

# Table of Contents

Table of Contents .....	i
Table of Figures.....	v
Summary.....	ix
Chapter 1. Introduction .....	1
1.1 Motivation.....	3
1.2 Problem Statement .....	4
1.3 Related Studies.....	5
1.4 Overview of the research project .....	5
Chapter 2. Background.....	7
2.1 Molecular representing .....	7
2.1.1. SMILES .....	7
2.1.2. Molfile.....	11
2.1.3. Molecular fingerprints .....	14
2.2 ChemDraw related formats.....	18
2.2.1. ChemDraw Binary Format (CDX) .....	19
2.2.2. Embedded ChemDraw in Microsoft Word documents.....	23
2.2.3. ChemDraw XML (CDXML) .....	28
2.3 Literatures Review .....	29
2.3.1. Chemical Entity Recognition .....	29
2.3.2. Optical Chemical Structure Recognition .....	32
2.3.3. ChemDraw Mining .....	34
Chapter 3. ChemScanner.....	36
3.1 File Reader .....	37
3.1.1. ChemDraw Parser.....	39

3.2	Pre-processing .....	45
3.2.1.	Extract nested objects .....	45
3.2.2.	Graphic refinement .....	47
3.2.3.	Arrow refinement .....	49
3.2.4.	Molecule compilation .....	52
3.3	Scheme Analysis .....	55
3.3.1.	ChemScanner Arrow.....	55
3.3.2.	Analysis Procedure .....	56
3.3.3.	Arrow area builder.....	57
3.3.4.	Role detection.....	59
3.4	Special case processing .....	61
3.4.1.	Molecule group.....	61
3.4.2.	Multiple line chain reactions .....	63
3.5	Text processing .....	65
3.5.1.	Abbreviations and superatoms .....	65
3.5.2.	Text assignment.....	66
3.5.3.	Molecular text processing .....	68
3.5.4.	Reaction text processing .....	71
3.6	Post-processing .....	72
3.6.1.	Text as molecules .....	72
3.6.2.	Multi-step reaction.....	74
3.6.3.	Generation of new elements.....	76
Chapter 4.	ChemScanner User Interface .....	79
4.1	Overview.....	79
4.2	Abbreviation and Superatom management.....	82
4.3	Scanned Output .....	85
4.3.1.	Results display.....	86

4.3.2. Results edit .....	89
4.4 File Storage .....	93
4.4.1. Backend classes .....	93
4.4.2. File Storage interface .....	93
Chapter 5. Chemotion ELN integration .....	100
5.1 Chemical searching .....	100
5.1.1. Exact structure search .....	100
5.1.2. Substructure search.....	101
5.1.3. Similarity search.....	104
5.2 Computational molecule properties interface .....	108
5.3 Green Chemistry .....	113
5.3.1. Atom Economy.....	114
5.3.2. Environmental (E) factor .....	115
5.3.3. Green Chemistry Interface .....	116
Chapter 6. Summary and conclusions.....	118
Further Impact .....	119
Areas for future work.....	119
Chapter 7. Abbreviations.....	121
Chapter 8. Literatures.....	125
Chapter 9. Appendix .....	133
9.1 ChemScanner Usage and APIs .....	133
9.1.1. Usage .....	133
9.1.2. API.....	133
9.2 Curriculum Vitae.....	138
9.3 Acknowledgments .....	140