G4L0097	C ₂₉ H ₄₂ N ₄ O ₃	CC[N+](C)(CC)CCC(=O)Nc1ccc3c(c1)c(=O)c2cc(NC(=O)CC[N+](C)(CC)CC)ccc23
G4L0098	C ₂₉ H ₃₈ N ₄ O ₅	C[N+]5(CCC(=O)Nc1ccc4c(c1)c(=O)c3cc(NC(=O)CC[N+]2(C)CCOCC2)ccc34)CCOCC5
G4L0099	C ₃₁ H ₄₂ N ₄ O ₃	C[N+]5(CCC(=O)Nc1ccc4c(c1)c(=O)c3cc(NC(=O)CC[N+]2(C)CCCCC2)ccc34)CCCCC5
G4L0103	C ₃₃ H ₄₆ N ₄ O ₅	C[N@@+]5(CCC(=O)Nc1ccc4c(c1)c(=O)c3cc(NC(=O)CC[N@@+]2(C)CCCC[C@@H]2CO)ccc34)CC[C@@H]5CO
G4L0108	C ₂₇ H ₄₀ N ₅ O ₂	CC[NH+](CC)CCC(=O)Nc3ccc2cc1ccc(NC(=O)C[C[NH+](CC)CC)cc1[nH+]c2c3
G4L0109	C ₂₃ H ₃₂ N ₅ O ₂	C[NH+](C)CCC(=O)Nc3ccc2cc1ccc(NC(=O)CC[NH+](C)C)cc1[nH+]c2c3
G4L0110	C ₂₉ H ₄₀ N ₅ O ₂	O=C(CC[NH+]1CCCCC1)Nc5ccc4cc3ccc(NC(=O)CC[NH+]2CCCCC2)cc3[nH+]c4c5
G4L0112	C ₂₇ H ₃₆ N ₅ O ₂	O=C(CC[NH+]1CCCC1)Nc5ccc4cc3ccc(NC(=O)C[C[NH+]2CCCC2)cc3[nH+]c4c5
G4L0113	C ₃₃ H ₄₈ N ₅ O ₂	[H][N@+]5(CCC(=O)Nc4ccc3cc2ccc(NC(=O)CC[N@@+]1([H])CCCC[C@H]1CC)cc2[nH+]c3c4)CCCC[C@H]5CC
G4L0114	C ₃₁ H ₄₄ N ₅ O ₂	[H][N@+]5(CCC(=O)Nc4ccc3cc2ccc(NC(=O)CC[N@@+]1([H])CCCC[C@H]1C)cc2[nH+]c3c4)CC[C@H]5C
G4L0115	C ₃₁ H ₄₄ N ₅ O ₂	[H][N@+]5(CCC(=O)Nc4ccc3cc2ccc(NC(=O)CC[N@@+]1([H])CC[C@@H](C)CC1)cc2[nH+]c3c4)CC[C@@H](C)CC5

G4L0116	C ₃₁ H ₄₄ N ₅ O ₂	[H][N@@+]5(CCC(=O)Nc4ccc3cc2ccc(NC(=O)C C[N@@+]1([H])CCC[C@H](C)C1)cc2[nH+]c3c4) CCC[C@@H](C)C5
G4L0117	C ₃₁ H ₄₄ N ₅ O ₄	[H][N@+]5(CCC(=O)Nc4ccc3cc2ccc(NC(=O)CC[N@+]1([H])CCCC[C@@H]1CO)cc2[nH+]c3c4)C CCC[C@H]5CO
G4L0118	C ₃₃ H ₄₈ N ₅ O ₄	[H][N@@+]5(CCC(=O)Nc4ccc3cc2ccc(NC(=O)C C[N@@+]1([H])CCCC[C@@H]1CCO)cc2[nH+]c 3c4)CCCC[C@H]5CCO
G4L0119	C ₂₉ H ₄₀ N ₅ O ₄	[H][N@+]5(CCC(=O)Nc4ccc3cc2ccc(NC(=O)CC[N@@+]1([H])CC[C@@H](O)CC1)cc2[nH+]c3c4) CC[C@@H](O)CC5
G4L0120	C ₃₁ H ₄₄ N ₅ O ₂	O=C(CC[NH+]1CCCCC1)Nc5ccc4cc3ccc(NC(= O)CC[NH+]2CCCCC2)cc3[nH+]c4c5
G4L0124	C ₂₂ H ₁₇ F ₂ N ₂	CC3CC4C1CC(F)CCC1[NH+](C)C5[C@@H]2C[C@@H](F)CC[C@@H]2[N@@H+](C)[C@@H](C3)C45
G4L0133	C ₃₂ H ₄₂ N ₄ O ₆	C[N@+]5(CCC(=O)Nc3ccc(NC(=O)CC[N@@+]1 (C)CCCC[C@H]1O)c4c(=O)c2cccc2c(=O)c34)C CCC[C@H]5O
G4L0142	C ₃₄ H ₄₄ N ₄ O ₆	O=C(CCN1CCC(CCO)CC1)Nc5ccc4c(=O)c3cc(N C(=O)CCN2CCC(CCO)CC2)ccc3c(=O)c4c5
G4L0154	C ₂₃ H ₂₈ N ₄	CN(C)CCc4ccc3nc2c1cccc1n(CCN(C)C)c2cc3c4
G4L0155	C ₂₀ H ₁₇ N O ₄	CN(C)CCOc1ccc3c(c1)oc4c(=O)c2cccc2c(=O)c34
G4L0158	C ₂₈ H ₃₀ N ₄	CCCNc5ccc4nc3c(ccc2nc1ccc(CNCCC)cc1cc23) cc4c5
G4L0159	C ₂₄ H ₂₁ N ₃	CCCNc4cc2nc1cccc1cc2c5cc3cccc3nc45
G4L0160	C ₂₆ H ₂₆ N ₄	CN(C)CCCNc4cc2nc1cccc1cc2c5cc3cccc3nc4 5
G4L0161	C ₂₈ H ₃₀ N ₄	CCCNc5ccc4nc3ccc2nc1ccc(CNCCC)cc1cc2c3cc 4c5
G4L0162	C ₂₈ H ₃₀ N ₄	CCCNc5ccc4nc3c(ccc2cc1cc(CNCCC)ccc1nc23) cc4c5
G4L0163	C ₃₂ H ₄₀ N ₆	CN(C)CCCNc5ccc4nc3c(ccc2cc1cc(CNCCCN(C)C)ccc1nc23)cc4c5
G4L0164	C ₂₈ H ₂₈ N ₄	c6ccc5nc4cc(CNCCN1CCCC1)c3nc2cccc2cc3c 4cc5c6
G4L0166	C ₃₂ H ₄₀ N ₆	CN(C)CCCNc5ccc4nc3c(ccc2nc1ccc(CNCCCN(C)C)cc1cc23)cc4c5
G4L0167	C ₂₁ H ₂₀ N ₃	CC[n+]4c(c1cccc1)c2cc(N)ccc2c3ccc(N)cc34
G4L0168	C ₂₉ H ₂₇ N ₆	CC[n+]5c(c1cccc1)c3cc(N/N=N/c2ccc(C(C)=N)cc 2)ccc3c4ccc(N)cc45

G4L0169	C ₂₈ H ₂₆ N ₇	CC[n+] ₅ c(c1cccc1)c3cc(N/N=N/c2cccc(C(=N)N)c2)ccc3c4ccc(N)cc45
G4L0170	C ₂₈ H ₂₆ N ₇	CC[n+] ₅ c(c1cccc1)c3c(/N=N/c2cccc(C(=N)N)c2)c(N)ccc3c4ccc(N)cc45
G4L0171	C ₂₇ H ₂₈ N ₆	Cc5ccc(c4c1cc(N)ccc1c3ccc(Nc2cc(C)[n+](C)c(N)n2)cc3[n+] ₄ C)cc5
G4L0172	C ₂₇ H ₃₄ N ₅ O ₂	O=C(CCN1CCCC1)Nc5ccc4cc3ccc(NC(=O)CCN2CCCC2)cc3[nH+] ₄ c4c5
G4L0174	C ₃₅ H ₄₄ N ₇ O ₂	CN(C)c6ccc(Nc4c2ccc(NC(=O)CCN1CCCC1)cc2[nH+] ₅ cc(NC(=O)CCN3CCCC3)ccc45)cc6
G4L0175	C ₃₃ H ₄₀ N ₇ O ₂	Nc6ccc(Nc4c2ccc(NC(=O)CCN1CCCC1)cc2[nH+] ₅ cc(NC(=O)CCN3CCCC3)ccc45)cc6
G4L0185	C ₂₆ H ₁₈ F N ₃ O ₅	N[C@H]7CCN(c4c(F)cc5c(=O)c(C(=O)O)c[n+] ₆ c3cc2oc1cccc1c2cc3[o+] ₄ c4c56)C7
G4L0189	C ₂₁ H ₃₀ N ₈	C[N+](C)(C)c3ccc(Nc2nc(N)nc(Nc1ccc([N+](C)(C)C)cc1)n2)cc3
G4L0190	C ₁₉ H ₁₈ N ₈ S ₂	CN5CS[C@H]4C[C@@H](Nc3nc(N)nc(N[C@H]1CC[C@@H]2[C@@H](C1)SCN2C)n3)CC[C@H]45
G4L0191	C ₂₃ H ₂₂ N ₈	C[n+] ₅ cccc4cc(Nc3nc(N)nc(Nc2ccc1c(ccc[n+] ₁ C)c2)n3)ccc45
G4L0192	C ₂₈ H ₃₁ N ₉ S	CSc5nc(Nc2ccc1nc(C)cc(N(C)C)c1c2)nc(Nc4ccc3nc(C)cc(N(C)C)c3c4)n5
G4L0193	C ₂₄ H ₂₃ N ₅ O ₂	CN(C)CCCNc4nc1cccc1c5[nH]c3c2cc(N(=O)=O)ccc2ccc3c45
G4L0194	C ₂₅ H ₂₆ N ₄ O	COc5ccc4ccc2c([nH]c3c1cccc1nc(NCCCN(C)C)c23)c4c5
G4L0195	C ₂₄ H ₂₄ N ₄ O	COc5ccc4ccc2c([nH]c3c1cccc1nc(NCCN(C)C)c23)c4c5
G4L0196	C ₂₅ H ₂₆ N ₄ O	COc5ccc1c(ccc3c1[nH]c4c2cccc2nc(NCCCN(C)C)c34)c5
G4L0197	C ₃₅ H ₄₃ N ₇ O ₂	CN(C)c6ccc(Nc4c2ccc(NC(=O)CCN1CCCC1)cc2nc5cc(NC(=O)CCN3CCCC3)ccc45)cc6
G4L0213	C ₂₄ H ₂₃ N ₂	CC[N@@H+] ₃ [C@H]1CCCC[C@H]1C5C2C(CC(C)C[C@@H]23)C4CCCCC4[NH+] ₅ CC
G4L0214	C ₂₈ H ₂₆ N ₃ O ₂	C[N@@H+] ₃ [C@@H]1CCCC[C@@H]1C6C2C(CCC[C@H]23)C5CC(/C=C/C(=O)N4CCOCC4)C5[NH+] ₆ C
G4L0238	C ₁₆ H ₁₃ N ₂	C[n+] ₄ c1cccc1cc3[nH]c2cccc2c34
G4L1025	C ₂₄ H ₂₈ N ₄ O ₄	CN(C)CCC[NH+] ₁ C(=O)c2ccc3C(=O)[NH+](CCCN(C)C)C(=O)c4ccc(C1=O)c2c34
G4L1027	C ₂₄ H ₃₀ N ₄ O ₄	C[N+](C)(C)CC[NH+] ₁ C(=O)c2ccc3C(=O)[NH+](CC[N+](C)(C)C)C(=O)c4ccc(C1=O)c2c34

G4L1028	C ₂₆ H ₃₄ N ₄ O ₄	C[N+](C)(C)CCC[NH+]1C(=O)c2ccc3C(=O)[NH+](CCC[N+](C)(C)C)C(=O)c4ccc(C1=O)c2c34
G4L1029	C ₂₅ H ₃₀ N ₄ O ₄	CCCCNCC[NH+]1C(=O)c2ccc3C(=O)[NH+](CCNCCC)C(=O)c4ccc(C1=O)c2c34
G4L1032	C ₃₄ H ₃₀ N ₂ O ₄	CN(C)CCC1C(=O)c2ccc3c5ccc6C(=O)C(CCN(C)C)C(=O)c7ccc(c4ccc(C1=O)c2c34)c5c67
G4L1033	C ₄₀ H ₃₈ N ₂ O ₄	O=C2c4ccc5c7ccc8C(=O)C(CCN1CCCCC1)C(=O)c9ccc(c6ccc(C(=O)C2CCN3CCCCC3)c4c56)c7c89
G4L1034	C ₃₄ H ₃₀ N ₂ O ₆	O=C1c2ccc3c5ccc6C(=O)C(CCNCCO)C(=O)c7ccc(c4ccc(C(=O)C1CCNCCO)c2c34)c5c67
G4L1035	C ₃₆ H ₃₆ N ₂ O ₄	C[N+](C)(C)CCC1C(=O)c2ccc3c5ccc6C(=O)C(CC[N+](C)(C)C)C(=O)c7ccc(c4ccc(C1=O)c2c34)c5c67
G4L1036	C ₄₀ H ₄₆ N ₂	c3cc5c7ccc8CC(CCN1CCCCC1)Cc9ccc(c6ccc4CC(CCN2CCCCC2)Cc3c4c56)c7c89
G4L1178	C ₂₃ H ₂₆ Cl N ₃ O ₃	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCN(C)CCl
G4L1179	C ₂₅ H ₂₆ Cl N ₃ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCN6CCOCC6Cl
G4L1180	C ₂₆ H ₃₀ Cl N ₃ O ₃	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCN6CCCCC6Cl
G4L1181	C ₂₁ H ₂₁ Cl N ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCOCl
G4L1182	C ₂₇ H ₂₅ Cl N ₂ O ₃	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCc6ccccc6Cl
G4L1183	C ₂₄ H ₂₈ Cl N ₃ O ₃	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCCN(C)CCl
G4L1184	C ₂₆ H ₃₀ Cl N ₃ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCN6CCOCC6Cl
G4L1185	C ₂₇ H ₃₂ Cl N ₃ O ₃	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCN6CCCCC6Cl
G4L1186	C ₂₆ H ₃₀ Cl N ₃ O ₃	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCN6CCCC6Cl
G4L1187	C ₂₂ H ₂₄ Cl N ₃ O ₃	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5NCCNCl
G4L1311	C ₂₇ H ₃₁ Cl N ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCCN6CCCCC6Cl
G4L1312	C ₂₉ H ₃₁ Cl N ₃ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCC[n+]6ccc(N(C)C)cc6Cl
G4L1313	C ₂₇ H ₂₆ Cl N ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCC[n+]6ccccc6Cl

G4L1314	C ₂₉ H ₃₀ Cl N ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC CC[n+]6cccc6Cl
G4L1315	C ₃₀ H ₃₂ Cl N ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC CCC[n+]6cccc6Cl
G4L1316	C ₂₈ H ₂₈ Cl N ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC [n+]6cccc(C)c6Cl
G4L1317	C ₂₈ H ₂₇ Cl FN ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC [n+]6cccc(CF)c6Cl
G4L1318	C ₂₉ H ₃₁ Cl N ₃ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC [n+]6cccc(N(C)C)c6Cl
G4L1319	C ₂₈ H ₂₈ Cl N ₂ O ₅	COc6ccc[n+](CCCOc4c(OC)ccc5cc3c2cc1OCOc1c c2CC[n+]3cc45)c6Cl
G4L1320	C ₂₆ H ₂₇ Cl N ₃ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC N6CCN(C)C6Cl
G4L1321	C ₃₁ H ₂₈ Cl N ₂ O ₄	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC [n+]7ccc6cccc6c7Cl
G4L1322	C ₂₆ H ₂₅ Cl N ₄ O ₅	COc5ccc4cc3c2cc1OCOc1cc2CC[n+]3cc4c5OCCC [NH+]6/C=C\N(N)=N/C6=OCl
G4L1377	C ₂₄ H ₃₃ N ₅ O ₂	CCN(C)CC(=O)NCc2cccc3cc1cccc(CNC(=O)CN(C)C)c1nc23
G4L1378	C ₂₆ H ₃₇ N ₅ O ₂	CCN(C)CCC(=O)NCc2cccc3cc1cccc(CNC(=O)CC N(C)C)c1nc23
G4L1381	C ₂₉ H ₃₅ N ₃ O ₂	O=C(CC1CCCC1)NCc4cccc5cc3cccc(CNC(=O)C C2CCCC2)c3nc45
G4L1382	C ₃₁ H ₃₉ N ₃ O ₂	O=C(CCC1CCCC1)NCc4cccc5cc3cccc(CNC(=O) CCC2CCCC2)c3nc45
G4L1383	C ₃₃ H ₄₃ N ₃ O ₂	O=C(CCCC1CCCC1)NCc4cccc5cc3cccc(CNC(=O))CCCC2CCCC2)c3nc45
G4L1385	C ₃₃ H ₄₃ N ₃ O ₂	O=C(CCC1CCCCC1)NCc4cccc5cc3cccc(CNC(=O))CCC2CCCCC2)c3nc45
G4L1386	C ₃₁ H ₃₉ N ₃ O ₂	O=C(CC1CCCCC1)NCc4cccc5cc3cccc(CNC(=O) CC2CCCCC2)c3nc45
G4L1387	C ₃₃ H ₄₅ N ₅ O ₂	CN5CCC(CCC(=O)NCc3cccc4cc2cccc(CNC(=O) CCC1CCN(C)CC1)c2nc34)CC5
G4L1388	C ₂₃ H ₂₉ I ₂ N ₅ O ₂	C[N+](C)(I)CC(=O)NCc2cccc3cc1cccc(CNC(=O) C[N+](C)(C)I)c1nc23
G4L1389	C ₂₅ H ₃₃ I ₂ N ₅ O ₂	C[N+](C)(I)CCC(=O)NCc2cccc3cc1cccc(CNC(=O))CC[N+](C)(C)I)c1nc23

Appendix G

Complete results of CORAL QSAR model

	Correlation	RMSE
1	0.7778	0.3480
2	0.7973	0.4360
3	0.7510	0.5110
4	0.8012	0.3740
5	0.3224	0.4030
6	0.8110	0.3070
7	0.7486	0.4030
8	0.6610	0.5100
9	0.8637	0.4150
10	0.8782	0.2880
MEAN	0.7412	0.3995
SD	0.1594	0.0749

	Correlation	RMSE
1	0.7980	0.3070
2	0.7574	0.4880
3	0.8863	0.3430
4	0.6631	0.4680
5	0.7833	0.3770
6	0.7950	0.6060
7	0.8233	0.4020
8	0.8506	0.4030
9	0.7020	0.1800
10	0.5756	0.4770
MEAN	0.7635	0.4051
SD	0.0931	0.1161

	Correlation	RMSE
1	0.8870	0.3840
2	0.6301	0.3320
3	0.4566	0.3480
4	0.8158	0.3880
5	0.9039	0.4010
6	0.7332	0.4050
7	0.8308	0.4330
8	0.7124	0.3850
9	0.8414	0.3210
10	0.8067	0.4530
MEAN	0.7618	0.3850
SD	0.1357	0.0420

	Correlation	RMSE
1	0.5844	0.5020
2	0.6720	0.5160
3	0.7086	0.4320
4	0.8266	0.1960
5	0.7929	0.3420
6	0.8739	0.3680
7	0.6818	0.3730
8	0.8915	0.3700
9	0.6866	0.6070
10	0.9280	0.2840
MEAN	0.7646	0.3990
SD	0.1138	0.1196

	Correlation	RMSE
1	0.8689	0.3530
2	0.6694	0.4700
3	0.7237	0.4060
4	0.8431	0.4160
5	0.8261	0.2860
6	0.8246	0.3960
7	0.8740	0.2120
8	0.7278	0.4860
9	0.7926	0.4540
10	0.5598	0.4280
MEAN	0.7710	0.3907

	Correlation	RMSE
1	0.7903	0.3970
2	0.9287	0.3270
3	0.6827	0.4460
4	0.7795	0.4490
5	0.7221	0.4190
6	0.6317	0.4500
7	0.6281	0.5170
8	0.8476	0.3390
9	0.8647	0.4100
10	0.6619	0.3480
MEAN	0.7537	0.4102

SD	0.1004	0.0856
-----------	---------------	---------------

SD	0.1048	0.0595
-----------	---------------	---------------

	Correlation	RMSE
1	0.7547	0.6050
2	0.7996	0.3270
3	0.6050	0.3130
4	0.7312	0.3560
5	0.8546	0.3370
6	0.8323	0.2890
7	0.7055	0.4990
8	0.7522	0.2750
9	0.6771	0.3600
10	0.8534	0.5490
MEAN	0.7566	0.3910
SD	0.0811	0.1162

	Correlation	RMSE
1	0.5741	0.5950
2	0.2827	0.4360
3	0.6878	0.3450
4	0.7406	0.4080
5	0.6419	0.3740
6	0.8828	0.3530
7	0.6882	0.3810
8	0.8189	0.4510
9	0.8688	0.3410
10	0.8789	0.4780
MEAN	0.7065	0.4162
SD	0.1835	0.0783

	Correlation	RMSE
1	0.6266	0.4610
2	0.6294	0.4750
3	0.7869	0.3560
4	0.7797	0.3510
5	0.7421	0.4810
6	0.7936	0.4040
7	0.8053	0.2910
8	0.7862	0.5700
9	0.8570	0.3360
10	0.8072	0.3990
MEAN	0.7614	0.4124
SD	0.0758	0.0841

	Correlation	RMSE
1	0.7909	0.4230
2	0.8516	0.3990
3	0.7801	0.3760
4	0.7701	0.2240
5	0.7860	0.4980
6	0.8123	0.3260
7	0.8368	0.3220
8	0.8062	0.4030
9	0.6213	0.2570
10	0.7077	0.4110
MEAN	0.7763	0.3639
SD	0.0672	0.0822

Gradient Boosting Tree		
Average Correlation	Average RMSE	SD
0.7557	0.3557	0.0197

Appendix H

Complete list of additional ligands

<u>Index</u>	<u>Formula</u>	<u>Structure</u>
G4L0030 ³⁴	C ₂₈ H ₃₂ N ₄ O ₆
G4L0035 ³⁴	C ₂₈ H ₃₂ N ₄ O ₆
G4L0040 ³⁴	C ₂₈ H ₃₂ N ₄ O ₆
G4L0045 ³⁴	C ₂₈ H ₃₂ N ₄ O ₆
G4L0082 ³⁵	C ₂₇ H ₃₂ N ₄ O ₅

G4L0092 ³⁵	C ₄₁ H ₄₄ N ₄ O ₃
G4L0093 ³⁵	C ₄₃ H ₄₈ N ₄ O ₃
G4L0094 ³⁵	C ₃₉ H ₄₂ N ₆ O ₃
G4L0095 ³⁵	C ₃₇ H ₄₀ N ₈ O ₃
G4L0111 ³⁶	C ₂₇ H ₃₆ N ₅ O ₄
G4L0121 ³⁶	C ₃₃ H ₄₈ N ₅ O ₂

G4L0122 ³⁶	C ₃₉ H ₅₆ N ₅ O ₂
G4L0123 ³⁶	C ₄₃ H ₅₂ N ₅ O ₂
G4L0136 ³⁸	C ₃₂ H ₄₄ N ₄ O ₄
G4L0137 ³⁸	C ₂₈ H ₃₂ N ₄ O ₄
G4L0140 ³⁸	C ₂₆ H ₂₈ N ₄ O ₆
G4L0173 ⁴²	C ₃₃ H ₄₆ N ₅ O ₂

Appendix I

Complete results of ligand-based ESPH model using Test set 1

Ligand-based 4-subgroups ESPH model

Gradient Boosting Tree		
	Correlation	RMSE
1	0.5580	1.0656
2	0.8535	0.5425
3	0.7384	0.8010
4	0.3794	0.9349
5	0.8760	0.5029
6	0.8012	0.7692
7	0.8951	0.5229
8	0.0956	0.9927
9	0.6996	0.3687
10	0.8814	0.3283
MEAN	0.6778	0.6829
SD	0.2626	

Random Forest		
	Correlation	RMSE
1	0.7788	0.5756
2	0.8638	0.5679
3	0.6638	0.8441
4	0.5161	1.1164
5	0.8405	0.5922
6	0.4562	0.8791
7	0.6894	0.4448
8	0.8277	0.6636
9	0.7257	0.7868
10	0.7057	0.9778
MEAN	0.7068	0.7448
SD	0.1350	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.9162	0.6329
2	0.8082	0.7951
3	0.5017	1.0637
4	0.9289	0.4531
5	0.7355	0.5684
6	0.7922	0.6323
7	0.2675	1.0255
8	0.6947	0.5236
9	0.6581	0.5069
10	0.6806	0.8430
MEAN	0.6984	0.7045
SD	0.1970	

Random Forest		
	Correlation	RMSE
1	0.9401	0.6635
2	0.6325	0.5346
3	0.7756	0.5721
4	0.9189	0.6361
5	0.9445	0.5849
6	0.5552	1.0591
7	0.6436	0.7332
8	0.4422	1.0543
9	0.7642	0.5300
10	0.6828	0.8559
MEAN	0.7299	0.7224
SD	0.1705	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8199	0.7328
2	0.6187	0.8274
3	0.6696	0.7500
4	0.2567	0.8172
5	0.8598	0.4183
6	0.6239	0.5207
7	0.9046	0.6238

Random Forest		
	Correlation	RMSE
1	0.8066	0.7239
2	0.5333	0.8223
3	0.8719	0.6251
4	0.7419	0.8903
5	0.2761	0.8519
6	0.3092	1.0748
7	0.8876	0.6337

8	0.4477	1.2528
9	0.8893	0.4898
10	0.7419	0.7761
MEAN	0.6832	0.7209
SD	0.2082	

8	0.8136	0.6436
9	0.8793	0.4297
10	0.5577	0.7760
MEAN	0.6677	0.7471
SD	0.2340	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.5706	0.8581
2	0.7465	0.6933
3	0.7618	0.5567
4	0.8537	0.4672
5	0.9132	0.5587
6	0.1849	1.0045
7	0.7858	0.8629
8	0.7086	0.7018
9	0.8455	0.5385
10	0.7424	0.7836
MEAN	0.7113	0.7025
SD	0.2072	

Random Forest		
	Correlation	RMSE
1	0.4279	0.7376
2	0.8580	0.4929
3	0.9390	0.6230
4	0.8692	0.5213
5	0.6109	1.0706
6	0.7298	0.7321
7	0.3900	1.0017
8	0.9256	0.3743
9	0.7777	0.7031
10	0.6937	0.8984
MEAN	0.7222	0.7155
SD	0.1949	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7031	0.8237
2	0.4957	0.7795
3	0.7687	0.5925
4	0.7980	0.7697
5	0.8568	0.7370
6	0.7732	0.4739
7	0.7006	0.7681
8	0.5470	0.8469
9	0.8419	0.5621
10	0.7587	0.8315
MEAN	0.7244	0.7185
SD	0.1188	

Random Forest		
	Correlation	RMSE
1	0.8233	0.8176
2	0.5319	0.9016
3	0.8758	0.5812
4	0.6754	0.7238
5	0.7490	0.8273
6	0.7081	0.7643
7	0.5251	0.8468
8	0.6408	0.4529
9	0.6270	0.8106
10	0.7747	0.5811
MEAN	0.6931	0.7307
SD	0.1164	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7739	0.7441
2	0.2876	1.2768
3	0.8297	0.4043
4	0.3979	0.7435
5	0.8731	0.5483

Random Forest		
	Correlation	RMSE
1	0.9340	0.4537
2	0.8069	0.7257
3	0.1489	1.1521
4	0.7958	0.7649
5	0.3419	0.9201

6	0.5846	0.8718
7	0.8294	0.5636
8	0.8130	0.5240
9	0.7654	0.6614
10	0.9165	0.6073
MEAN	0.7071	0.6945
SD	0.2127	

6	0.6083	0.5528
7	0.7709	0.7302
8	0.9283	0.5704
9	0.8530	0.6351
10	0.7865	0.5519
MEAN	0.6974	0.7057
SD	0.2590	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7556	0.4320
2	0.7373	0.6790
3	0.8628	0.3956
4	0.8421	0.4994
5	0.4027	1.2018
6	0.7138	0.6408
7	0.7171	0.7741
8	0.6856	0.8427
9	0.7907	0.7422
10	0.8563	0.5941
MEAN	0.7364	0.6802
SD	0.1332	

Random Forest		
	Correlation	RMSE
1	0.7776	0.5325
2	0.7045	0.7877
3	0.5136	1.1186
4	0.6517	0.7892
5	0.6101	0.6092
6	0.0203	0.8811
7	0.5182	0.9589
8	0.9373	0.4890
9	0.8082	0.5215
10	0.5959	0.8864
MEAN	0.6137	0.7574
SD	0.2474	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6356	0.9166
2	0.2660	0.9355
3	0.8631	0.5916
4	0.6917	0.4780
5	0.6481	0.9283
6	0.7566	0.7466
7	0.8459	0.5416
8	0.8119	0.7074
9	0.5465	0.8407
10	0.8932	0.4660
MEAN	0.6958	0.7152
SD	0.1885	

Random Forest		
	Correlation	RMSE
1	0.7117	0.6910
2	0.8974	0.5207
3	0.9245	0.5367
4	0.8437	0.7146
5	0.7742	0.6878
6	0.3045	1.0768
7	0.3146	0.6421
8	0.6636	0.5894
9	0.3749	1.1594
10	0.7129	0.5694
MEAN	0.6522	0.7188
SD	0.2369	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.9015	0.4931
2	0.9238	0.5513

Random Forest		
	Correlation	RMSE
1	0.7218	0.7486
2	0.9238	0.5305

3	0.4082	0.8660
4	0.5158	1.0123
5	0.8574	0.5333
6	0.7249	0.7605
7	0.6034	0.8084
8	0.3233	1.1085
9	0.6315	0.4978
10	0.8921	0.4883
MEAN	0.6782	0.7120
SD	0.2169	

3	0.2817	0.8580
4	0.6825	0.8028
5	0.3550	0.9929
6	0.9516	0.3795
7	0.8319	0.5953
8	0.9236	0.4164
9	0.7487	0.9497
10	0.5103	0.9998
MEAN	0.6931	0.7274
SD	0.2388	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7939	0.5904
2	0.6941	0.6766
3	0.6065	0.9071
4	0.3008	1.0055
5	0.7894	0.6440
6	0.5899	0.7474
7	0.9492	0.4219
8	0.8533	0.5065
9	0.8362	0.6430
10	0.6738	0.8748
MEAN	0.7087	0.7017
SD	0.1829	

Random Forest		
	Correlation	RMSE
1	0.6791	0.4656
2	0.3978	1.0254
3	0.8272	0.7180
4	0.5399	0.7560
5	0.8693	0.5348
6	0.4593	0.9735
7	0.8489	0.7289
8	0.9271	0.6900
9	0.8494	0.7328
10	0.7353	0.6430
MEAN	0.7133	0.7268
SD	0.1873	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7021	0.7033	0.0194	0.6889	0.7297	0.0354

Ligand-based 7-subgroups ESPH model

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6171	0.8295
2	0.8176	0.8090
3	0.5805	0.9424
4	0.7500	0.6504
5	0.8176	0.6023
6	0.9325	0.4463
7	0.7171	0.7775
8	0.5480	0.6815
9	0.4275	0.8494
10	0.8452	0.5876
MEAN	0.7053	0.7176
SD	0.1577	

Random Forest		
	Correlation	RMSE
1	0.8188	0.7962
2	0.3618	1.0722
3	0.8362	0.6600
4	0.4489	0.8066
5	0.9133	0.4150
6	0.6721	0.7865
7	0.2920	0.9433
8	0.6892	0.7561
9	0.6835	0.8321
10	0.8589	0.6356
MEAN	0.6575	0.7704
SD	0.2187	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7694	0.5582
2	0.8147	0.6695
3	0.7179	0.9180
4	0.6788	0.6744
5	0.7523	0.7495
6	0.8759	0.4875
7	0.7540	0.7459
8	0.5812	0.8068
9	0.7670	0.8080
10	0.1153	0.9291
MEAN	0.6826	0.7347
SD	0.2142	

Random Forest		
	Correlation	RMSE
1	0.4782	0.6447
2	0.4325	0.8952
3	0.4211	0.8181
4	0.6932	0.7252
5	0.8054	0.6655
6	0.4980	0.7340
7	0.5893	0.9764
8	0.8502	0.7052
9	0.6012	1.0157
10	0.6467	0.7319
MEAN	0.6016	0.7912
SD	0.1495	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.2571	1.2089
2	0.8880	0.5610
3	0.7069	0.8118
4	0.7877	0.6692
5	0.8774	0.5422
6	0.8275	0.5001
7	0.3787	0.8146
8	0.9474	0.2687
9	0.4775	0.8935

Random Forest		
	Correlation	RMSE
1	0.7608	0.7175
2	0.3638	1.0482
3	0.8061	0.5634
4	0.6525	0.6633
5	0.3891	0.8505
6	0.8360	0.7320
7	0.8177	0.8188
8	0.8610	0.6320
9	0.6673	0.8341

10	0.8609	0.6194
MEAN	0.7009	0.6889
SD	0.2419	

10	0.4779	0.8452
MEAN	0.6632	0.7705
SD	0.1892	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8273	0.6855
2	0.9031	0.4220
3	0.2869	1.1403
4	0.7004	0.6441
5	0.8314	0.5542
6	0.6179	0.8244
7	0.6041	0.9231
8	0.9611	0.3232
9	0.7220	0.7920
10	0.8299	0.4052
MEAN	0.7284	0.6714
SD	0.1940	

Random Forest		
	Correlation	RMSE
1	0.5982	0.6724
2	0.5585	1.0704
3	0.8363	0.5057
4	0.7444	0.7347
5	0.4889	0.9861
6	0.7168	0.6788
7	0.7296	0.8440
8	0.8013	0.8682
9	0.5161	0.8285
10	0.7824	0.5659
MEAN	0.6772	0.7755
SD	0.1258	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7421	0.6960
2	0.8285	0.6972
3	0.2697	0.8960
4	0.7447	0.6914
5	0.5822	0.8283
6	0.7799	0.7500
7	0.9072	0.4832
8	0.8266	0.5861
9	0.8043	0.6465
10	0.7652	0.8150
MEAN	0.7250	0.7090
SD	0.1806	

Random Forest		
	Correlation	RMSE
1	0.7995	0.7530
2	0.2644	1.1872
3	0.8514	0.5184
4	0.7533	0.7091
5	0.6553	0.7421
6	0.7951	0.8186
7	0.6473	0.6869
8	0.8123	0.6393
9	0.9200	0.5462
10	0.8593	0.4891
MEAN	0.7358	0.7090
SD	0.1864	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8300	0.5609
2	0.6797	0.7345
3	0.6355	0.6210
4	0.6473	1.0681
5	0.7461	0.5954
6	0.9310	0.3817

Random Forest		
	Correlation	RMSE
1	0.1890	1.0808
2	0.7033	0.9494
3	0.6328	0.6128
4	0.4114	0.6795
5	0.8291	0.5027
6	0.8893	0.6347

7	0.2527	1.1350
8	0.8478	0.6735
9	0.7828	0.6237
10	0.8357	0.5069
MEAN	0.7189	0.6901
SD	0.1895	

7	0.4831	0.9836
8	0.7033	0.8179
9	0.8801	0.6512
10	0.8265	0.7133
MEAN	0.6548	0.7626
SD	0.2301	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6241	0.7938
2	0.5761	1.0182
3	0.9275	0.4918
4	0.4539	0.8800
5	0.7276	0.7953
6	0.6637	0.7268
7	0.8696	0.4554
8	0.8073	0.5433
9	0.8644	0.6001
10	0.6101	0.8423
MEAN	0.7124	0.7147
SD	0.1525	

Random Forest		
	Correlation	RMSE
1	0.8109	0.4686
2	0.9592	0.3719
3	0.7842	0.9242
4	0.4213	0.9129
5	0.6748	0.8697
6	0.3306	0.8936
7	0.7659	0.6864
8	0.8839	0.6598
9	0.5478	0.6664
10	0.6689	0.9030
MEAN	0.6847	0.7357
SD	0.2004	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8329	0.4777
2	0.8225	0.6258
3	0.5509	0.7414
4	0.8348	0.6904
5	0.7845	0.6694
6	0.7828	0.6768
7	0.4800	0.9221
8	0.8699	0.5383
9	0.7985	0.7874
10	0.3856	0.9540
MEAN	0.7142	0.7083
SD	0.1735	

Random Forest		
	Correlation	RMSE
1	0.7572	0.7553
2	0.4256	0.8553
3	0.8038	0.6834
4	0.8372	0.5907
5	0.6186	0.9754
6	0.8951	0.7351
7	0.8594	0.6485
8	0.4493	0.8316
9	0.8915	0.5749
10	0.6768	0.6586
MEAN	0.7214	0.7309
SD	0.1744	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6652	0.7700
2	0.6037	0.8724
3	0.9360	0.4364
4	0.8992	0.5837

Random Forest		
	Correlation	RMSE
1	0.7979	0.6455
2	0.7857	0.8635
3	0.7138	0.9139
4	0.4122	0.9699

5	0.6168	0.9262
6	0.9408	0.3504
7	0.4805	0.9360
8	0.8116	0.4738
9	0.9031	0.5599
10	0.4038	0.9577
MEAN	0.7261	0.6867
SD	0.1982	

5	0.5291	0.8062
6	0.6007	0.7759
7	0.4415	0.6002
8	0.7415	0.6767
9	0.6841	0.9090
10	0.8207	0.6824
MEAN	0.6527	0.7843
SD	0.1491	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8089	0.4418
2	0.5107	1.0055
3	0.8896	0.5403
4	0.8612	0.5131
5	0.6525	0.4938
6	0.6079	0.7350
7	0.5426	1.0979
8	0.7766	0.5387
9	0.7378	0.8193
10	0.7846	0.9335
MEAN	0.7172	0.7119
SD	0.1319	

Random Forest		
	Correlation	RMSE
1	0.8280	0.7823
2	0.8619	0.5768
3	0.7935	0.5258
4	0.6132	0.7882
5	0.5806	0.9714
6	0.9362	0.6625
7	0.6648	0.7208
8	0.6380	0.6049
9	0.7108	0.5900
10	0.2783	1.1263
MEAN	0.6905	0.7349
SD	0.1859	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7131	0.7033	0.0139	0.6739	0.7565	0.0379

Appendix J

Complete results of ligand-based ESPH model using Test set 2

Ligand-based 4-subgroups ESPH model

Gradient Boosting Tree			Random Forest		
	Correlation	RMSE		Correlation	RMSE
1	0.8138	0.4139	1	0.2827	0.8234
2	0.7531	0.5734	2	0.8255	0.5039
3	0.8520	0.2884	3	0.7247	0.5556
4	0.4910	0.5564	4	0.9276	0.5837
5	0.6903	0.6018	5	0.6712	0.5239
6	0.8717	0.5398	6	0.8772	0.6021
7	0.9028	0.4759	7	0.8564	0.5044
8	0.9001	0.4893	8	0.7251	0.3815
9	0.8234	0.4835	9	0.8363	0.3767
10	0.3973	0.3053	10	0.8165	0.4306
MEAN	0.7495	0.4728	MEAN	0.7543	0.5286
SD	0.1750		SD	0.1832	

Gradient Boosting Tree			Random Forest		
	Correlation	RMSE		Correlation	RMSE
1	0.8282	0.5572	1	0.7115	0.3803
2	0.8613	0.3660	2	0.9258	0.4797
3	0.7038	0.4745	3	0.8589	0.4699
4	0.8650	0.6187	4	0.9016	0.5311
5	0.6797	0.4661	5	0.6476	0.5342
6	0.8498	0.5727	6	0.2740	0.7238
7	0.6772	0.5876	7	0.8439	0.5072
8	0.6740	0.2971	8	0.8813	0.5187
9	0.4641	0.5589	9	0.8879	0.5881
10	0.8786	0.4038	10	0.7885	0.5996
MEAN	0.7482	0.4903	MEAN	0.7721	0.5333
SD	0.1325		SD	0.1961	

Gradient Boosting Tree			Random Forest		
	Correlation	RMSE		Correlation	RMSE
1	0.8923	0.3572	1	0.8543	0.5856
2	0.8569	0.4379	2	0.8336	0.4051
3	0.6263	0.4568	3	0.6537	0.6818
4	0.8762	0.5665	4	0.7100	0.5695
5	0.9059	0.5275	5	0.7255	0.5045
6	0.7470	0.5098	6	0.8090	0.4487
7	0.7261	0.4603	7	0.5939	0.4989

8	0.4770	0.4765
9	0.6710	0.6043
10	0.8784	0.5673
MEAN	0.7657	0.4964
SD	0.1424	

8	0.7758	0.9294
9	0.8638	0.4283
10	0.3822	0.5687
MEAN	0.7202	0.5621
SD	0.1480	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8171	0.6472
2	0.9480	0.3541
3	0.3700	0.7381
4	0.8200	0.3996
5	0.7983	0.3772
6	0.6491	0.4690
7	0.9295	0.4584
8	0.6765	0.5337
9	0.6154	0.6214
10	0.7800	0.6019
MEAN	0.7404	0.5201
SD	0.1704	

Random Forest		
	Correlation	RMSE
1	0.8473	0.4646
2	0.7525	0.7279
3	0.8682	0.5513
4	0.7468	0.4169
5	0.6033	0.7765
6	0.4184	0.4891
7	0.8965	0.3753
8	0.8447	0.3191
9	0.7737	0.6982
10	0.2788	0.6294
MEAN	0.7030	0.5448
SD	0.2071	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6812	0.6037
2	0.8172	0.4630
3	0.8606	0.4385
4	0.7505	0.6402
5	0.6925	0.5430
6	0.7320	0.5276
7	0.6433	0.4807
8	0.9170	0.4092
9	0.8581	0.3802
10	0.9243	0.4225
MEAN	0.7877	0.4909
SD	0.1013	

Random Forest		
	Correlation	RMSE
1	0.5114	0.5229
2	0.7764	0.7499
3	0.5882	0.6704
4	0.8805	0.4720
5	0.8868	0.2406
6	0.4241	0.4168
7	0.7601	0.6864
8	0.6026	0.6271
9	0.6719	0.8035
10	0.9774	0.3763
MEAN	0.7079	0.5566
SD	0.1788	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6220	0.4538
2	0.4063	0.6496
3	0.8573	0.4593
4	0.9226	0.3785
5	0.7352	0.5945

Random Forest		
	Correlation	RMSE
1	0.9595	0.5718
2	0.2538	0.6693
3	0.8437	0.3454
4	0.8601	0.5470
5	0.9165	0.3590

6	0.8566	0.4807
7	0.8898	0.6637
8	0.8249	0.3745
9	0.7655	0.4161
10	0.8193	0.5782
MEAN	0.7699	0.5049
SD	0.1541	

6	0.4799	0.7809
7	0.7264	0.5548
8	0.8857	0.6996
9	0.7157	0.5242
10	0.6420	0.4345
MEAN	0.7283	0.5487
SD	0.2208	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.5869	0.4524
2	0.7980	0.5970
3	0.7303	0.5417
4	0.7385	0.6497
5	0.8340	0.3640
6	0.8195	0.7683
7	0.7756	0.3803
8	0.8644	0.4527
9	0.8341	0.4607
10	0.9147	0.2453
MEAN	0.7896	0.4912
SD	0.0904	

Random Forest		
	Correlation	RMSE
1	0.8677	0.3403
2	0.6167	0.6578
3	0.8092	0.4151
4	0.8633	0.5186
5	0.6580	0.6712
6	0.7970	0.4957
7	0.7025	0.5955
8	0.7230	0.4508
9	0.8153	0.5665
10	0.7382	0.6276
MEAN	0.7591	0.5339
SD	0.0851	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6520	0.5121
2	0.7439	0.5371
3	0.8377	0.4042
4	0.8851	0.4187
5	0.1638	0.6100
6	0.8566	0.3448
7	0.8714	0.5444
8	0.8260	0.6076
9	0.8260	0.6076
10	0.8529	0.4555
MEAN	0.7515	0.5042
SD	0.2179	

Random Forest		
	Correlation	RMSE
1	0.7627	0.3240
2	0.8669	0.5691
3	0.8885	0.4056
4	0.7740	0.6221
5	0.8375	0.5453
6	0.7087	0.5804
7	0.8380	0.6002
8	0.9174	0.4648
9	0.4065	0.5230
10	0.6586	0.6093
MEAN	0.7659	0.5244
SD	0.1500	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7203	0.5604
2	0.8087	0.4984

Random Forest		
	Correlation	RMSE
1	0.8551	0.3283
2	0.8445	0.5222

3	0.8662	0.5324
4	0.7772	0.4387
5	0.7604	0.4906
6	0.6749	0.4717
7	0.8138	0.5032
8	0.8624	0.5167
9	0.8637	0.3104
10	0.8577	0.6020
MEAN	0.8005	0.4925
SD	0.0667	

3	0.3204	0.6826
4	0.9493	0.6066
5	0.4246	0.5749
6	0.8799	0.4646
7	0.7080	0.7549
8	0.8693	0.4926
9	0.8056	0.5628
10	0.4040	0.4976
MEAN	0.7060	0.5487
SD	0.2325	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.2933	0.5758
2	0.7369	0.4495
3	0.8760	0.3517
4	0.8591	0.4485
5	0.7965	0.5626
6	0.7137	0.4813
7	0.7570	0.5105
8	0.8047	0.5314
9	0.8167	0.6341
10	0.9030	0.6987
MEAN	0.7557	0.5244
SD	0.1734	

Random Forest		
	Correlation	RMSE
1	0.5837	0.5091
2	0.9284	0.5108
3	0.6577	0.6547
4	0.8915	0.5967
5	0.9233	0.3497
6	0.5162	0.6257
7	0.4050	0.5706
8	0.8471	0.3980
9	0.7249	0.6559
10	0.8573	0.4579
MEAN	0.7335	0.5329
SD	0.1858	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7659	0.4988	0.0205	0.7350	0.5414	0.0261

Ligand-based 7-subgroups ESPH model

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6568	0.5911
2	0.8751	0.2927
3	0.6990	0.5980
4	0.6754	0.4563
5	0.7122	0.4446
6	0.9246	0.4048
7	0.8027	0.4559
8	0.8927	0.6532
9	0.8327	0.6012
10	0.6436	0.4924
MEAN	0.7715	0.4990
SD	0.1060	

Random Forest		
	Correlation	RMSE
1	0.7006	0.5373
2	0.5880	0.6395
3	0.8907	0.4911
4	0.6685	0.6110
5	0.6059	0.6862
6	0.9455	0.5393
7	0.8524	0.5130
8	0.7367	0.5908
9	0.8364	0.3855
10	0.4968	0.5131
MEAN	0.7321	0.5507
SD	0.1465	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.9810	0.1492
2	0.7602	0.4387
3	0.8945	0.3428
4	0.7693	0.4645
5	0.8640	0.5644
6	0.8779	0.4874
7	0.5179	0.5475
8	0.8494	0.7178
9	0.8577	0.3346
10	0.8252	0.6906
MEAN	0.8197	0.4738
SD	0.1231	

Random Forest		
	Correlation	RMSE
1	0.5796	0.5080
2	0.8678	0.6029
3	0.6908	0.5556
4	0.8570	0.3758
5	0.9191	0.5781
6	0.8431	0.6290
7	0.7084	0.5499
8	0.6804	0.5774
9	0.7625	0.6688
10	0.6925	0.4502
MEAN	0.7601	0.5496
SD	0.1076	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8726	0.5248
2	0.8828	0.4975
3	0.7792	0.3029
4	0.5130	0.4198
5	0.8292	0.5341
6	0.9202	0.3559
7	0.6907	0.4181
8	0.6408	0.5240
9	0.7043	0.7376

Random Forest		
	Correlation	RMSE
1	0.7631	0.7768
2	0.7804	0.5635
3	0.5786	0.8219
4	0.7668	0.4104
5	0.8811	0.6025
6	0.8348	0.3596
7	0.6456	0.5405
8	0.8111	0.4936
9	0.9290	0.2748

10	0.7569	0.6179
MEAN	0.7590	0.4933
SD	0.1255	

10	0.5428	0.5330
MEAN	0.7533	0.5377
SD	0.1268	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.9257	0.3153
2	0.6838	0.5168
3	0.8814	0.4098
4	0.7466	0.6180
5	0.7939	0.3747
6	0.8569	0.5592
7	0.6805	0.5674
8	0.7006	0.3401
9	0.6494	0.6789
10	0.8363	0.4728
MEAN	0.7755	0.4853
SD	0.0967	

Random Forest		
	Correlation	RMSE
1	0.3643	0.5220
2	0.6759	0.5594
3	0.8297	0.4651
4	0.3496	0.5628
5	0.8939	0.5847
6	0.9205	0.5466
7	0.7506	0.5535
8	0.9145	0.6245
9	0.6532	0.4943
10	0.9211	0.4783
MEAN	0.7273	0.5391
SD	0.2189	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7787	0.3702
2	0.7607	0.5963
3	0.7027	0.3886
4	0.6798	0.4646
5	0.8814	0.5690
6	0.8623	0.5675
7	0.9149	0.3486
8	0.8235	0.4641
9	0.6378	0.7322
10	0.6341	0.4812
MEAN	0.7676	0.4982
SD	0.1019	

Random Forest		
	Correlation	RMSE
1	0.6322	0.6124
2	0.8989	0.6532
3	0.7031	0.5865
4	0.8792	0.7118
5	0.1908	0.4874
6	0.8671	0.3704
7	0.8530	0.5525
8	0.7728	0.5585
9	0.4911	0.5537
10	0.6264	0.4197
MEAN	0.6914	0.5506
SD	0.2212	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.6025	0.5881
2	0.8523	0.5536
3	0.7742	0.5765
4	0.7939	0.5610
5	0.8655	0.4846
6	0.8297	0.4391

Random Forest		
	Correlation	RMSE
1	0.7960	0.3909
2	0.8560	0.3554
3	0.5849	0.6994
4	0.8967	0.6243
5	0.9054	0.4432
6	0.8939	0.4623

7	0.3484	0.3922
8	0.6288	0.5773
9	0.9015	0.4568
10	0.7860	0.3971
MEAN	0.7383	0.5026
SD	0.1677	

7	0.7408	0.6225
8	0.7026	0.6381
9	0.2846	0.5594
10	0.7890	0.5639
MEAN	0.7450	0.5359
SD	0.1907	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.8579	0.6693
2	0.5137	0.6598
3	0.5970	0.4305
4	0.7720	0.5854
5	0.8758	0.3563
6	0.6691	0.5270
7	0.8516	0.5386
8	0.8847	0.2733
9	0.8810	0.5415
10	0.9086	0.3848
MEAN	0.7811	0.4967
SD	0.1394	

Random Forest		
	Correlation	RMSE
1	0.7363	0.4418
2	0.9057	0.4610
3	0.7222	0.8096
4	0.7983	0.4533
5	0.5662	0.5053
6	0.7936	0.5665
7	0.5916	0.4006
8	0.8221	0.6699
9	0.7790	0.4143
10	0.8618	0.5598
MEAN	0.7577	0.5282
SD	0.1087	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7367	0.5063
2	0.7261	0.5431
3	0.7057	0.4493
4	0.8994	0.3322
5	0.7495	0.5251
6	0.9166	0.4019
7	0.8751	0.5361
8	0.8112	0.7529
9	0.7349	0.5344
10	0.6235	0.5389
MEAN	0.7779	0.5120
SD	0.0947	

Random Forest		
	Correlation	RMSE
1	0.7978	0.3871
2	0.8062	0.6257
3	0.8590	0.7352
4	0.8175	0.5562
5	0.8641	0.4361
6	0.8800	0.3398
7	0.7058	0.6305
8	0.6694	0.5943
9	0.5578	0.6078
10	0.8569	0.2690
MEAN	0.7814	0.5182
SD	0.1047	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.7457	0.3005
2	0.5941	0.6511
3	0.8733	0.5083
4	0.9242	0.4443

Random Forest		
	Correlation	RMSE
1	0.7125	0.5785
2	0.9451	0.2963
3	0.6270	0.5150
4	0.8935	0.4600

5	0.8222	0.5045
6	0.7274	0.5328
7	0.7574	0.5793
8	0.7634	0.6439
9	0.9103	0.3067
10	0.7641	0.4277
MEAN	0.7882	0.4899
SD	0.0984	

5	0.5136	0.6143
6	0.8251	0.6079
7	0.8152	0.5431
8	0.7330	0.6147
9	0.8336	0.6451
10	0.8159	0.5051
MEAN	0.7715	0.5380
SD	0.1282	

Gradient Boosting Tree		
	Correlation	RMSE
1	0.4604	0.3653
2	0.2425	0.6605
3	0.8844	0.3663
4	0.8790	0.4236
5	0.8280	0.5349
6	0.7938	0.4855
7	0.8055	0.7166
8	0.9193	0.3946
9	0.6490	0.4744
10	0.7578	0.5154
MEAN	0.7220	0.4937
SD	0.2156	

Random Forest		
	Correlation	RMSE
1	0.8952	0.5666
2	0.6812	0.6048
3	0.7444	0.5490
4	0.7915	0.5876
5	0.6580	0.3656
6	0.8789	0.4646
7	0.8550	0.7528
8	0.8836	0.3903
9	0.5971	0.4456
10	0.5439	0.5651
MEAN	0.7529	0.5292
SD	0.1279	

Gradient Boosting Tree			Random Forest		
Average Correlation	Average RMSE	SD	Average Correlation	Average RMSE	SD
0.7701	0.4944	0.0268	0.7473	0.5377	0.0255