TABLE OF CONTENTS

COVER PAGE ..... **i**
APPROVAL PAGE ..... **ii**
REVISION FREE LETTER OF STATEMENT ..... **iii**
PLAGIARISM FREE STATEMENT ..... **iv**
ACKNOWLEDGEMENTS ..... **v**
PREFACE ..... **vi**
TABLE OF CONTENTS ..... **viii**
LIST OF FIGURES ..... **xii**
LIST OF TABLES ..... **xiv**
LIST OF ABBREVIATIONS ..... **xv**
LIST OF APPENDICES ..... **xviii**
ABSTRACT ..... **ii**
CHAPTER 1 ..... **1**
A. Background 1
B. Problem Statements 4
C. Research Objectives 4
D. Research Benefits 5
E. Literature Review 6
1. SARS-CoV 2 6
2. COVID-19: Epidemiology, Clinical Manifestation, Pathology, and Treatment 8
3. SARS CoV-2 RP1a as a Target Therapy 12
4. Drug Repurposing 14
5. Cheminformatics Analysis for Drug Repurposing 17
a. Definition, Comparison between Bioinformatics and Their Parts 17
1). Cheminformatic database 18

2). Databases Searching Tools 19
3). Property Prediction Tools 20
b. Ligand-Based Virtual Screening (LBVS) 20
1). Definition, List of LBVS Methods 20
2). Descriptor-Based Virtual Screening 21
a). Substructure Keys-based Fingerprints 23
b). Topological Keys-based Fingerprints 24
c). Circular Fingerprints 25
3). Scaffold Mining & Clustering Method 26
4). Maximum Common Substructure (MCS) Aligmnet and SMiles ARbitrary Target Specification (SMARTS) 31
c. Machine Learning for Ligand-based Virtual Screening 33
1). Definition, Purpose, Evaluation of Machine Learning 33
a). Confusion Matrix 36
b). Cohen's Kappa 37
2). List of Most Common Supervised Classification ML 38
a). Random Forest Classifier 38
b). Artificial Neural Network 40
c). Support Vector Machine 41
6. Protein-Protein Interaction (PPI) Analysis 44
7. Programmatic Data Access with Open-Source Facilities 45
8. Theoretical Basics 46
9. Empirical Statement 49
CHAPTER II 50
A. Research Plan 50

B. Research Variables Operational Definition 51
C. Research Materials 52
D. Research Tools 53
1. Computer Specification 53
2. Software 53
a. Konstanz Information Miner (KNIME) 53
b. Cytoscape 54
E. Research Location 54
F. Research Steps 55
1. Preparation of RP1a Protein 56
2. Programmatic Data Access: Retrieval of Protein-ligand structure from Protein Data Bank 56
3. Programmatic Data Access: Bioactivity Research of Potential Ligands from ChEMBL and PubChem 57
4. Gadaleta Standardization 59
5. Fetch substructure using Bemis-Murcko Scaffolds & Hierarchical Clustering 60
6. Generate and Calculate Maximum Common Substructure (MCS) 60
7. Preparation of Potential Compounds Database and Virtual Screening: Substructure Search 61
8. Ligand-based Machine Learning Model Preparation 61
9. Model Usage to Predict Activity Score 62
10. Protein-protein Interaction Network Generation and Literature Analysis 62
CHAPTER III 64
A. Research Concept 64
B. Retrieval & Analysis of Protein & Ligand Target Data 66
C. Scaffold Mining, Hierarchical Clustering, and MCS Alignment 70
D. Candidate Compound Screening with Created MCS 77
E. Ligand-based Machine Learning Model Analysis 80

1. Random Forest 83
2. Artificial Neural Network 84
3. Support Vector Machine 84
F. Machine Learning Model Testing 85
G. Protein-Protein Interaction Analysis of Target Protein 87
H. Literature Analysis of Screened Drug Results 93
I. Molecular Mechanism of Screened Drug in RP1a Inhibition 101
J. Study Limitations 104
CHAPTER IV 106
A. Conclusions 106
B. Recommendations 107
REFERENCES 109
APPENDICES 126

LIST OF FIGURES

Figure 1. Classification of SARS-CoV-2 Virus 7
Figure 2. Pathophysiology of SARS-CoV-2 Leads to RAAS Disturbance and Inflammation 11
Figure 3. Genetic Material (RNA) of SARS-CoV-2 Enters the Host Cell and Borrows Ribosomes to Translate Into 16 Non-Structural Proteins (NSPs) 13
Figure 4. Traditional Drug Discovery vs. Drug Repurposing Method 16
Figure 5. List of Virtual Screening Approach 21
Figure 6. Example of Substructure-based Fingerprint Similarity 24
Figure 7. Example of Hashed Fingerprint Similarity Approach 25
Figure 8. Definitions of Different Types of Fragments in a Molecule 28
Figure 9. All Types of Linkage Method of Hierarchical Clustering 30
Figure10. Illustration of Bemis-Murcko Fragmentation Algorithm with Hierarchical Clustering 31
Figure 11. Illustration of implementation of MCS 33
Figure 12. Illustration of Confusion Matrix 36
Figure 13. Illustration of Random Forest Classifier 39
Figure 14. Illustration of Artificial Neural Network 41
Figure 15. Illustration of Support Vector Machine 42
Figure 16. KNIME Drug Repurposing Workflow 65
Figure 17. KNIME Workflow of PDB Protein-ligand Retrieval 66
Figure 18. KNIME Workflow of ChEMBL Bioactivities Retrieval 67
Figure 19. KNIME Workflow of PubChem Bioactivities Retrieval 68
Figure 20. KNIME Workflow of Creating Murcko Scaffold 71
Figure 21. KNIME Workflow of Hierarchical Clustering 72
Figure 22. KNIME Workflow of MCS Calculation 73
Figure 23. KNIME Substructure Screening Workflow 78

Figure 24. Chart Visualization of Screened Drug List from DrugBank for Each Target 79
Figure 25. KNIME Workflow of Fingerprint Similarity using MACCS and ECFP4 (Morgan) and Machine Learning Model Creation with 10-fold k-Cross-validation..... 82
Figure 26. ROC Curve of Each ML Algorithm with MACCS Fingerprint Method: (a) Random Forest, (b) Artificial Neural Network, and (c) Support Vector Machine..... 83
Figure 27. KNIME Workflow of Model Usage for Compounds Prediction..... 86
Figure 28. List of Screened Candidate Compounds with >90% Confidence 87
Figure 29. Protein-protein Interaction Network of NSP5 or 3CLpro 90
Figure 30. Top 10 Proteins that Interact with NSP5 after analyzed by 'Degree' Hubba Node Parameter in Cytoscape..... 90
Figure 31. Suggested Molecular Mechanism of RP1a with Other Proteins..... 103

LIST OF TABLES

Table I. Comparisons Between Databases, Data Formats, Prediction Methods, Visualization Software and Manipulation Tools Used in Bioinformatics and Cheminformatics 17

Table II. Research Variables 51

Table III. Number of Compounds Available from Different Databases 69

Table IV. Top 15 list of Clusters of Enriched Maximum Common Substructures from Murcko Framework with Its Number of Successfully Screened Candidate compounds 74

Table V. Summary of Machine Learning Parameter with Consideration to Use the Model for Prediction for 3CLpro Targeted Therapy 81

Table VI. Top 10 Protein Interaction of 3CLpro by 'Degree' Hubba Node with Their Physiological Roles 90

Table VII. Literature Analysis of Top 5 Compounds from Top 5 Scaffolds from RF MACCS Model Predictor 95

LIST OF ABBREVIATIONS

3CLpro	3-Like-Chymotrypsin Protease
AAK1	Adaptor-associated Protein Kinase 1
AID	PubChem Assay ID
ANN	Artificial Neural Network
API	Application Programming Interface
ACE2	Angiotensin-Converting Enzyme II
ATP	Adenosine Tri Phosphate
AUC	Area Under Curve
CatB/L	Cysteine Protease Cathepsin B & L
CART	Classification and Regression Tree
CCR5	C-C Chemokine Receptor Type 5
CDER	Centre of Drug Evaluation and Research
ChEMBL	Chemical European Molecular Biology Laboratory
CID	PubChem Compound ID
COVID-19	Coronavirus Disease 2019
DDX39B	DEAD-box helicase 39B
DNA	Deoxy Ribonucleic Acid
ECFP	Extended-Connectivity Fingerprints
EIF2S1	Eukaryotic Translation Initiation Factor 2 subunit 1
EIF4A1	Eukaryotic Translation Initiation Factor 4A1
FACT	Facilitates Chromatin Transcription
FDA	Food and Drug Administration
HCoV	Human Coronavirus
HE-protein	Membrane Glycoprotein
InChI	IUPAC-international Chemical Identifier
IND	Investigational New Drug

IFN	Interferon
ISG15	Interferon-Stimulated Gene 15
IUPAC	International Union of Pure and Applied Chemistry
JKN	Jaminan Kesehatan Nasional
JSON	JavaScript Object Notation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNIME	Konstanz Information Miner
LBVS	Ligand-based Virtual Screening
LBML	Ligand-based Machine Learning
LRPPRC	Leucine-Rich Pentatricopeptide Repeat Containing
MACCS	Molecular Access System
MAPK	Mitogen-activated Protein Kinase
MAVS	Mitochondrial Antiviral-Signaling Protein
MCS	Maximum Common Substructure
MYO1B	Myosin IB
N-protein	Nucleocapsid Protein
NDA	New Drug Application
NEMO	NFκB Essential Modulator
NSAID	Non-Steroidal Anti-Inflammatory Drug
NSP	Non-Structural Protein
ORF	Open Reading Frame
PDB	Protein Data Bank
PharmGKB	Pharmacogenomic Knowledge Base
PLpro	Papain Like Protease
PSMB9	Proteasome 20S Subunit Beta 9
RAAS	Renin Angiotensin Aldosteron System
RDKit	Rational Discovery Kit
RF	Random Forest
RIG-I	Retinoic acid-inducible gene I

RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
S-protein	Spike Protein
SAR	Structure-Activity Relationship
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus
SDF	Structured Data Format
SDHA	Succinate Dehydrogenase complex flavoprotein subunit A
SM-protein	Small Membrane Protein
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular-Input Line-Entry System
SMOTE	Synthetic Minority Over-sampling Technique
SPTBN2	Spectrin Beta, Non-Erythrocytic 2
SVM	Support Vector Machine
TFAM	Mitochondrial transcription factor A
TMRPSS2	Transmembrane Protease Serine 2
TNF-α	Tumor Necrosis Factor-alpha
tRNA	Transcriptional RNA
UniProt	Universal Protein Resources
VOC	Variant of Concern
Xpath	XML Path Language
WHO	World Health Organization
RP1a	Replicase Polyprotein 1a

LIST OF APPENDICES

Appendix 1. QR Code (Github) to Access the Thesis Files 126
Appendix 2. Detailed Version of Parameter Set in Controlled Variable in KNIME Application 126
Appendix 3. Confusion Matrix of RF Classifier for 3CLpro 130
Appendix 4. Confusion Matrix of ANN Classifier for 3CLpro 130
Appendix 5. Confusion Matrix of SVM Classifier for 3CLpro 130
Appendix 6. Binary Labelling KNIME Workflow 131
Appendix 7. KNIME Workflow of Gadaleta Standardization. 132
Appendix 8. Step 1 of Ligand Pre-processing from PDB 133
Appendix 9. PDB Property Downloader and Extractor 134
Appendix 10. Step 3 PDB Retrieval 135
Appendix 11. ChEMBL Step 2 to Find Bioactivity Data 136
Appendix 12. ChEMBL Step to Request the PubMed ID and Download the ChEMBL Document to Enhance the Evidence 137
Appendix 13. ChEMBL and PubChem Workflow to Extract Canonical SMILES 138
Appendix 14. PubChem Step to Request Compound Name 139
Appendix 15. Ligand Retrieval Post-Processing Workflow 140
Appendix 16. ML Model Evaluation 141
Appendix 17. Complete List of Screened Drug Compounds by RF MACCS model with Prediction Confidence Score >90% 141