Predictive Models for Thermal Stability and Explosive Properties of Chemicals from Molecular Structure

THÈSE N^{O} 6970 (2016)

PRÉSENTÉE LE 1ER AVRIL 2016

À LA FACULTÉ DES SCIENCES DE BASE
GROUPE DE SÉCURITE CHIMIQUE ET PHYSIQUE
PROGRAMME DOCTORAL EN CHIMIE ET GÉNIE CHIMIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Nadia BAATI

acceptée sur proposition du jury:

Prof. V. Hatzimanikatis, président du jury
Dr T. Meyer, Prof. F. Stoessel, directeurs de thèse
Prof. G. Storti, rapporteur
Prof. G. Reniers , rapporteur
Prof. K. Sivula, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

To my grandparents…

Acknowledgements

For this PhD project, I have received the advices, help and support of many people and I would like to sincerely thank them as none of it would have been the same without their contributions.

I would like to thank MER Dr. Thierry Meyer and Prof. Francis Stoessel for giving me the opportunity to conduct this PhD thesis under their supervision. I am highly grateful for their mentorship and guidance and honored by their trust. I appreciated the fine balance between the independence they gave me, the critics and comments that helped me move the project forward and the motivating encouragements.

Dr. Annik Nanchen deserves my deepest gratitude for all her contributions to this project, for suggesting it in the first place, for the constructive discussions and creative ideas, and for being a source of inspiration and motivation, even beyond the PhD.

I am very thankful to Prof. Genserik Reniers, Prof. Giuseppe Storti, and Prof. Kevin Sivula for being available as jury members, and Prof. Vassily Hatzimanikatis, as jury president of this dissertation. I am grateful for their interest and their valuable feedback. In particular, Prof. Sivula and Prof. Hatzimanikatis are warmly thanked for taking part in the first year exam of this PhD, and for their mentorship since many years, the enlightening discussions, the useful tips and hints, and for always being helpful, even on matters out of the thesis scope. It was also a great pleasure to be involved as a teaching assistant on various courses with both of them.

I would like to express my appreciation to Dr. Christoph Heuberger for welcoming me during a few days at the Novartis Safety Lab and for enabling this research to reach an unexpected extent thanks to the data he made available; and to Dr. Bertrand Roduit and Prof. Rafiqul Gani, who made available two catalysts for this work, the AKTS and ICAS softwares.

I also learned a lot with the collaborators at Swissi Process Safety and for that I would like to express my gratefulness to all of them, especially to Dr. Georg Suter, Dr. Jean-Michel Dien, Dr. Mischa Schwaninger, Dr. Serge Forestier, Manfred Eiche, and Delphine Berset.

It was a real pleasure to work with all EPFL SB-SST members: Dr. Sebastian Brückner and

Dr. Kirstin Friedrich with whom I shared the office, Dr. Valeria Granata, Dr. Amela Grošo, but also Paule Anken and her crew Sandrine Pfister, Kim Godat, Maëlle Traufer, and Laurine Vermeille. I am also happy I shared the PhD experience with Dr. David Plüss, Charles Guinand, and Laetitia Mage. With her master thesis, Laetitia greatly contributed to this project, and with her wittiness and cheerfulness, she was the life of the party in the office. To the super-energetic Charles, I wish the best of luck for the upcoming challenges. These years would not have been the same without David's friendship and our mutual support through the ups and downs of PhD, many thanks!

From other labs and departments at EPFL, I would like to thank people who always welcomed me with an open door and helpful advice, in particular, Dr. Christophe Roussel, Prof. José Millan Del Rocio, Dr. Noushin Hadadi, Meriç Ataman, Anne Lene Odegaard, and Yoann Dind, and Prof. Pierre-Alain Carrupt from Université de Genève.

Then, to my aunt Mouhiba Jamoussi, and my mother Nabila who read my thesis entirely - more than once - for proofreading and and helping me present this work in a readable dissertation, thank you so much!

Finally, I have been lucky to be surrounded with wonderful friends and a loving family and for this I owe them all praise and thankfulness. I thank all my friends who made these years really enjoyable, especially Eda, Younes, Rima, Flavien and Helena. Pauline has been so supportive and comforting through the craziest parts of my PhD; for that and for the realized dreams and joyful memories, I am really grateful. At last, I would like to thank my family in Tunisia, my sister, and above all my parents, for their constant support and guidance throughout my studies, actually throughout my entire life, and their invaluable love.

Lausanne, March 2016 N. B.

Abstract

Industrial chemical processes may involve thermal risks as most of the reactions performed are exothermic, the chemicals used are often thermally unstable, and the operating conditions are set to induce high conversion and throughput. Besides the reactive steps, all operations from mixing to storage and from processing to transport involving sensitive chemicals should be conducted under strictly controlled conditions ensuring safe operations. Performing an efficient risk assessment and implementing the proper risk mitigation measures are essential to avoid, or at least reduce, accidents and their potentially disastrous consequences.

For optimal design and implementation of safety measures, it is important that these considerations are taken into account at early stages of process development. The required data should be made available at the appropriate time so it can be properly accounted for and efficiently serve the design. Yet, at early design phases, some information may be unavailable due to several reasons: the process design being still in development, some parameters can be unknown; experimental analysis of chemicals could be hindered or impossible due to products availability in required quantities; several alternatives are under investigation which raises the necessary resources (time and material) for experimental tests.

Predictions would be the appropriate response to such scenario. The aim of this dissertation is to develop predictive models for two hazardous behaviors of chemicals: explosive sensitivity and thermal stability. For the models to be applicable at early development phases, it is preferable to minimize the information feed requirements, and therefore, structure-based approaches are applied. Two methods were identified: Quantitative Structure-Property Relationships (QSPR) and Group Contributions Method (GCM).

The hazardous behaviors are studied through characteristic measurements: the Minimal Ignition Energy (MIE) to represent explosive sensitivity and Differential Scanning Calorimetry (DSC) for thermal stability. These measurements are widely employed in safety studies and deliver necessary information to identify potential hazards. Moreover, their specificities call for predictive models: MIE tests require repetitive analysis and hence are time and material consuming; regarding DSC experiments, experts have noted that they seem to exhibit structurally dependent features, and so far no study has comprehensively investigated this phenomenon.

This work presents in a first part the structure-based approaches that are applied and the elements of Data Analysis necessary for developing predictive models and simulating experimental results. Secondly, both experimental analysis are detailed and the important information our models should be able to represent will be exposed. Finally, the third and last part is dedicated to the applications: the obtained predictive models are presented, evaluated and discussed. Most of the initial objectives are met as efficient solutions are proposed, nonetheless, some improvement strategies may also be considered.

Key words: Safety, Modeling, Molecular Simulation, Data Analysis, Machine Learning, Thermal risks, Explosion risks.

Résumé

Les procédés de l'industrie chimique présentent souvent des risques thermiques, notamment dus aux faits que les réactions sont exothermiques, les réactifs instables et les conditions favorables à des conversions et des rendements importants. Mises à part les étapes de réactions, toute autre étape d'un procédé, que cela soit le stockage, le mélange ou le transport de produits chimiques sensibles, doit se faire dans des conditions contrôlées et maitrisées afin d'assurer la sécurité des opérations. Il est alors nécessaire d'évaluer les risques encourus et de mettre en place des mesures de prévention et de protection afin d'éviter, ou tout du moins de minimiser, les accidents et leurs conséquences potentiellement désastreuses.

Il est alors important que ces éléments soient pris en considération dès le début du développement d'un procédé afin d'optimiser la conception de mesures de sécurité adéquates et leur intégration au procédé. Un grand nombre de données de sécurité sont alors nécéssaires. Or, lors des premières étapes de développement de procédé, ces informations ne sont pas toujours connues : le procédé dans son ensemble étant en cours de développement, certains paramètres sont inconnus ; il se peut que des produits chimiques ne soit pas disponibles en quantité suffisante pour être analysés expérimentalement ; finalement, si plusieurs produits représentent des alternatives à une même fonction, déterminer le meilleur candidat expérimentalement représente un coût non négligeable (en termes de temps et de produits).

L'utilisation de modèles prédictifs pourrait apporter une réponse à ce type de problèmes. Cette dissertation a pour but de développer des modèles prédictifs pour deux phénomènes dangereux : l'explosivité et la stabilité thermique. Afin de rendre accessibles ces modèles dès le début du développement d'un procédé, ils ne devront nécessiter que des informations facilement disponibles, telles que la structure chimique du composé étudié. Il existe principalement deux types de modèles se basant sur la structure chimique : la méthode des incréments de groupes et la méthode des modèles quantitatifs de relation structure-propriétés (en anglais, Group Contributions Method (GCM) et Quantitative Structure-Property Relationships (QSPR), respectivement).

Les deux phénomènes dangereux étudiés sont caractérisés par deux analyses expérimentales : l'explosivité peut se mesurer par l'Energie Minimale d'Ignition (EMI), quant à la stabilité thermique, elle est souvent évaluée par Calorimétrie Différentielle à Balayage (en anglais, Differential Scanning Calorimetry ou DSC). Ces deux analyses sont très largement employées lors des études de sécurité car elles permettent d'identifier les dangers liés aux produits analysés. Cependant, la mesure d'une EMI requiert un grand nombre de répétitions d'un même test, ainsi, le temps et les quantités de produits nécessaires pour une mesure pourraient être réduits par l'emploi de modèles prédictifs. Quant à la DSC, les experts notent que les résultats expérimentaux suggèrent une dépendance à la structure chimique. Or, il n'y pas eu d'étude complète de ce phénomène pour révéler cette dépendance et proposer des modèles. Dans une première partie, les types de modèles basés sur la structure chimique qui sont employés ici sont présentés, ainsi que quelques approches d'analyse de données qui permettent de développer des modèles prédictifs. Dans une seconde partie, nous revenons sur les analyses expérimentales, leurs procédures et leurs résultats afin de déterminer les informations importantes que les modèles doivent refléter. Enfin, la troisième partie comprend l'exposé des résultats : les modèles prédictifs obtenus sont présentés, évalués et discutés. Les objectifs initiaux sont pour la plupart atteints, toutefois, quelques améliorations envisageables sont à discuter.

Mots clefs : Sécurité des Procédés, Modélisation, Simulation Moléculaire, Analyse de Données, Apprentissage Automatique, Risques Thermiques, Risques d'Explosion.

Zusammenfassung

Industrielle chemische Prozesse können thermische Risiken beinhalten, da die meisten Reaktionen von exothermischer Natur sind, die Chemikalien oft thermisch instabil und die Reaktionsbedingungen wegen der hohen Kapazität oftmals kritisch sind. Abgesehen vom eigentlichen chemischen Prozess verlangen aber auch alle anderen involvierten Prozesse, wie Transport, Lagerung oder Vermischung nach einer streng kontrollierten Umgebung, welche für einen sicheren Betrieb unabdingbar ist. Dafür ist die Durchführung einer wirkungsvollen Risikobeurteilung und die Einführung von entsprechenden Massnahmen zur Risikominderung zwingend notwendig um Umfälle und deren möglichweise katastrophalen Auswirkungen zu verhindern oder zumindest zu vermeiden.

Um die Konzeption und Implementierung von Sicherheitsmassnahmen zu erleichtern, sollten diese an einem möglichst frühen Zeitpunkt der Prozessentwicklung bedacht werden. Die dazu benötigten Daten müssen rechtzeitig zur Verfügung stehen um entsprechend in die Entwicklung einzufliessen, jedoch stehen diese Informationen gerade in frühen Phasen der Entwicklung aus verschiedenen Gründen nicht zur Verfügung: der Prozess befindet sich noch in Entwicklung, gewisse Parameter sind noch unbekannt, experimentelle Analysen werden erschwert oder verhindert aufgrund mangelnder Verfügbarkeit von Rohprodukten in ausreichender Menge, verschiedene Alternativen stehen zur Diskussion, was Ressourcen bindet.

Vorhersagen sind eine mögliche Herangehensweise um den Mangel an Daten zu verringern. Das Ziel dieser Dissertation ist deshalb die Entwicklung von Modellen zur Vorhersage zweier gefährlicher Eigenschaften von Chemikalien: Explosionsempfindlichkeit und thermische Stabilität. Damit diese Modelle in frühen Entwicklungsphasen anwendbar sind, werden die Anforderungen an Messdaten geringgehalten und struktur-basierte Herangehensweisen angewendet (Quantitative Struktur-Wirkungs-Beziehung und Gruppenbeitragsmethoden).

Die gefährlichen Eigenschaften werden durch charakteristische Messungen untersucht: Tests der Mindestzündenergie um die Explosionsempfindlichkeit und die dynamische Differenzkalorimetrie um die thermische Stabilität zu bestimmen. Diese Methoden werden häufig in Sicherheitsstudien angewendet und können wertvolle Informationen zur Identifikation von möglichen Gefahren liefern. Zusätzlich eignen sich ihre Eigenschaften für Vorhersagemodelle: zum einen sind zur Bestimmung der Mindestzündenergie Testreihen nötig und damit ist diese Methode zeit- und materialintensiv. Zum anderen weist die dynamische Differenzkalorimetrie auf mögliche strukturabhängige Eigenschaften hin, was jedoch noch in keiner Studie vollständig analysiert wurde.

Zusammenfassung

Diese Arbeit präsentiert in einem ersten Teil die strukturabhängigen Herangehensweisen und die Elemente der Datenanalyse welche für die Entwicklung der Vorhersagemodelle verwendet werden. In einem zweiten Teil werden die experimentellen Analysen beschrieben und die durch Vorhersagen gewonnen Parameter beschrieben. Im dritten und letzten Teil werden die Vorhersagemodelle angewendet: die Modelle werden vorgestellt, evaluiert und besprochen. Die meisten Ziele dieser Arbeit werden dabei erfüllt und als wirksame Methoden für verschiedene Anwendungen vorgeschlagen; die möglichen Strategien zur Verbesserung dieser Methoden werden ebenfalls besprochen.

Stichwörter: Prozesssicherheit, Modellierung, Molekulare Modellierung, Datenanalyse, Maschinelles Lernen, Thermische Risiken, Explosionsrisiken.

Contents

Acknowledgements **i**

Abstract (English/Français/Deutsch) **v**

List of Figures **xvii**

List of Tables **xix**

Introduction **3**

Context 3
Aim 6
Outline 7

I Predictive Modeling from Molecular Structure **11**

1 Molecular Structure Based Modeling **13**

1.1 Group Contribution Methods 14
1.1.1 Background 14
1.1.2 Basic Principles 15
1.1.3 Frameworks and Groups 16
1.2 Quantitative Structure-Property Relationship 19
1.2.1 Background 19
1.2.2 Principles 21
1.2.3 Descriptors 23
1.3 Conclusion 26

2 Data Mining and Machine Learning: Approaches and Good Practices **29**

2.1 Definitions 29
2.2 Feature Selection and Dimension Reduction 30
2.2.1 Filter 31
2.2.2 Wrapper 31
2.2.3 Principal Component Analysis 32
2.3 Classification 33
2.3.1 Cluster Analysis 33

2.3.2 Decision Trees 36
2.3.3 Linear Classifiers 38
2.4 Regression Models 39
2.4.1 Multiple Linear Regression 39
2.4.2 Stepwise Regression 39
2.4.3 Neural Networks 40
2.5 Good Practices 42
2.5.1 Partitioning Training-Validation 45
2.5.2 Model Evaluation Criteria 48
2.5.3 Statistics for Experimental Data Analysis 52
2.6 Conclusion 55

II Experiments, Data Preprocessing and Extraction 59

3 Minimum Ignition Energies 61
3.1 Definition and Use 61
3.2 Experimental Measurement 66
3.3 Influencing Parameters 69
3.4 Data Collection and Treatment 70
3.5 Conclusion 70

4 Differential Scanning Calorimetry 71
4.1 DSC Principles 73
4.1.1 Power Compensation DSC 75
4.1.2 Heat-Flux DSC 77
4.2 Thermogram Analysis 79
4.3 Data Collection, Treatment and Property Extraction 84
4.3.1 Data Collection 85
4.3.2 Thermogram Parametrization 87
4.3.3 Property Extraction and Data Treatment 89
4.4 Conclusion 91

III Applications 93

5 MIE Modeling 95
5.1 Literature Review 95
5.2 Resulting Models 99
5.2.1 Global Models 99
5.2.2 Local Models 101
5.2.3 Sensitivity Classification 106
5.3 Model Interpretation 110
5.4 Conclusion 111

6 DSC Models **115**
- 6.1 Literature Review . . . 115
  - 6.1.1 Group Contribution Models of Thermal Properties . . . 115
  - 6.1.2 Quantitative Structure-Property Relationships . . . 118
- 6.2 Experimental Error and Confidence Interval . . . 122
- 6.3 Local and Global Models . . . 126
  - 6.3.1 Objectives . . . 126
  - 6.3.2 Nitro Compounds Study . . . 126
  - 6.3.3 First Generalization . . . 130
  - 6.3.4 Chemical Families . . . 135
  - 6.3.5 Global Models . . . 137
- 6.4 Systematic Construction of Local Subsets . . . 140
  - 6.4.1 Modified Modeling Procedure . . . 140
  - 6.4.2 Features Space Clustering . . . 142
  - 6.4.3 DSC-based Clustering and Classification . . . 143
- 6.5 Model Application . . . 147
- 6.6 Conclusion . . . 156

Conclusions **161**

Appendices **167**

Minimal Ignition Energies Data **169**

Differential Scanning Calorimetry **183**
- B.1 Tolerance Zone Construction . . . 183
- B.2 Nitro Compounds Study . . . 185

Confidential Data **189**
- C.1 Chemical Families Study . . . 189
- C.2 Models on Clusters . . . 200
- C.3 DSC Properties Clustering . . . 201

Bibliography **228**

Abbreviations **229**

Copyright Credits **233**

Curriculum Vitae **235**

List of Figures

1.1 Molecular Structure Based Modeling II . . . 13
1.2 Representation Example in Marrero-Gani Framework . . . 18

2.1 Filtering . . . 32
2.2 Principal Component Analysis . . . 33
2.3 Cluster Analysis . . . 34
2.4 Hierarchical Clustering . . . 35
2.5 K-means Cluster Analysis . . . 36
2.6 Decision Tree . . . 37
2.7 Linear Classifiers . . . 38
2.8 Stepwise Regression . . . 41
2.9 Artificial Neural Network . . . 42
2.10 Error Dependance to Number of Parameters . . . 43
2.11 Information Content in Parameters . . . 44
2.12 Model Selection . . . 45
2.13 Cross-Validation . . . 46
2.14 Bootstrap Validation . . . 47
2.16 Molecular Structure Based Modeling Procedure II . . . 57

3.1 MIE and Electrostatic Discharges . . . 65
3.2 MIKE 3 Apparatus . . . 67
3.3 Statistic Energy . . . 68

4.1 Calorimeter of Lavoisier and Laplace . . . 73
4.2 Power-compensated DSC . . . 76
4.3 Heat-Flux DSC . . . 77
4.4 DSC Thermogram Example . . . 80
4.5 Purity Effect on Melting . . . 80
4.6 Onset Temperature and Reaction Enthalpy . . . 82
4.7 Kinetic Behavior on DSC . . . 84
4.8 DSC for Kinetic Evaluation . . . 84
4.9 Effect of Enclosing Atmosphere on DSC . . . 86
4.10 Effect of Sample Weight on DSC . . . 87
4.11 Gaussian and Fraser-Suzuki Models . . . 89

4.12 DSC Key Properties . . . 90

5.1 MIE Global Model . . . 100
5.2 Local Models for MIE of Dusts . . . 102
5.3 Best Local Model for MIE of Dusts . . . 103
5.4 Local Model for MIE of Liquids . . . 104
5.5 Local Models for MIE of Gases . . . 105
5.6 Local Model for MIE of Liquids and Gases . . . 105
5.7 Classification Tree . . . 107

6.1 Tolerance Zone . . . 125
6.2 DSC Nitro Models Responses . . . 128
6.3 DSC Reconstructions from Nitro Models . . . 129
6.4 DSC Models for Miscellaneous Set . . . 131
6.5 DSC Reconstructions from Miscellaneous Set Models . . . 132
6.6 DSC Reconstructions Comparison . . . 134
6.7 Chemical Families Models Responses . . . 137
6.8 DSC Properties Distribution . . . 139
6.9 Global Models Responses . . . 139
6.10 Modified Modeling Procedure . . . 141
6.11 DSC Decision Tree . . . 144
6.12 DSC Clusters Models Responses . . . 146
6.13 DSC Reconstructions from DSC Cluster Models . . . 147
6.14 Average Thermogram for Cluster 2 . . . 149
6.15 Risk Matrix . . . 152
6.16 Examples of DSC Reconstructions . . . 156

A.1 MIE Classification Tree . . . 176

C.1 DSC Models for Nitroso and Nitrites . . . 194
C.2 DSC Models for Azo and Tetrazoles . . . 195
C.3 DSC Models for Phenylamines . . . 196
C.4 DSC Models for Nitriles . . . 197
C.5 DSC Models for Ethers . . . 198
C.6 DSC Clusters Representation . . . 201
C.7 DSC Models for DSC Cluster 1 . . . 206
C.8 DSC Models for DSC Cluster 2 . . . 207
C.9 DSC Models for DSC Cluster 3 . . . 208
C.10 DSC Models for DSC Cluster 4 . . . 209
C.11 DSC Models for DSC Cluster 5 . . . 210
C.12 DSC Models for DSC Cluster 6 . . . 211
C.13 DSC Models for DSC Cluster 7 . . . 212
C.14 Examples of DSC Reconstructions Comparison . . . 214

List of Tables

1.1 Standard Enthalpy of Formation 16
1.2 Marrero-Gani Groups for 6-hydroxy-2-methylbenzoxazole 18
1.3 Examples of Structural Descriptors for 6-hydroxy-2-methylbenzoxazole 25
1.4 Comparison of Molecular Structure Based Modeling Methods 27

2.1 Comparative Evaluation of Regression Models 51

3.1 Zones Definitions of Explosive Atmospheres 66

5.1 Comparative Summary of Literature Models for Explosive Characteristics . . . 96
5.2 Comparison of E_{ES} from Literature Models 98
5.3 Global Model Evaluation Summary 101
5.4 Comparative Evaluation Summary of Dusts Models for MIE 102
5.5 Ignition Sensitivity Categories of Powder Suspensions 106
5.6 Classification Tree Parameters 108
5.7 MIE Post-Classification Models 109
5.8 Cyclohexanone Peroxide Structure 110

6.1 Comparative Summary of Literature Models for DSC-Derived Data 121
6.2 Repeatability Coefficients 123
6.3 Tolerance Zone Coefficients 125
6.4 Model Investigation Plan 127
6.5 Nitro Models Evaluation Summary 129
6.6 Miscellaneous Set Models Evaluation Summary 130
6.7 Global Models Evaluation Summary 138
6.8 Reduced Feature Space 142
6.9 Clusters Models Evaluation Summary 143
6.10 DSC Clusters Models Evaluation Summary 145
6.11 Molecules treated as Examples 148
6.12 Example of DSC Cluster Assignment 149
6.13 Examples of Enthalpy Calculation with Chemical Family Models 153
6.14 Examples of Enthalpy Calculation with Cluster Models 155

A.1 MIE Values for Gaseous Compounds 169

A.2 MIE Values for Liquid Compounds 170
A.3 MIE Values for Solid Compounds 172
A.4 MIE Values Sources 173
A.5 Global MIE Parameters 174
A.6 MIE Dust Model C Parameters 175
A.7 MIE Classes Models Parameters 177
A.8 Models Responses 178

B.1 Nitro Compounds Models Parameters 185
B.2 Nitro Models Responses 186
B.3 Miscellaneous Set Models Parameters 187
B.4 Miscellaneous Set Models Responses 188

C.1 Chemical Families 190
C.2 Chemical Families Models Evaluation Summary 191
C.9 Clusters Models Evaluation Summary 200
C.10 DSC Clusters Models Evaluation Summary 203

Introduction

Context

For the chemical industry, risk management and hazard assessment represent a priority and a necessity to ensure safety for the workers, the company, the society and the environment. In the past decades, awareness rose among all involved stakeholders and the regulations and legislation on chemicals and their use are continuously revised, harmonized, and tightened in order to minimize the risks through increasing knowledge of chemical intrinsic properties and behaviors. Besides the legal implications, it is also a moral obligation for chemical engineers to take into account process safety and manage the faced hazards.

Yet, this did not come naturally. Most of the current legislation ensues from the responses to major disasters.

In 1976, in Seveso, Italy, during a production of 2,4,5-trichlorophenol, an inappropriate heating system was employed and it caused an overheating of the reaction mass that could not be cooled down before the plant was shut down for the week-end. The heat excess triggered an exothermic decomposition reaction which lead to a consequent temperature rise, pressure build-up and production of a toxic compound, 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), or dioxin. Eventually, the pressure accumulation triggered the emergency pressure relief system and the reactor's rupture disk broke open, releasing the poisonous gas in the atmosphere [Kletz, 1999]. As the accident occurred over the week-end, no direct fatalities were caused by the accident, nonetheless indirect damages were caused to the populations of the neighboring cities and the environment: skin lesions, thousands of dead animals and elimination of the remaining to avoid further contamination of the food chain, soil-contamination, etc [Homberger et al., 1979]. Besides, long-term effects of the exposure to TCDD were evaluated and have shown that cancer risks were significantly higher in the affected regions [Pesatori et al., 2009].

Following this catastrophic event, the European Union passed a new law known as the Seveso Directive in 1982 to improve safety on industrial sites managing hazardous substances presenting major-accident risks. The Directive was revised in 1996 and 2012 (Seveso III [Council of the European Union, 2012]).

Less than ten years later than the Seveso accident, the Bhopal disaster occurred in 1984 in India, at a Union Carbide Corporation (UCC) production plant. A runaway reaction generated important temperature and pressure rise that eventually caused the vessel explosion and let the leak of its content into the atmosphere. The exact composition of the poisonous gas cloud is unclear, but comprised mainly methyl isocyanate (MIC). However, according to internal and external reports, very few basic safety principles were in force that day: the synthesis route employed was hazardous and the company had the possibility to conduct production through another route that did not involve MIC; contrary to company policies in different sites in Europe or North America, their Bhopal site stored large quantities of MIC for long periods; the safety instruments were either absent or unavailable such as the Vent Gas Scrubber which was of inappropriate dimensions and not in use, the cooling system was not in use either, the monitoring system which was neither comprehensive nor automatized; finally the workers were poorly trained and the overall site maintenance was practically nonexistent and multiple signs of corrosion were showing on the equipment [Chouhan, 2005, Eckerman, 2005]. The Bhopal disaster holds the highest death toll of major industrial accidents, claiming the lives of thousands of people within hours following the exposure and causing permanent injuries to hundreds of thousands. Estimations announce a total of 14 000 deaths and 730 000 injuries [Eckerman, 2004].

More recently, major accidents with heavy consequences for people (fatalities) and the environment also occurred in Schweizerhalle in Switzerland, Enschede in the Netherlands, Toulouse in France, and Buncefield in the United Kingdom. These events raised the awareness of the hazards of handling chemicals at industrial scales with major consequences on neighboring population and environment. International and national authorities of many countries revised their regulations in order to prevent other major accidents of this extent.

The above-mentioned Seveso Directive obliges operators of hazardous chemicals to take into their responsibilities to prevent the occurrence of major accidents by implementing all the preventive measures and limit their consequences on human health and environment. Safety reports shall be delivered to certify the application of the preventive policies and the implementation of a safety management system. In cases of modification of an existing system, the necessary adaptations and updates to the safety measures shall be taken into account and reported. Emergency plans must be prepared and tested. The information to the public, its consultation in decision-making and the measures to be taken following a major accident are also covered by the Directive.

In Switzerland, the Ordinance on Protection against Major Accidents was adopted in 1991 and revised several times since then, following the first Seveso Directive implementation by the EU to align with regulations. The Ordinance stipulations are highly similar to the obligations in force under the Seveso Directive [Swiss Confederation, 1991].

Besides the legislative framework and the moral obligations, industrial stakeholders are also increasingly sensitive to the costs incurred and the damages to the brand image that arise from accidents, even minor ones. Therefore, Risk Assessment and Risk Mitigation methods are more systematically implemented to ensure the operational and economical well-being of the companies. This may be integrated within broader safety management framework together with training and information of employees, characterization of materials, safer processes, design of the plant and the buildings, safe organization and working environment.

Particularly in the context of process design for the chemical industries, risk assessment is crucial and should be rigorously conducted to identify all hazards included to the process, the damage they could induce, the targets they threat and the conditions under which they could be eliminated. Moreover, where potential threats are identified, the safety measures to prevent their fulfillment should be implemented, and their reliability verified.

For instance, when a process unit or operation may generate an explosive atmosphere, particular precautions should be put in place in order to avoid the ignition of the explosive atmosphere. For this, information regarding the process, the chemicals, and the operating conditions must be gathered and thoroughly analyzed to estimate the probability of formation of the explosive cloud, its sensitivity or flammability, and the ability of the surrounding equipment or operations to provide a sufficient energy input to ignite it. Mitigation measures shall be taken to avoid all these elements: venting or diluting to impeach the explosive atmosphere formation, use of particular equipment and organizational measures that will avoid contact with ignition source, etc. These considerations fall under the EU ATEX Directive relating to equipment and protective systems intended for use in potentially explosive atmospheres [Council of the European Union, 1999].

In both Seveso or Bhopal accidents, runaway reactions occurred and were at the final stages of the event unfoldings, preceding only the vessel bursting and the discharge of its content. Thermal runaway reactions usually are highly exothermic reactions that produce higher amounts of heat than the surrounding system can remove, resulting in a heat accumulation and temperature rise, which intensifies the reaction rate which in turn leads to higher heat production. Due to the temperature rise and the products generation, the pressure builds up and can eventually lead to the explosion of the container. This depends on all elements interacting here, the initial reactant, the decomposition products, that may decompose as well, the potential energy release, and the reactor pressure and temperature management systems, their capacity and availability.

Additionally, the most effective safety measures are planned on the earlier stage of process development. Indeed, the early phases involve important decision-making, and the integration of safety considerations at this level implies highly influential choices and cost effectiveness. Ultimately, process safety in general should tend to propose inherently safer designs. Inherent Safety concept has been formalized by Kletz [2003] and relies on four pillars :

- minimize the quantities of hazardous materials handled;
- substitute a hazardous compound with a less hazardous one;

- moderate the potential effects of residual risks;
- simplify the system and avoid additional equipment or features.

These principles offer great potential for safety enhancement, and should be iteratively implemented at various stages of process development. Their impact is optimal when influencing initial choices regarding the selection and development of the process reactions and equipment [CCPS, 2009]. Moreover, at early design stages, the flexibility is still high and the costs of change low, whereas modifications brought later would incur higher financial costs.

During the risk analysis, very specific data are required to enable decision making regarding the necessary safety measures to implement. This information may be gathered mainly from three different sources: literature, knowledge and experimental analysis. Scientific works, Material Safety Data Sheets, company own or public databases represent large collections of the physico-chemical properties of chemicals and various characteristic data. When this information does not meet the specific conditions of the particular process considered, one could rely on its own knowledge and previous experience to evaluate the safety of integration of a given chemical into a given process. Nevertheless, the most reliable information source would be to experimentally investigate the properties to be determined by emulating the process conditions at smaller scales.

However, the experimental evaluation could be practically impossible. Nowadays, simulations are widely used for the process, product and production design. A classical product design example is drug discovery, which relies on *in silico* modeling to target compounds with specific structural features that indicate the compound could have corresponding biological activities. Among several products with potential interest, only a few will pass all the simulations screening phase and be physically synthesized and made available for laboratory analysis. With the extension of this approach to various fields, an increasing number of products are investigated virtually before being produced. Hence, only their properties for which simulations exist could be estimated.

Main Goal of the Project

The main intention throughout this project is to propose predictive models of process safety related data, in particular explosive sensitivity to ignition and thermal stability, in order to enable their early estimations, or their integration into the context of product design. Hence a major requirement is that these models would rely on restricted information that could be easily available. Therefore, the molecular structure of the chemicals is set as the primary information input to the models.

The safety data investigated here are the Minimal Ignition Energy (MIE) and the Differential Scanning Calorimetry (DSC) thermograms of chemicals. These two characteristics were chosen for their relevance in Process Safety. Moreover, the MIE procedure is time extensive and could be unburdened through simulations. Regarding DSC thermograms, evidence of correlation to the molecular structure has been established, but deserves to be extended further.

Outline of the Project

This work is divided in three parts.

Part I: Predictive Modeling from Molecular Structure

Chapter 1: Molecular Structure Based Modeling details the two main methodologies to apprehend the modeling of physico-chemical properties from the structure of chemicals, namely, the Group Contribution Methods, and the Quantitative Structure-Property Relationships. Their differences and similarities are exposed through examples of various applications. The principles of each method are explained in order to assess how they could be the most helpful to fulfill our objectives.

Chapter 2: Data Mining and Machine Learning presents the necessary mathematical and statistical data manipulations that could be applied in Predictive Modeling. Indeed, there is not a single protocol to develop correlations and many of these techniques are interchangeable or can be used in combination. This theoretical review will support the proposition of the most adapted protocol to the problem tackled here.

Part II: Experiments, Data Preprocessing and Extraction

Chapter 3: Minimum Ignition Energy defines this characteristic property and its applications. Then, the experimental analysis through which MIE are measured is explained, and the influential factors are reviewed. The data gathered for this present study were collected from literature, thus we will see the collection and treatment process to build a readily practicable dataset.

Chapter 4: Differential Scanning Calorimetry presents the basic principles and experimental analysis of DSC. As there are two different functioning principles to DSC apparatus, both will be detailed. However, both techniques lead to similar results, essentially the curve of the heat-flow as a function of temperature, namely a thermogram. An example of analysis of a typical thermogram is then performed to highlight the information that can be collected. The thermogram is interpreted as a combination of few key characteristics. These extracted properties are modeled separately, and this allows recovering the full DSC curve, limiting thus the data loss.

Part III: Applications

Chapter 5: MIE Models presents the first results of this project, as the modeling techniques and the experimental data come together and as the correlations are developed. Several models are proposed, as the studied set is varied. A global model is developed from the entire dataset of collected MIE without distinctions in order to maximize the generalization. Then, subsets are created based on the physical state and specific models are developed, referred to as local models. Finally a sensitivity classification method is proposed as a simple decision tree. An interpretation of the proposed models is discussed.

Chapter 6: DSC Models exposes the models developed in the case of the DSC study. As for the MIE, a global model is proposed for all the available data. Then, local models are proposed for the specific cases. Several criteria serve to divide the set into subsets: local subsets were defined based on chemical families, analysis of their structural similarities, and finally analysis of their DSC similarities. The most suitable protocol that comes out from these attempts is a combination of global classification and local regression and results in the most accurate predictions.

I

Predictive Modeling from Molecular Structure

1 Molecular Structure Based Modeling

There are two main methodologies to correlate physico-chemical properties of chemicals to their molecular structures, namely, Group Contributions Methods and Quantitative Structure-Property Relationships (herein referred to as GCM and QSPR respectively). The procedure to develop relationships between physico-chemical features of compounds and their molecular structure is simplified in Figure 1.1 and is common to both methods.

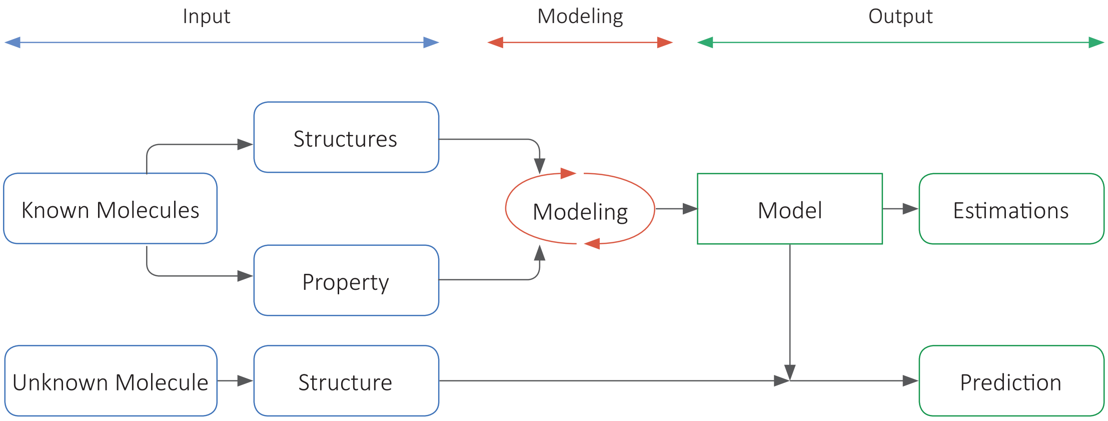


Figure 1.1 – Schematic Procedure for Development of Structure-Based Predictive Models

First of all, the set of "Known Molecules" to be studied is selected. This selection can either be motivated by the interest in their similar - or different- macroscopic behavior or their similar structural characteristics. The property of interest is then experimentally determined or measured for all observations and gathered in the "Property" space. "Structures" refer to the structural characteristics of the observed molecules that will be used to describe the "Property". In this case, the structures will differ between GCM and QSPR, as GCM will rely on Groups to describe molecules, while QSPR rely on numerical descriptors. The mathematical equations between the "Property" and "Structures" spaces are named "Models". The typical model for

GCM is a linear regression, while QSPR methods include all kinds of mathematical correlations. Therefore the "Modeling" step degree of complexity may vary, and for simplification purposes, it is here represented as an iterative loop.

Finally, the model is applied to the structures of molecules of unknown properties, the "Unknown Molecules", and allows to predict them. When applied to the molecules of known properties, it delivers estimate values that can be compared to the actual "Property" values to evaluate the model. For each independent property investigated, another model is developed.

The efficiency of these estimations depends on all steps of the procedure. As the choice of molecules and the experimental determination of the "Property" space are related to the applications of the method, they are detailed later on in Parts II and III. On the other hand, the various ways to represent the structures and develop predictive models are invariant tools and will be the subject of this first part.

In this chapter, the two main methodologies to correlate structures of chemicals and physico-chemical properties, GCM and QSPR, are introduced. The purpose is the same and the applications are highly similar, yet the underlying principles differ and each offers certain advantages or limits compared to the other. Therefore, they are here distinguished and the following sections will present the historical evolution of these methods, their basic principles and some of their most encountered applications while highlighting their similarities, differences and complementarity and thus to grasp the interest to apply both of them in this work.

1.1 Group Contribution Methods

1.1.1 Background

Chemists have always been concerned by the hazardous behaviors of compounds and since the 1950's, they attempted to understand their relationship to chemical structure. One of the earliest articles found reporting such work is this of Calcote et al. [1952] which analyzes the effect of molecular structure on the minimum spark ignition energies for various fuels. They highlighted the influence on ignition sensitivity of branching or unsaturation by comparing homologous series of hydrocarbons, or the effects of closed and/or aromatic compounds. Finally, the presence of hetero-atoms in various substituents (alcohol, nitro or thiol groups) was also analyzed.

Later, thermal decompositions were the focuses of a similar study. Blake and colleagues conducted thermal analysis of more than 100 organic compounds by measuring the vapor pressure rise induced by thermal decompositions and release of volatile products [Blake et al., 1961]. The results are presented in a parallel analysis between the collected thermal data, such as the onset temperatures of the decomposition reactions, and the structural features of the compounds or the influence of the different reaction paths on the thermal stability.

Among the influential structural characteristics discussed are the number of substituents, the steric crowding they cause and the aromaticity. They also divided the compounds into several chemical categories, and present for each some possible decomposition pathways to complement the qualitative discussion of their results.

These studies were the first to demonstrate that the molecular structure indeed influences the macroscopic properties and behavior of compounds. The qualitative tendencies they highlighted maybe serve as descriptive models. They note quantitative predictions were attempted but appeared inaccurate. Despite the faulty estimations they obtained from their models, several elements of their approaches are still in use nowadays.

In the meanwhile, quantitative approaches were emerging, and in particular, group additivity methods were developed in the late 1950's with Lydersen work on critical properties [Lydersen, 1955] and Benson and Buss' on heat of formations [Benson and Buss, 1958]. These two methods were ground laying for the development of numerous studies to estimate various thermodynamic properties and they have been transformed and improved gradually to adapt to novel applications.

1.1.2 Basic Principles

The group contribution methods are laid on the principle that each fragment of the molecular structure, be it an atom, a functional group or a larger substructure, participates to the molecular property and that this contribution is particular to the fragment. Then, for any other molecule composed of fragments of known contributions, its property value will be the sum of the groups' contributions. In simple words, it can be considered as the chemists' Lego[1] and when the bricks are assembled, the obtained entity is the sum of its constitutional parts, both on the structural aspects and on the physico-chemical features.

The principle of additivity of the groups' contributions also determines that, by definition, all GCM applications should give rise to linear models where all groups contributions are multiplied by the group's appearance and summed.

To illustrate the general rule of GCM, Table 1.1 reproduces an example from Benson and Buss' study mentioned above [Benson and Buss, 1958]. From their analysis, the authors determined that the incrementation of the structure by one $\mathrm{CH_2}$ group decreases the standard enthalpy of formation by approximately 5 kcal/mol and their fitted values for $\Delta H_f°$ of several saturated alkanes are rather accurate (given in kcal/mol in Table 1.1).

---

[1] Analogy made by Prof. P. Vogel, during *Organic Functions and Reactions I* lectures, EPFL, 2007.

Table 1.1 – Standard Enthalpy Change of Formation Estimates by Group Contribution [kcal/mol] from Benson and Buss [1958]

Molecule	$\Delta H_f°$	$\Delta H_f°(est)$	Error
CH_4	-17.9	-15.3	2.6
C_2H_6	-20	-20	0
C_3H_8	-24.8	-25.2	-0.4
nC_8H_{18}	-49.8	-49.8	0
$nC_{11}H_{24}$	-64.6	-64.6	0

The models proposed by Lydersen for critical temperature pressure and volume [Lydersen, 1955] are not simply additive as they include the group contributions sum into non-linear equations: for instance, the critical pressure P_{crit} presented below is dependent on the molecular weight M, and the inverse of a constant parameter c and the group contributions G_i:

$$P_{crit} = \frac{M}{(c + \sum G_i)^2} \tag{1.1}$$

Despite the slight difference - either the property is related to the sum or to a function of the sum of group contributions, the fundamental idea remains the same.

The major challenge beyond these simple principles is the definition of the groups and the empirical determination of their respective contributions. In their work, Benson and Buss already define their groups at three different levels: the zeroth-order groups are atoms; first-order groups are partially substituted atoms and small molecular frameworks (e.g. CH_2, NH or CO) that don't necessarily correspond to the common functional groups; finally, the second-order groups comprise two substituted atoms or neighboring groups (e.g. $-CH_2CO-$). And yet, they note that this framework is limited as several substructures cannot be represented such as rings or double and triple bonded carbons. Hence, various frameworks were developed in order to address this issue and to extend the application of group contributions to other compounds and to other properties, and the next section will present few examples.

1.1.3 Frameworks and Groups

Following the pioneering methods in group contributions, several frameworks were developed in the consecutive years. As an indicator of the extent of model multiplication, nowadays software packages enabling thermochemical property estimations propose to select among more than 60 group contribution methods [DDBSST, 2009]. The aim here is not to present a comprehensive review of all the possibilities, but rather to understand the criteria that help selecting among them.

The variety of models developed targeted different features and thus relied on different sets of experimental data, that determined in a way the definitions of the groups. Kolská et al. [2012] make an extensive inventory of GCM applications noting that, besides critical properties and enthalpy of formation, models were developed for parameters of state equations, activity coefficients, vapor pressure, gas or liquid viscosity, etc.

Most of these models apply the previously explained principles and are additive as for instance the method proposed by Joback and Reid [1987] which is an extension of Lydersen method both in terms of number of defined groups and predicted properties. Methods derived from UNIQUAC as UNIFAC [Fredenslund et al., 1975] took the method a step further and included group interactions that are neglected by other methods [Hooper et al., 1988, Larsen et al., 1987, Tiegs et al., 1987]. This results in more complex models as more parameters are to be determined on top of the group contributions. At the same time, some methods tended towards models simplification by only considering atoms and molecular weights as parameters [Klincewicz and Reid, 1984],though, at the cost of accuracy.

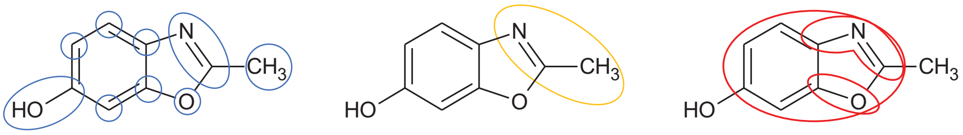
The studied compound sets were determined depending on the property to estimate, hence some of these models are generated from large sets of organic compounds while others are constructed on narrow sets of chemicals or with particular physical features or specific functional groups: highly branched hydrocarbons [Chickos et al., 1995], fluorinated compounds [Brown et al., 2010] or ionic liquids [Lazzús, 2012]. So, some of the latter methods cannot apply to polar compounds, heavily halogenated compounds or molecules of high molecular weights [Joback, 2001].

In the present case, the purpose is not to estimate given properties with an existing model nor to develop a novel framework, but to develop predictive models of thermal stability with a large application range. Therefore the framework required needs to be broadly applicable to various kinds of chemicals, and preferably applicable to thermodynamic properties. This leaves several possibilities, among which the framework intentionally developed by Marrero and Gani [2001] for this purpose.

Considering the available methods at the time [Joback and Reid, 1987, Klincewicz and Reid, 1984, Lydersen, 1955], Marrero and Gani report that several limitations were yet to be solved: most methods are unable to distinguish between isomers, applicability is limited due to over-simplified structural representations, and they considered some to be "of questionable accuracy". Building from previous attempts to solve these issues [Constantinou et al., 1994, Gani and Constantinou, 1996, Marrero-Morejón and Pardillo-Fontdevila, 1999], they developed a new framework that allows for more accurate estimations and that can describe a wider range of compounds, including large and complex chemicals. Their method relies on three levels of molecular description: the first level of simple groups, as for previous methods, that can not distinguish isomers or neighboring effects; the groups of the second level are larger fragments of the molecular structure, assemblies of first order groups, that allow for better consideration of isomers or neighboring groups; the third order groups are able to describe compounds for which other methods usually fail, such as complex heterocyclic or large acyclic but polyfunctional compounds. Later, the framework was enlarged even further with the inclusion of connectivity indices that provide structural information on the "special cases", that would be a molecule that can not be fully represented by any of the groups available [Hukkerikar et al., 2012]. This combined approach is referred to as GC^+ and is the one that will be applied for this project.

To determine the groups and their frequencies within a molecule, they developed a software package, ICAS [Gani, 1999], which includes a specific module, ProPred which allows drawing the 2D molecular structures and automatically generates the corresponding GC^+ groups.

To illustrate the groups identification within the GC^+ framework, a molecule's representation, namely 6-hydroxy-2-methylbenzoxazole is detailed in Figure 1.2.



(a) First Order **(b)** Second Order **(c)** Third Order

Figure 1.2 – Representation Example in Marrero-Gani Framework

Table 1.2 – Marrero-Gani Groups for 6-hydroxy-2-methylbenzoxazole

1^{st} Order Groups		2^{nd} Order Groups	
Times	Group	Times	Group
1	CH_3	1	(N=C)cyc-CH_3
3	aCH		
2	aC		
1	aC-OH		
1	C=N(cyc)		
1	O(cyc)		
3^{rd} Order Groups		Connectivity Indices (CI)	
Times	Group	Value	Index
1	aC-(N=CHn)cyc fused rings	3.9	${}^0\chi$
1	aC-Ocyc fused rings	1.1	${}^1\chi$
1	Aromatic fused rings [2]s2		

1.2 Quantitative Structure-Property Relationship

1.2.1 Background

As group contributions methods quantitatively correlate structures of chemicals and their physico-chemical properties, they are thus often considered as part of QSPR methods. However the distinction made here is to mark the fact that QSPR methods rely on descriptors and not on constitutional fragments of the molecule. Descriptors are defined by Todeschini and Consonni [2012] as follows:

> *The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.*

Descriptors can be classed according to their origin or their dimensions:

They either derive from the

- topological,
- geometrical,
- electronic,
- quantum-chemical,
- or thermodynamic properties of the molecule,

or may be classed as

- zero-dimensional descriptors (constitutional descriptors derived from chemical formula, like the atom count for instance);
- 1D descriptors (constitutional descriptors for fragments and groups);
- 2D descriptors (topological descriptors derived from molecular graph);
- or 3D descriptors (geometrical features or quantum-chemical descriptors that require non-trivial computation) [Dehmer et al., 2012].

From this perspective, GCM are special cases of QSPR that only employ 0D and 1D descriptors.

QSPR methods, or QSAR methods originally for Quantitative Structure-Activity Relationships, have roots in studies even precedent to the Group Contributions discussed above and also started first by focusing on a hazardous behavior: toxicity. Cros has reported in 1863 that toxicity of aliphatic alcohols is a function of their water solubility which is itself a function of molecular structure [Liang et al., 2011].

Toxicity has been one of the main focus of QSPR modeling for decades especially following the development by Hansch [Hansch et al., 1968] of an equation to quantitatively relate biological activity -i.e. the minimal effective concentration at which a biological activity is observed- and as structurally determined feature - i.e. a relative hydrophobicity measure through partition coefficients in octanol/water system- [Albert, 2013]. QSAR found applications both in environmental [Harder et al., 2003, Karcher and Devillers, 1990, McCarty et al., 1985, Nendza and Russom, 1991] and medical studies [Gebauer et al., 2003, Hansch and Dunn, 1972, Palm et al., 1998, Panthananickal et al., 1978].

Until the end of the 1960's, the relationships developed to link molecular properties and molecular structure were essentially Group contributions and based on 0D and 1D descriptors, with simple molecules, homologous series with a common molecular skeleton, or with descriptors developed for specific substituents according to Selassie et al. [Selassie et al., 2003].

'Whole-molecules' approaches emerged in the 1970's, with the application of graph theory to represent chemical compounds which gave rise to the developments and applications of topological descriptors. The mathematics were not only used to develop the relationships but also the descriptors themselves. Among several studies, Randiç work on branching degree and the development of what is later referred to as Randiç index, is based on a molecular representation in edges and vertices instead of atoms and bonds [Randić, 1997]. Explicit hydrogens are removed and only the $\mathrm{C-C}$ connections remain at the center of the representation. Randiç defines the branching index as the sum for all edges degrees as shown in Equation 1.2 :

$$\chi = \sum_{edges} \frac{1}{(v_i \cdot v_j)^{\frac{1}{2}}} \tag{1.2}$$

where v_i and v_j are the valences of the vertices i and j at each end of each edge. The valency is also meant in the context of graph theory i.e the number of connections. Randiç successfully correlated this branching index to boiling points, formation enthalpies and Antoine's constants for a set of C_2 to C_7 alkane isomers. At that time, several topological indices were developed and were increasingly involved in QSAR modeling of various properties as reported by Rohrbaugh and Jurs and later Selassie et al. [Rohrbaugh and Jurs, 1987, Selassie et al., 2003].

By the mid 1980's, the geometrical structure of compounds was also engulfed into QSAR techniques with the emergence of 3D descriptors that encode the surface, volume or charge distribution. The geometrical, electronic and quantum mechanical descriptors are derived from molecular orbital functions and therefore their broad application emerged when the quantitative calculations of molecular orbital became possible. The resolution of Schrödinger equation for many-electron systems was enabled by the Born-Oppenheimer approximation first, then by the Hartree-Fock method - or self-consistent field method (SCF)- already in the late 1920's; however, these methods were iterative and required enduring computation. Cances et al. note that it was only when computational power could enable reasonable computing time that molecular structures could be effectively determined. Besides the increase of computational power, geometries and electronic structure calculations were significantly simplified thanks to other approximations methods that allow to by-pass *ab initio* resolution of Schrödinger equations through Hartree-Fock method [Dewar and Thiel, 1977, Pople and Segal, 1965, Stewart, 1989]. Despite their acknowledged inaccuracies, they reduced computational time and resources to determine the molecular electronic structures and gave rise to several numerical descriptors that served for a better understanding of 3D geometries and could be included, among other applications, into QSPR studies [Karelson et al., 1996].

The QSPR models evolved with the increasing number of available descriptors, followed trends and faced skepticism. Nonetheless, successful applications to a variety of fields, lead to progressive establishment of QSPR methods. A wide spectrum of possibilities exists, in terms of descriptors, models, studied properties, that contribute to their popularity and motivate their use to novel investigations.

1.2.2 Principles

The basic principles of QSPR methods are rather simple, flexible and permissive. The models are the equations that express the physico-chemical properties as function of the structure. The aim being to correlate the property space and the descriptors space, all mathematical techniques to identify and develop correlations are allowed: from mono-parametric linear discrimination that gives rise to a 2-class classification based on 1 criteria, to artificial neural networks resulting in complex non-linear weighted sums of multi-parametric nodal functions models (i.e. a function of the function of a sum of functions)[Dehmer et al., 2012].

The typical procedure to follow is similar to what was previously described in Figure 1.1. The few rules that apply to the model construction phase are presented here but further details are given in chapter 2 as each step of the method can be performed through different possibilities.

Dataset creation: the dataset consists in the property space and the chemical (or structure) space, i.e. the collection of the physico-chemical property to be studied and the descriptors for all observed molecules. The dataset size and quality have a high impact on the quality of models. Ideally, consistency in the experimental protocols and data collection should be ensured and the dataset size should be large enough to contain information representative of the potentially relationship existing between property and structure.

However, dataset standardization is sometimes traded off for size, as to ensure the availability of a sufficient number of observations, data collection from several sources is often necessary and decreases the strict standardization of experimental data. For a dataset to be considered of sufficient statistical relevance, it is preferable it comprises at least 20 observations and up to several hundreds. Nonetheless, this is not possible with all applications, especially those for which data availability might be scarce. It also implies drawbacks as statistical correlation methods tend to give rise to complex models when dealing with large data sets, that might end up to be uninterpretable. Moreover, the excess of information is also unfruitful if it over expresses certain features to the detriment of other. For instance, if a specific substructure appears once within a set, its effect on the studied property will be better examined within a narrow set, rather than a large set where this information could be lost within an high amount of information. Hence, an ideal dataset should gather data representative of structural and property spaces with well-balanced, continuous distributions. If not, particular attention should be paid to treat gaps and less-represented classes within the property space.

Descriptors: As seen in the previous section, development of new models and new descriptors have followed a fruitful cycle and nowadays, the number of available descriptors exceeds thousands. The generation of the descriptors for the dataset creation is usually performed with specialized software packages (several suggestions are found in [Milano Chemometrics & QSAR Research Group, 2007]) in order to enable fast and accurate calculations for a large number (≈ 1000) of descriptors and to process several molecules.

Training: The construction of a model is referred to as training. Indeed, most of the generally used methods are iterative and necessitate several successive steps in order to improve the model from the initial to the final step to determine the relationships, hence the terminology. The search for the most appropriate descriptors among the available pool can be rather challenging and selection should be drastic: for instance to select 5 descriptors among a base of 100 leads to more than 75 million possible combinations. This selection is the aim of the training stage, nonetheless, it is advisable to control training and to impose a ratio of 1:5 between descriptors and observations to generate robust models.

The quality of the model will depend on the quality of the property and descriptors data , however the choice of the modeling method, the descriptor selection and the model evaluation are critical to the development of robust models. Algorithms are set to vary descriptors parametrizations to minimize the error between the responses and the targeted property values. For this, they tend to maximize information inclusion by increasing the parametrization while it would be preferable to keep it to a lowest. Thousands of descriptors are available, many are redundant or inter-correlated, while models should only rely on a few independent descriptors; this results in a vast number of possibly equivalent combinations of descriptors as models. Therefore, the training stage can result in various outcomes depending on the sequence of decisions taken here. These issues will be addressed more thoroughly in Chapter 2.

Validation: Validation requires dividing the available observations prior to training, to train the model with a subset, the training set, and then apply the model to the remain-

> ing subset, the validation set, to ensure its ability not only to fit but also to predict observations. The performance, validation and reliability of the models depend on the training-validation proportions, the validation methods and statistical indicators employed.

Overall, at each stage of the process several alternatives are possible, probably too many alternatives, so that decisions have important impact on the final outcome. Nonetheless, this high degree of freedom is both the benefit but also the drawback of QSPR methods: it opens large horizons and enables to find underlying correlations, but the results may appear impossible to interpret if they are too complex or cannot be repeated when the decision sequence is altered.

Indeed, beyond validation, models are legitimized by physical interpretations, and model complexity hinders interpretation as influence of the different parameters is silenced by the excess of information. The central idea is to maintain balances: balance in the dataset size, balance in chemical and property spaces representations, balance in descriptors selection to maintain sufficient but reasonable information, balance in training and validation sets, balance in model evaluation, etc. This is referred to as the "Tao of QSPR building" by Dehmer et al. [2012].

Several statistical machine learning methods (i.e. statistical modeling algorithms) are exposed in Chapter 2. They allow developing classification or regression models. The method of choice is mainly determined by the objectives of the project, which also holds when GCM are employed.

1.2.3 Descriptors

Several types of numerical descriptors exist and can be calculated from molecular structure, and while 0D, 1D or 2D descriptors can be determined without requiring advanced computation, it is not the case of 3D descriptors. Therefore, software packages are made available to draw 2D structures, optimize the 3D geometry and evaluate the descriptors. Some focus on either one of these tasks, while others integrate all of these functions, and propose supplementary options as data preparation, visualization, and even model building and evaluation. Descriptors calculations should not vary with the choice of software employed, however, descriptors included or methods can be slightly different and imply numerical differences, yet tendencies are consistent. For instance, electronic partial charges can either be calculated from Gasteiger-Marsili, Zefirov or Mulliken methods. Not all software packages propose to calculate with these 3 methods, nor do the methods present the same results.

0D and 1D descriptors are mainly constitutional descriptors, e.g. total atom count, elements atom counts, molecular weight, etc. These descriptors are straightforwardly determined from the condensed molecular formula, and the use of software for this purpose is only for convenience and rapid treatment of large sets of compounds.

2D descriptors derive from the developed molecular formula. For this, software packages offer to draw the structure if known, to convert it from the name or browse online databases. The 2D descriptors types are the topological descriptors that reflect on the connectivity and branching of the molecule and the information-content type which considers integration of molecules subsets and how much "information" (atoms) they contain relatively to the entire structure.

Finally, to obtain 3D descriptors, a pre-processing step of 3D geometry generation and optimization from the 2D structure is necessary in order to describe the molecular conformations and derive descriptors. 3D descriptors such as the molecular volume, the surface area, or moments of inertia along three dimensions can be generated but also serve for the computing of more complex features as the electronic structure or charges distribution. As QSPR were first developed for applications in biology, and enzymatic activities in particular, descriptors related to local surface properties of molecules are highly important and generated in great number by the packages as they are crucial for the understanding of enzymes folding and active sites characteristics.

Quantum chemical descriptors can also be used since their calculations have also been significantly simplified and most QSPR software packages allow to estimate them. Even though the *ab initio* methods result in better and more accurate calculations, they are computationally expensive which motivated the development of several approximate but simpler methods referred to as "semi-empirical" methods. Semi-empirical methods are also based on Hartree-Fock approximation but neglect several parameters from the exact calculations and replace them with empirically determined parameters. Karelson et al. [1996] cite several examples:

- Extended Hückel Theory (EHT): neglects electronic and nuclear repulsion, gives good qualitative description but results in unrealistic charge distributions [Grüber and Buß, 1989]
- Complete Neglect of Differential Overlap (CNDO): neglects of both diatomic and single-atom atomic orbital overlap [Pople and Segal, 1965, 1966, Pople et al., 1967].
- Modified Neglect of Di-Atomic Overlap (MNDO)[Dewar and Thiel, 1977], Austin Model 1 (AM1) [Dewar et al., 1985], and Parametric Model 3 (PM3) [Stewart, 1989]: neglect diatomic differential overlap only but still take into account two-electron repulsion integrals when electrons are on the same atom.

Another alternative to *ab initio* methods that also allows for saving in computational expense when solving for electronic structures is the Density Functional theory (DFT) [Kohn and Sham, 1965].

The local, relative, or transformed descriptors give rise to a high enlargement of the descriptor space. They represent the same features for a sub-part of the molecule, a particular atom or are the logarithm of a given descriptor. They are necessarily highly correlated to the original descriptors from which they derive, but their physical meaning is not as explicit as they mathematically combine several factors or focus on substructures.

The same example, 6-hydroxy-2-methylbenzoxazole, treated earlier in the Marrero-Gani framework, is processed by a descriptors generating software package, Codessa Pro [Petrukhin et al., 2001]. Codessa Pro package includes MOPAC software which enables 3D geometries optimization from 2D molecular structures, then derives over a thousand numerical descriptors. A first selection eliminates the descriptors with the lowest variances or with missing values resulting in a working set of approximately 350 descriptors per structure. Table 1.3 shows few examples, selected to represent the various types and dimensions.

In summary, descriptors encode for the molecular structure from various aspects and each aspect is treated extensively by numerous descriptors. Unlike GCM, descriptors categories are not exclusive and, unless it is a committed stance, there are no objections to develop models including different types of descriptors. Moreover, whereas the GCM lies on the additivity principle, the QSPR methodology does not bear similar constraints and therefore, the models are not restricted to linear regressions, or sums of descriptors contributions.

All together they offer a broad space for QSPR to evolve and a tremendous number of possible combinations. It is therefore important to select them with care and rigor when building the models, otherwise they are able to fit anything and give meaningless results.

Table 1.3 – Examples of Structural Descriptors for 6-hydroxy-2-methylbenzoxazole

Dimension	Type	Name	Value
0D , 1D	Constitutional	Total number of atoms	18
		Relative number of C atoms	0.444
2D	Topological	Wiener index	143
		Randiç index (order 0)	7.84
3D	Geometrical	Moments of inertia A	0.109
		Molecular volume	132
		TMSA Total molecular surface area	353
	Partially Charged SA	PPSA1 Partial positive surface area	263
	Molecular Orbitals	HOMO energy	-8.91
	Quantum	Total dipole of the molecule	0.431
	Thermodynamic	Total entropy (300K)	94

1.3 Conclusion

After reviewing the historical paths that lead to the development of QSPR and GCM, the basic principles were briefly described. Both methods rely on a relatively simple common procedure and it was noticed that GCM are actually a special case of QSPR. With GCM, two elements of the procedure are pre-determined: the structure representation is to be performed within a particular framework of groups and the models usually involve the sum of the groups contribution. On the other hand, QSPR refer to the general procedure in which all elements are yet to be fixed, and thus offer high flexibility. Vast categories of numerical descriptors are available, and no restrictions are made regarding the choice of possible models.

GCM give rise to explicit models: structural fragments contribute to the increase or decrease of certain macroscopic features; whereas QSPR might imply more complex descriptors within an intricate mathematical relation. This could result in models for which a straightforward explanation cannot be drawn or complex models for which interpretation is hindered. Hence, the GCM can be considered to have more accessible interpretation. Nonetheless, for complex behaviors, additivity might not be appropriate and thus, the flexibility in modeling alternatives proposed by QSPR could be profitable. It is important though, to note that GCM are not immutably linear regressions, and nowadays, some studies tend to take GCM off beaten tracks and propose non-linear models based on group contribution [Albahri, 2014].

Regarding the structural representations, the geometrical descriptors, especially those related to the electronic or quantum properties of the molecules could be far more powerful than constitutional descriptors to describe complex reactivity mechanisms and enclose information that GCM can only skim over. Moreover, despite all efforts towards comprehensiveness of GCM, frameworks still encounter limitations to define all molecular structures and if a compound cannot be fully represented it cannot be studied, while numerical descriptors can describe most structures considering the various aspects covered. The chosen framework for GCM application, GC^+ by Marrero-Gani, was developed to overcome this challenge, and this issue should not be faced.

Finally, another element of comparison that has not been discussed so far is the necessity for appropriate software tools for the generation of the structural representations adequate to each method. With the appropriate software tools, structure space can be rapidly and accurately generated. When the software tools are unavailable, however, only the groups, if the entire framework is known, can be simply drawn from the molecular structure while the descriptors that require computation will not be available. Even computation of topological descriptors can be very fastidious, let alone the electronic distribution related descriptors. In the context of model construction, the software availability could determine the method choice and will be guaranteed in most cases. However, if one is to apply an existing model, to a compound of unknown properties, one has to determine the representation of the considered compound within the context in which the models have been developed. In this case, GCM would be preferable to QSPR: visually, clearly defined groups can be identified, and their frequency

Table 1.4 – Comparison of Molecular Structure Based Modeling Methods

	GCM	QSPR
Principle	Additivity rule	No specific rule
Structural Features	Groups	Descriptors
Generation	Simple computation	Advanced computation
Can describe all molecules	Within framework	Yes
Models	Linear, sums	All types
	Simple	Simple to complex
Interpretation	Explicit	Not always explicit
Main advantages	Explicit	Comprehensive
	Simple	Diverse application range
	Visual identification of groups	Flexibility
Main drawbacks	Limited to defined framework	Descriptors software necessary
	Additivity rule	Model Interpretation

assessed. Then, the calculation for the desired property of the considered compound is recovered with a simple addition. If the model is developed in the QSPR context, the ease of applicability will depend on the included descriptors, or the availability of means to compute the complex descriptors, if needed.

Both cases have their own strengths and limits, and considering the discussions above, they seem equally appropriate for our purpose to develop predictive models for the thermal stability and explosive properties of chemicals. The next chapter will present some supporting mathematical techniques that help toward data analysis, allow to develop different versions of models and evaluate robustness.

2 Data Mining and Machine Learning: Approaches and Good Practices

Data Mining and Machine Learning are fields of computational science that surround algorithms development and application to statistically analyze (large) databases. Data Mining focuses on the exploratory part of identifying potential patterns or tendencies within a given dataset. Machine Learning methods are applied to "learn" from fed information, to observe and characterize correlations within or between several spaces and extrapolate these observations to data outside the input sets and predict their features. Depending on the field of application, data mining is usually pushed to the furthest, which often implies various machine learning methods and hence, the distinction between data mining and machine learning becomes thinner, even unnecessary.

Different approaches exist and correspond to different types of problems: ranking, feature selection, clustering, classification or regression. In this chapter, a review of the current techniques of data analysis is exposed. A selection of these tools will serve for the procedure which will be applied for model construction. Thus, it is necessary to investigate their operating modes and their possible outcomes.

Moreover, there are given manipulations that can generalize over the different techniques to ensure optimal application of the learning algorithms, that will be discussed as well, in order to ensure our procedure comes in adequacy with "best practices".

2.1 Definitions

The field of Machine Learning uses specific terminology for concepts [Mohri et al., 2012] and to avoid any confusion some elements are explicitly defined here.

Observations: the observations, or examples, or samples are the elements of the dataset whose characteristics are studied.
In this case: each studied molecule is an observation.

Features: Or attributes, are the describing characteristics of observations. They represent the

major input required by any learning algorithm. To each observation correspond an ensemble of attributes, usually collected as vectors to be fed to the algorithm.
In this case: the attributes of the molecules are the molecular structures representations. When considering Group Contribution methods, the attributes of the molecule are the frequencies of occurrence of the constitutional groups, and when considering QSPR, the attributes are the numerical values of the descriptors.

Labels: Property value of the observations. They can be either numerical, categorical or nominal. Labels should also be fed to the learning algorithms as input, but are not a requirement in all cases (see supervised vs unsupervised learning).
In this case: Labels correspond to studied properties, thermal behavior and sensitivity to ignition. Besides the real values which correspond to experimental measurements, ranking categories (of the type "High - Medium - Low") are also employed.

Unsupervised Learning: Methods that analyze unlabeled features to highlight hidden patterns. Features are the only input, and outputs are usually classes, discrimination factors or ranking of features influence. As labels are either not fed or nonexistent, outcomes cannot be quantitatively evaluated in the sense that there are no right answers. An example of unsupervised learning is k-means clustering that will be developed later.

Supervised Learning: Methods that analyze labeled features. As features and corresponding labels are fed to the algorithms, the outcomes are usually the relationships between the features to recover the labels. As labels are available, the responses can be compared to the real labels to evaluate the models. Regression problems are necessarily supervised as the label values are required to train the models.

Training and validation: These have already been exposed in the previous chapter. The data are split into 2 sets, the training and the validation sets. The first is fed to the algorithm and serves for the learning process and the model construction, then the model is applied to the latter to evaluate its performance.

2.2 Feature Selection and Dimension Reduction

For any type of modeling problem, the data set dimensions are a critical parameter. The number of observations should be much higher than the number of features involved in the model. As the learning algorithm usually browses the feature space to analyze hidden or latent structures, the search depends on the number of combinations to test and thus on the number of features. On the other hand, with complex systems and powerful tools to study them, the feature space can be rather large while the available observations number may be limited.

Therefore, in most cases, a prior dimensionality reduction is often necessary. Obviously, the dimension to be reduced is the number of features, whereas the number of observations should be maximized. To limit the dimension of the feature space, several feature selection methods are possible. This allows to hold only the most interesting features and discard the others, thus reducing the possible combinations, easing the learning process, and producing better results more efficiently.

Feature selection is normally performed in an unsupervised manner, so that the feature space is considered independently from the labels. Yet, ultimately, the selected features will be put in relation with the labels. Therefore some supervised methods can be employed as well for selection, as they take into account the possible correlations of features and labels. Some of the most applied methods are presented here.

2.2.1 Filter

Filtering is an evaluation of the feature space alone [Witten et al., 2011]. It is an analysis of all attributes based on a criteria and a threshold value, and all features that do not meet the threshold can be discarded. Features can be treated separately (univariate) or in groups (multivariate). For instance, the inter-correlations of features can be evaluated with their covariance as shown in Figure 2.1.

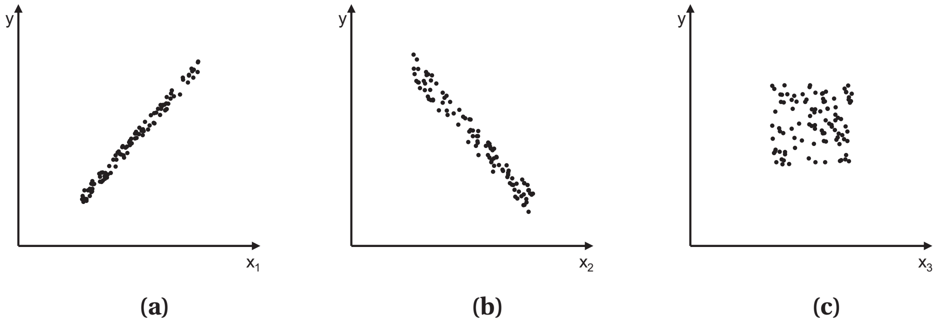
If several features exhibit high correlations to each other as x_1, x_2 and y in Figure 2.1 (a) and (b), they can be considered redundant, and only one of them is necessarily kept, for instance y, while the others are removed. This is a multi-variate and unsupervised manner to simply and rapidly eliminate features, however the correlation threshold is to be fixed wisely for this elimination not to be too drastic which could cause a significant data loss.

Supervised filtering is also possible: instead of evaluating the features inter-correlations, their correlations to the labels are evaluated individually (univariate) and those that are weakly correlating the label are discarded. For instance, if on Figure 2.1 (c) x_3 is a feature and y the label, then x_3 could be eliminated as it poorly correlates to the labels.

Inter-correlated features are often encountered with chemical structure describing features as several information are inter-dependent, such as the number of atoms and the molecular weight, or the occurrence of a given structural group and its composing atoms, let alone surface or volume normalized parameters that are by definitions the combinations of several other features. The choice of which of them to keep or discard is then quite challenging, and modeling is here needed.

2.2.2 Wrapper

Wrapper methods select the features through the construction of a model [Witten et al., 2011]. The set of features that give rise to the best performing models are thus considered the most relevant ones and withheld, whereas the others are discarded. This technique is widely applied, especially when the ultimate regressions are developed with non-linear methods, a prior linear regression can be performed for feature selection. Nevertheless, it can also be performed with the learner that is used for the model construction. The learner is run twice: once to select the



(a) (b) (c)

Figure 2.1 – Examples of Feature Selection by Filtering

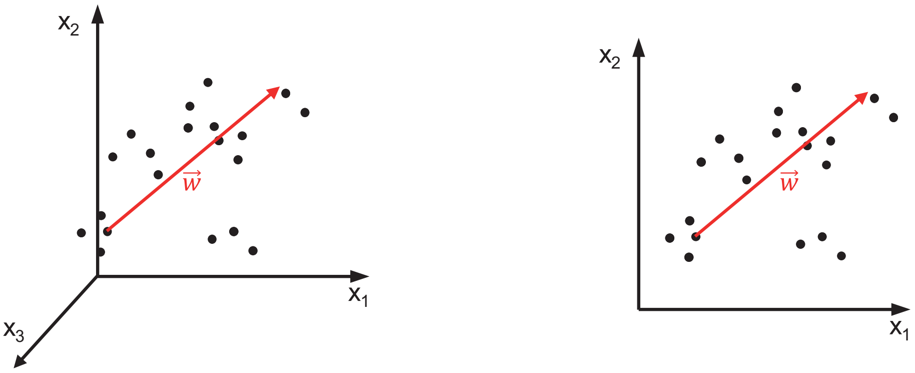
features and once to adjust the coefficients for the reduced space. An example of this method could be seen as in Figure 2.1 (a) and (b), if we let now y as the labels. Then x_1 and x_2 both correlate well to y they can be selected for modeling; however when the model is adjusted, perhaps only one feature is necessary either x_1 or x_2.

2.2.3 Principal Component Analysis

As the feature elimination induces data loss, other methods provide dimension reduction while preserving as much information as possible. Principal Component Analysis (PCA) consists in projecting the feature space onto a new space of lower dimension [Witten et al., 2011]. PCA algorithm normalizes the feature space, analyzes covariances, and computes eigenvectors and eigenvalues. The projection space is defined by the eigenvectors of the original space. Once again, a threshold is to be set as there are as many eigenvectors than original features, therefore only those with the highest eigenvalues are retained. For instance, the 3D space in Figure 2.2 (a) would be projected through PCA on the 1D vector $\overrightarrow{w}$ in red.

PCA is an efficient dimension reduction system, nonetheless, the projection space being defined by vectors that are linear combinations of the original space has a major drawback. In the transformed space, the dimensions do not bear the physical meaning of the original attributes. If the model is to be explained with the original features, the back-transformation would recover the original space, with risks of over-complex models. If the back-transformation is not performed, the models could simply have no possible interpretation as the eigenvectors do not have a meaning in the sense of the original information.

In order to benefit from PCA, without working in a space of meaningless vectors, a hybrid method may be performed. After evaluating and selecting the eigenvectors with the highest eigenvalues, the inter-correlation of the original space to the new space is computed. This highlights which features of the original space contribute to the principal components, and these features are selected for further investigation. In the example shown on Figure 2.2, the new space $\overrightarrow{w}$ is highly correlated to the original space dimension x_1 and x_2, hence the final result would be the 2D space presented in Figure 2.2 (b) as a compromise between the original



(a) Initial Space **(b)** Transformed Space

Figure 2.2 – Example of Principal Component Analysis

3D space and the 1D space lacking physical meaning.

Several other feature selection methods exist, however they are not covered nor applied here. The above-mentioned techniques can be applied in sequence to efficiently reduce the feature space dimension. It is important though to retain sufficient information for the learner to build the most adequate models.

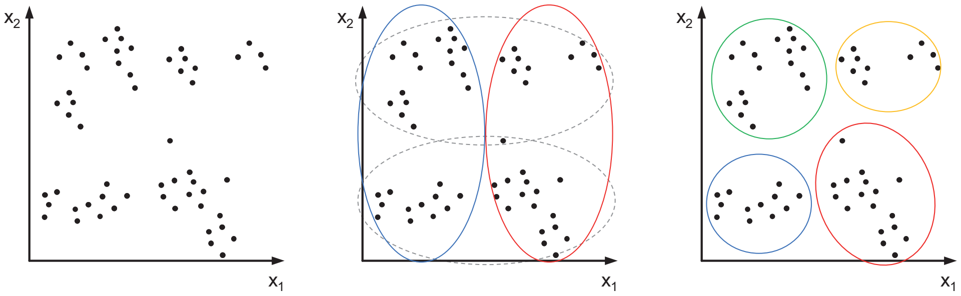
2.3 Classification

Classification problems are the cases in which the labels and the expected responses are categorical or nominal. For instance, a model to determine the structural features that make an observation exhibit "high", "medium" or "low" sensitivity to ignition is developed in Chapter 5.

2.3.1 Cluster Analysis

Clustering is not actually a classification method, it is the unsupervised version of the same type of problems, i.e. grouping observations into categories. Clustering is used on a set of features of unknown class repartition in order to determine if there are identifiable subgroups within the dataset. Hence, clustering plays an important role for the development of systematic grouping of unlabeled data. Among the various cluster analysis methods are the hierarchical clustering, K-means, Self-organizing Maps (SOM) or Gaussian Mixture Models (GMM).

The common goal to all these methods can be schematized as in Figure 2.3. For a set of observations, the learner divides the dataset according to their features and proceeds iteratively in order to identify the optimal groups.



(a) Example (b) Intermediate level (c) Final clustering

Figure 2.3 – Schematic Representation of Clustering

Hierarchical Clustering

Hierarchical clustering can be performed in two manners: either top-down - the whole dataset is considered as one group (top group) and divided progressively in 2 groups along an attribute dimension until each observation is in a separate cluster (bottom of the hierarchy), or bottom-up - starting from individual observations, the learner progressively merges them into groups based on their similar attributes until all data are gathered into the top group [Stramaglia et al., 2004].

The hierarchical clustering offers several advantages, therefore it is one of the most widely used cluster analysis. First of all, it does not require any *a priori* knowledge or guess on the real number of clusters. Second of all, it will point out at the most influential features, thus if feature selection was not performed until this point, it can be deduced from the hierarchical tree. The tree, or dendogram, is the output of hierarchical clustering algorithms and is the last but not least advantage of this method. It visually represents all clustering possibilities from the top group (including all observations) to the bottom groups (individual observation per cluster). The user can then visually analyze the clustering and decide at which level to stop the clustering, depending on the considered criteria (number of clusters, separation of the groups or data distribution).

Figure 2.4 shows an example of a dendogram corresponding to the clustering shown in Figure 2.3. The clustering procedure is completed, however, user can decide to retain the construction of an intermediate number of clusters if it is more adequate. The levels shown here with the construction of 2 or 4 clusters correspond to the cases presented in Figure 2.3. Each branching of the tree corresponds to a dataset division in two sub-groups depending on a threshold value and leafs or nodes correspond to clusters containing certain observations. If the procedure is taken to completion, the bottom clusters construction obtained would suggest as many clusters as there are observations.

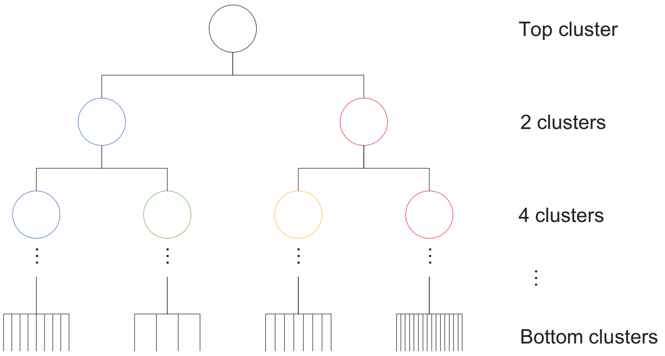


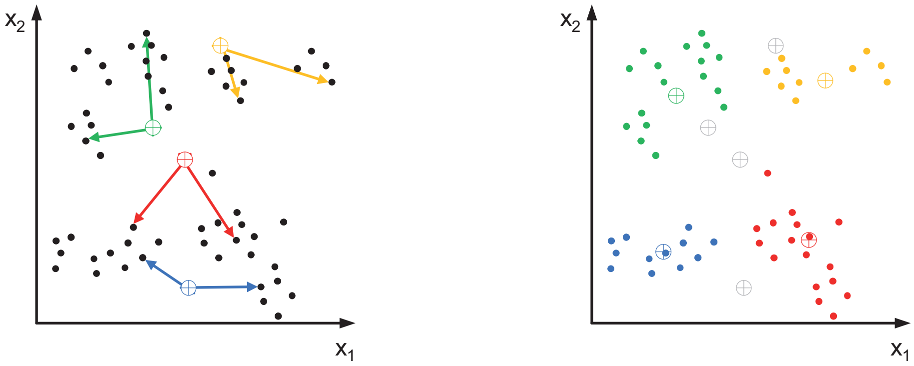
Figure 2.4 – Example of Dendogram obtained from Hierarchical Clustering

K-means Clustering

Another popular clustering method is the K-means technique. K-means creates k groups defined by their mean position in the feature space, called centroid. Initially, the centroids are randomly generated. Then, euclidean distances between each observation and all centroids are computed, and observations are assigned to the group with the closest centroid. As an observation is added to a given group, its centroid position is recalculated and the process continues with the modified centroid position. As the centroids positions are updated, the distances are recomputed and the observations are reassigned, if they become closer to another cluster, and so on until the centroids positions stabilize and all observations are assigned to one of the k clusters [Bishop, 2006].

This technique requires k, the number of clusters, as input. It is thus necessary either to know or have a good guess concerning k. Otherwise, an analysis in order to determine the most adequate number of clusters must be performed.

The most efficient way to determine k, is to vary it, perform the k-mean clustering and compare the results based on silhouette plots. Silhouette plots are the representation of the observed data grouped in the clusters they have been assigned to, and their relative distance to data in other clusters. This reveals how the data are distributed among the clusters. Thus, if a cluster contains few observations it can be considered unnecessary and those observations would be assigned to a different cluster. On the opposite, if a cluster seems larger than the others, it could imply that additional clusters would be beneficial and this cluster would be split further into sub-groups. Silhouette plots also provide information concerning the cluster separation. So, if a cluster contains data that could equally be assigned to another cluster, it reveals that these two clusters are neighboring and could perhaps either be merged, or that a third cluster would be more appropriate to assign data that "overlap" between the two first ones.



(a) Initialization of centroids **(b)** Final clustering

Figure 2.5 – Representation of K-means Clustering

Various Clustering Methods

Other methods are also available and widely used especially in the image recognition field, visualization or for mapping purposes. For instance, Self-Organizing Maps or Kohonen self-organizing maps (SOM) is an artificial neural network based method that is very useful to visualize high-dimensional data into a low-dimension space [Millán and Chavarriaga, 2013]. SOM analyze the distance and the topology of the data in the original d-dimensional feature space and develop a network of nodes in 2D or 3D. The nodes position in the low-dimensional space are not defined with values coordinates but with vectors of weights of the original space, hence the dimension reduction while maintaining the information concerning the neighboring observations. The weights are randomly initialized, and iteratively adjusted to resemble more closely to the sample until a stable map is obtained.

2.3.2 Decision Trees

Decision trees are simple and intuitive supervised classification tools. The principle is to analyze the feature space and determine the most discriminant features and threshold values to split the data into sub-groups that correspond to the labeled classes [Dehmer et al., 2012]. The simplest trees are univariate and binary, that is to say that each node corresponds to one feature criterion and separates the input data into 2 groups. At each node, the decisive criterion must be answered by "yes/no" or "true/false" types of answers. The sequence of queries or decisions is determined by the learner by identifying the feature and the corresponding threshold value that maximizes class separation. Between iterations, the learner evaluates if supplementary splits are required or if a class can be assigned to leaves. For this, the error rate of classification is evaluated. Supplementary splitting rules are added if they improve the tree estimations, if not, the tree construction is terminated and classes are assigned to leaves .

Figure 2.6 presents an example of classification using a decision tree. In the level 0, all data are gathered into one group, and as there are three known classes, the data are to be split

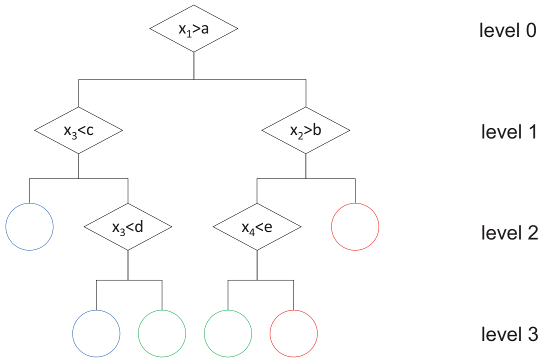


Figure 2.6 – Example of Classification by Decision Tree.
Diamonds: decision nodes; colored circles: classes

successively in order to recover the repartition into classes. The learner selects feature x_1 and threshold a as the first decision node. This criteria is the most discriminant as the blue and red classes are completely separated. Regarding the green class, several criteria are necessary to manage good separation between green and blue on one hand and green and red on the other hand.

Various constraints can be imposed to control the decision tree construction. The usual optimization criterion for a tree construction is the error rate or node purity. The algorithm selects the query in order to maximize node purity. Without a stopping criterion, the splitting is carried on until all terminal leaves are pure. A stopping criterion can be set as a limit on the branching degree, or on the terminal or parent leaves populations. If these criteria are not set, the tree could be over-branched with pure nodes that only contain one or a few observations. This would resemble the bottom level of a hierarchical tree, and would not bring much information regarding the data classification.

It is also possible to prune an over-branched tree by merging leaves that have a common parent node in order to decrease the branching degree. For instance in Figure 2.6, if the population of the green leaves is very small, it could be judicious to prune one or both of the pairs of leaves on level 3 and stop the tree construction at level 2. This decreases the branching and the nodes of the tree, at the cost of increased impurity of the new leaves. In this case, it would completely erase the green class, which would not be appropriate, nonetheless in complex cases it could be highly beneficial if the induced error increase is limited.

Overall, decision trees are rather simple yet efficient tools to class data. It clearly hierarchies the influential features as it selects the most discriminant ones for queries. The visual representation is also helpful to understand the structure of the feature space and to identify neighboring data as they follow similar paths along the branches. Finally, it is easily understandable and applicable which is valuable in exploratory studies and could serve as a good initial step prior to more complex methods.

2.3.3 Linear Classifiers

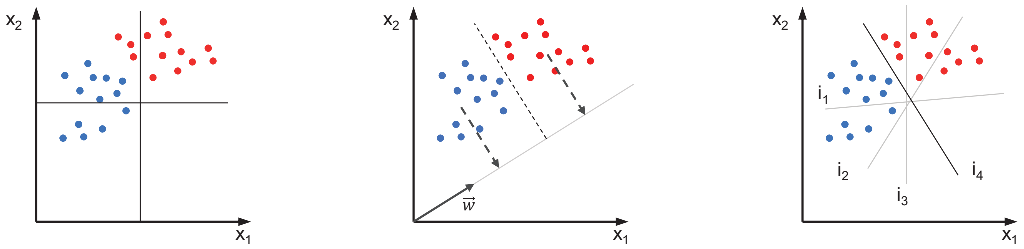
Linear classifiers are combinations of several weighted features to separate the data into their labeled classes. They define hyperplanes in the feature space that represent the boundaries between the different classes.

Bayesian methods analyze data distribution to determine the probability densities of the classes on the feature space. Then, any new observation is assigned a probability of membership into a class rather than an actual assignment. Non-probabilistic methods proceed differently: the data distribution serves to determine the most discriminant dimensions and observations are assigned to classes such as the error rate is minimized.

Figure 2.7 (a) presents a case where the features are insufficient to assign data to their labels if taken separately, whereas a combination of several features can successively differentiate classes. To determine the optimal classifier, several methods were developed.

Fisher Linear Discriminant (LD) for instance proceeds as shown in Figure 2.7 (b) [Bishop, 2006]. Data are projected onto a 1D vector $\vec{w}$ that maximizes inter-class scatter while minimizing the intra-class scatter. Then, a threshold value on $\vec{w}$ is fixed as the limit between classes. Fisher linear discrimination assumes that classes have a normal distribution and equal covariances [Dehmer et al., 2012].

Otherwise, it is also possible to search for the optimal classifier $\vec{w}$ with an iterative loop in which $\vec{w}$ is progressively adjusted in order to minimizes classification errors. An initial random $\vec{w}$ is set, then an evaluation of data dispersion around the hyperplane allows to compute errors and $\vec{w}$ parameters are updated. The iterations are conducted until the error is lower than an acceptable level. This is schematically represented in Figure 2.7 (c).



(a) Example **(b)** Fisher LD **(c)** Iterative algorithm

Figure 2.7 – Examples of Linear Classifiers

2.4 Regression Models

Regression problems are the cases where the targeted properties or labels are real numerical values. In those cases, the aim is not only to distinguish the data from each other as for classification problems, but the model responses should correspond as closely as possible to the real target values.

2.4.1 Multiple Linear Regression

Multiple Linear Regression (MLR) are the most fundamental and probably the most widely applied solution to regression problems. The main initial postulate of MLR is that the property y of observation i is approximated by $\hat{y}$ as a linear combination of the k attributes X of the considered observation [Dehmer et al., 2012]. This is formalized in the equation below where β_0 is the model constant.

$$\hat{y} = \alpha_1 \cdot x_{i1} + \alpha_2 \cdot x_{i2} + \cdots + \alpha_k \cdot x_{ik} + \beta_0 \tag{2.1}$$

$$\hat{y} = \sum_j \alpha_j \cdot x_{ij} + \beta_0 \tag{2.2}$$

Estimators are used to determine the coefficients α. The most commonly applied is the Ordinary Least Squares (OLS) which minimizes the sum of squared residuals to determine the coefficients. Other methods, noted robust regressions, are capable of discarding outliers before estimating errors and adjust the coefficients consequently. Stepwise regression allows to optimize the model choice simultaneously with its construction.

Variations of the MLR exist and apply for particular cases where the labels are vectors instead of values (General linear model), or limited values that are all positive or span over a given range (Generalized linear model).

MLR is a very popular technique that is simple to develop and to interpret. The features appear as the models parameters. Their contribution (positive, negative, large, limited or null) to the property is directly reflected by their respective coefficient in the final equation.

2.4.2 Stepwise Regression

As mentioned in the previous part, stepwise regression is a particular method to build Multiple Linear Regressions that does not rely on OLS. MLR is a very efficient method, nonetheless when the feature space is rather large and the studied behavior complex, the learner tends to over-parametrize the model by including numerous features in order to reduce the errors. Stepwise regression proceeds differently and allows for a feature selection simultaneous to the coefficient estimation. This procedure is schematically represented in Figure 2.8.

1. The stepwise procedure starts initially with a constant value.
2. Then,
    (a) For each parameter not included in the model (i.e. all of them at the first iteration), it tests the null hypothesis that this parameter, if included to the model, would have a null coefficient. It computes the p-values corresponding to these F-tests and analyzes the results. If the probability of the null hypothesis being true is higher than a previously set threshold (noted p-enter), the corresponding parameter is not included. Otherwise, it is included to the model and its coefficient is adjusted;
    (b) For a parameter already in the model, the tested hypothesis is that its coefficient should be zero. If the probability of this hypothesis being true is higher than another fixed threshold (p-remove), the parameter should be removed from the model.
3. The model is updated with the new parameters, and the process is repeated.
4. The algorithm iterates until no parameter should be included nor removed from the model [Wang and Jain, 2003]. This model is then evaluated and selected for further validation.

For stepwise regression to include fewer parameters than ordinary MLR, it should be closely controlled through the inclusion and removal thresholds p-enter and p-remove. Otherwise, the model would include numerous parameters and over-fit the training data.

2.4.3 Neural Networks

Non-linear regressions find a link function between labels and features that is a non-linear combination. As this opens the scope of possibilities, it is preferable to have *a priori* defined or determined the form or the type of the relationship in order to estimate the parameters. However, the application of neural networks can allow to find a non-linear relationship of unknown type between the targeted property y and the feature space X.

An artificial neural network consists in a system of connected nodes, or neurons, organized in different layers. The input of the network are the independent variables X and its output an approximation of the targeted property $\hat{y}$ after several transformations of the input [Rojas, 1996]. As shown in Figure 2.9, the features x_i are fed to the network. Each node is fed with the outcomes of the previous layer (feed-forward). At each node j, the ANN creates two functions:

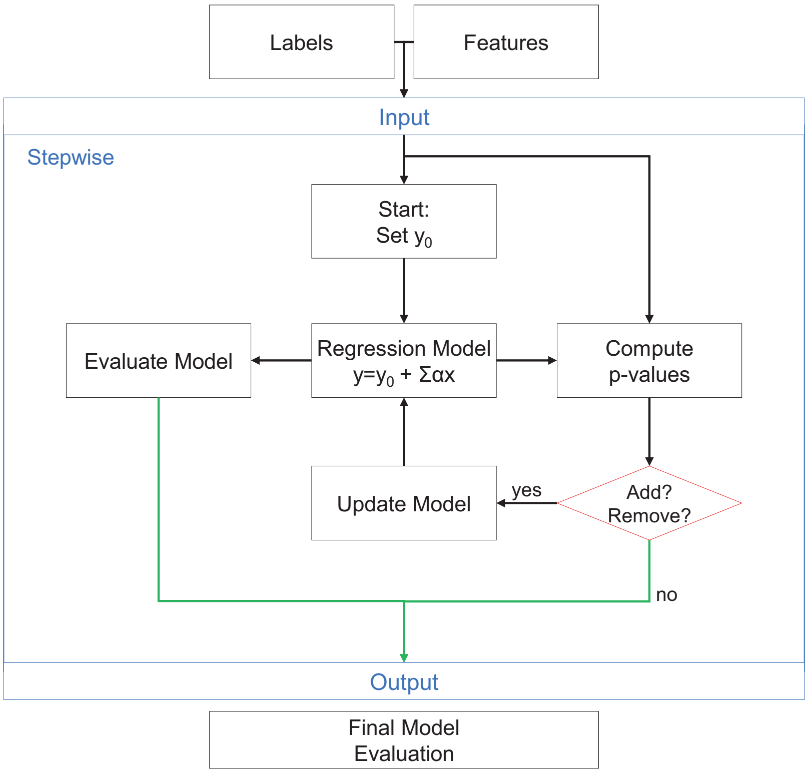


Figure 2.8 – Simplified Stepwise Regression Procedure

an activation function A_j and an output function O_j which depends on A_j and can take various forms, such as the identity function or most often a sigmoidal function.

$$A_j(x,w) = \sum_i x_i \cdot w_{ij} \tag{2.3}$$

$$O_j(x,w) = \frac{1}{1+\exp A_j(x,w)} \tag{2.4}$$

The final output function f is a combination of the output functions from the previous layer, which in turn are combinations of the outputs from the previous hidden layer, and so on. The approximate network response is improved through iterations in which an error function comparing the produced response is compared to the actual target, and fed back to the network in order to readjust the weights (back-propagation). Considering the targeted property, the back-propagation algorithm computes the error function E_j depending on the features, the

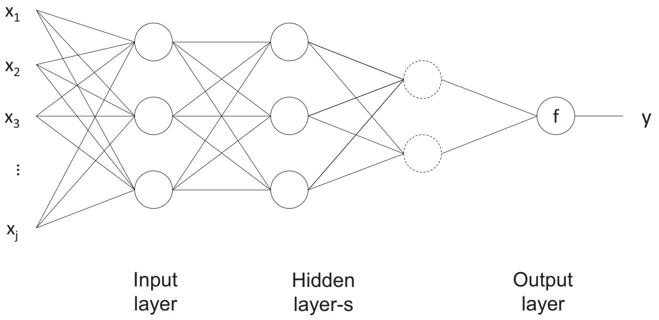


Figure 2.9 – Example of Artificial Neural Network

weights, and the desired output at each neuron j.

$$E_j(x, w, d) = \sum_j (O_j(x, w) - d_j)^2 \tag{2.5}$$

2.5 Good Practices

For all techniques presented above, the challenges faced in model construction are highly similar.

Firstly, large feature spaces carry high information content and though this is favorable in the context of unknown relationships- as one would like to maximize the chances of finding the adequate feature to describe the property- it causes the model computing to be time-consuming. Moreover, it might give rise to several equally probable models if the information is dispersed through the feature space or if several features are collinear.

Secondly, regression models aim at describing the studied dataset to the best by iteratively increasing the model's fit and decreasing the error. This procedure often produces over-fitting models. These models are highly tailored for the studied set and consequently fail to generalize to out-of-the-set data. These mechanisms are noted as "memorization" of the training set and "generalization" to external observations [Mohri et al., 2012]. To avoid over-fitting issues, there are few good practices that can be applied to ensure the model does not only perform well for the studied set but will also generalize better for new observations.

Application of Ockham's razor states that

"entia non sunt multiplicanda praeter necessitatem"
entities must not be multiplied beyond necessity

According to this principle, the simpler the theory to explain a certain phenomena, the better it is. Of course, the simplest theories are preferred if they give comprehensive explanation to the problem. An established translation of this principle in model construction states that among several models that fit the observations with equal or similar accuracies, it is preferable to opt for the simplest model, i.e. with fewer parameters, or with lesser intricate relationship.

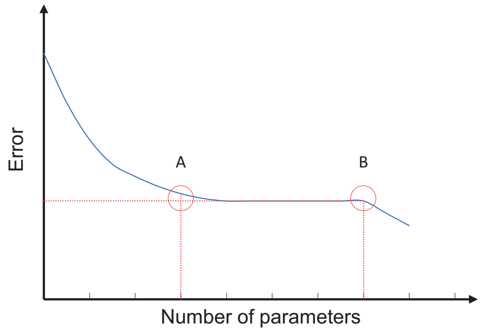


Figure 2.10 – Error Dependance to Number of Parameters Included in a Model. Reproduced from [Bishop, 2006]

To illustrate this discussion, Figure 2.10 shows the typical tendency followed by the model's error when the number of parameters is varied: the error decreases with the inclusion of more parameters to the model as the approximate response of the model is improved. Yet, at some point the inclusion of more parameters does not bring a significant improvement to the model and the error stabilizes. Eventually, with the inclusion of a great number of parameters, the approximate response can be further improved. Here, one can see that between the two potential models, A and B, the number of parameters roughly doubled while the error was only slightly decreased. Following Ockham's razor principle one would favor model A over model B, as they give equal (or highly similar) description of the considered data, and A is simpler than B [Bishop, 2006, Witten et al., 2011].

Feature Selection is helpful for the search of the parameters and decreases the necessary time. Moreover it can also be an efficient preliminary step to ensure that during model construction, the learner does not have to select among equally probable features i.e. features of the same statistical significance to the properties. This allows clear selection of the principal descriptors to include /discard from the model.

To better describe the problem potentially faced here, let's consider 2 datasets, an ideal and an imperfect one. These data sets are simulated only for the illustrative purpose here and are represented in blue and red, respectively ideal and non-ideal, in Figure 2.11. We will consider here that if all the features of the dataset were included in two macro-models, these models would be 100% correlated to the studied property.

The idea in model construction is to recover the maximum of the information contained

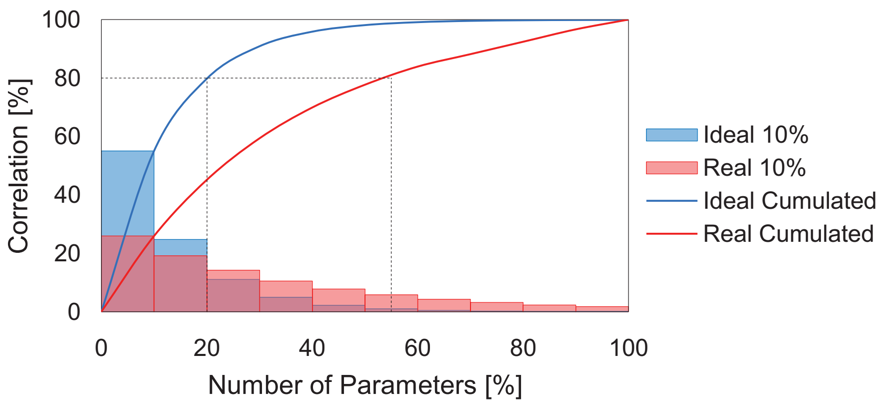


Figure 2.11 – Information Content Relative to Number of Parameters Included a Model

in the full dataset with only part of the features. An ideal set would present an information distribution respecting the Pareto rule [Pareto, 1971]: 20% of all features would hold 80% of the information [Berman, 2013]. The information that can be recovered from the different subsets of feature space is schematically represented by the histograms. For the ideal case in blue, the 10% most informative features contain about 60% of the information to correlate the property, and the next 10% contain about 20%. Together they allow recovering 80% of the whole information content. The accumulated information is represented by the curves.

For the less ideal case, represented in red, more than half the feature space is necessary to collect 80% of total information. Even the least significant features still hold important information without which the correlation does not reach the maximum.

In the ideal case, proper feature selection would be helpful to speed the model construction. The information held by the features is significantly different that learner should be able to appreciate this difference and properly identify the features that highly correlate the property and discard the others. In the other case, the differences in information content between the different features is not as well marked. Therefore, the learner could be limited to properly select the features to include or exclude from the model construction. Therefore, feature selection will serve here to determine among feature of similar influence which to feed to the learner in order to ensure a solution can be found.

Validation is the best way to evaluate the models ability to generalize. Some of the observations should be kept aside from the model training procedure, and after the construction gives one or several satisfactory models, validation serves to verify how they perform on out-of-the-set data. If the validation data serves to select among several possible models, it is considered that they played a role in the model construction. Therefore an additional external validation can be required to evaluate the selected model. This is detailed in Section 2.5.1

Model Evaluation may be performed through several indicators. The error between the model response and targeted property can be quantified through the evaluation of the goodness-of-fit, the Root Means Square Error (RMSE) or relative deviations. Some indicators are presented in Section 2.5.2.

Data Evaluation is necessary as, depending on the data acquisition method, there could be several sources for errors that will lead to uncertainty over the input of the modeling. This should be evaluated and its influence on the models considered. The statistical procedure for this evaluation is presented in Section 2.5.3.

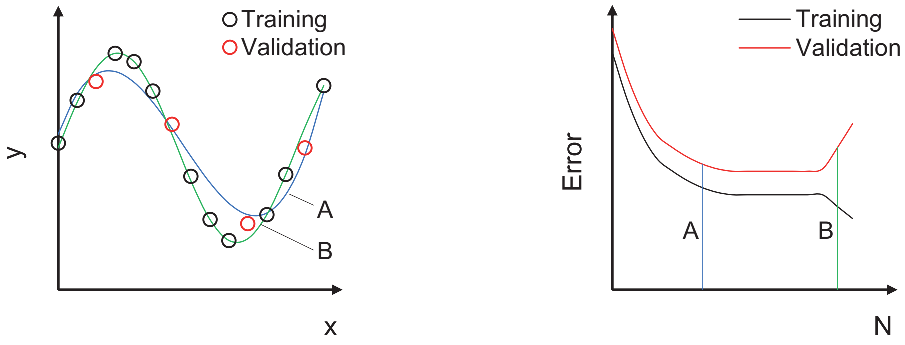
2.5.1 Partitioning Training-Validation

As mentioned above, most Machine Learning methods face the risk of producing over-fitting models due to the strong potential of learning and memorization of the training set, which leads to poorly generalizable models.

To prevent this problem, it is important to control and limit the model construction; the use of a validation set is the most efficient manner.

The validation set is an ensemble of observations removed from the entire dataset prior to model training. Once a model is determined, it is applied to the validation, to evaluate its ability not only to fit but also to predict observations. The training and validation set should be as representative of the overall dataset as possible.

Figure 2.12 presents two models, A and B, similar to the examples taken when discussing Ockham's razor principle. Figure 2.12 (a) shows that between the two models, model B in green fits better to the training data than the model A in blue. Yet, the validation data in red, have a slightly different behavior compared to the training set. Here, model B fails to describe these observations and their properties while the model A performs better. This is reflected in Figure 2.12 (b).



(a) Models **(b)** Error vs Number of parameters

Figure 2.12 – Model Selection

In this context, the validation set is used to compare two models and select the best one. As these data were actively used in the model selection, it is not considered as 'validation' but rather as 'testing'. It is necessary to perform an additional validation, or external validation, of model A with data unseen by the learner at this point.

The performance, validation and reliability of the models depend on the training-validation proportions, the validation methods and statistical indicators employed. It is recommended to split the set in a 80% -20% between the training and validation sets [Witten et al., 2011]. On the other hand, it is also advised that the training set size should not be lower than 20 observations for the model to be statistically meaningful. Therefore, in some cases where the amount of data is insufficient to proceed to this division, it is possible to apply cross-validation rather than simple validation as the withdrawal of 20% of a narrow set would not leave enough data for proper training.

Cross-validation consists in splitting the dataset in k equally populated subsets, named folds, to train the model on $k-1$ parts, and to validate it with the k^{th} set. Then, the validation set is rotated at least k times, so that all the data would have participated in the training and the validation [Witten et al., 2011]. Figure 2.13 schematizes this procedure: the dataset is split in 5 equal parts; at the first run, the 4 first parts serve to train the model while the fifth (blue shaded square) remains for the validation. At the following iterations, the training is performed using 4 different parts and the validation on the remaining set. Performing at least five iterations ensures that all data have been involved in both training and validation. 10-fold cross-validation and 5-fold cross-validation are commonly used, however 5-fold cross-validation is preferable as it respects the 80-20% proportions.

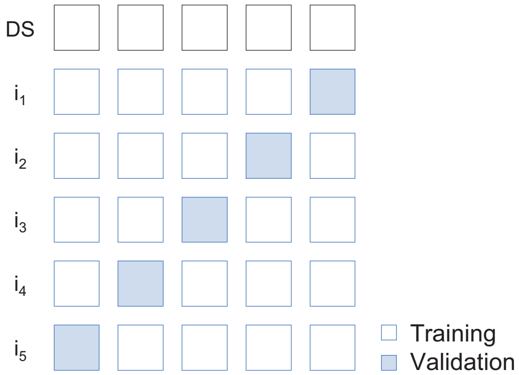


Figure 2.13 – Schematic Representation of Cross-Validation

The results of the k iterations are then averaged to output the final model. The principal issue of this procedure is the sensitivity of the training method to the training set. Indeed, the training set variations - even though 75% of the training set remains unchanged between 2 iterations- can strongly affect the model construction. Therefore, it is recommended to repeat several times the k-fold cross validation to minimize the influence of the training set. This does not guarantee to obtain a model that would efficiently generalize to data outside the studied set but favors it. However, if a k-fold cross-validation is to be repeated m times, it means that the whole model search has to be repeated $k - by - m$ times and could be computationally expensive.

Other validation methods include Leave-one-out validation and bootstrapping [Witten et al., 2011].

Leave-one-out, as the name indicates, leaves one observation from the dataset out for validation and trains the model with the remaining observations. It is a particular case of k-fold cross-validation where $k = n$, the number of observations. The process is iterated until all n observations have been excluded once. Depending on the dataset size, this might require even more iterations than the $k - by - m$ recommended for k-fold cross validations. Moreover, this method will lead to high error for the observations that differ strongly from the rest of the observations. Hence, this represents an efficient way to identify outliers, but the main advantages are the maximal training set size, and that it does not rely on a random data repartition between training and validation sets.

Another manner to face limited amount of data issue is to apply bootstrap validation. This sampling technique holds out a subset of the original dataset for validation, but instead of training the model on a reduced number of observations, it re-samples observations from the training set in order to compensate for the holdout data. This is illustrated in Figure 2.14. The procedure is carried out several times and the results averaged. The main advantage of re-sampling is to consider a training set of the size of the validation set. Besides, the iterations should prevent the effects of over-weighted data (as they are represented more than in the original space). Nonetheless, bootstrapping tends to underestimate real errors.

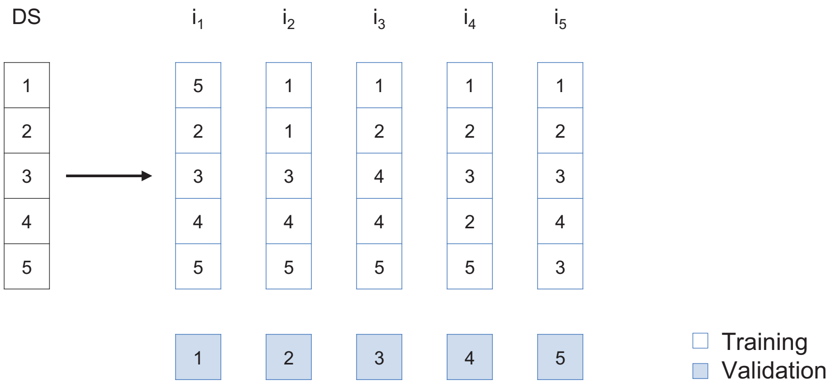


Figure 2.14 – Schematic representation of Bootstrap validation black: original dataset - white: training sets - shaded blue: validation sets

2.5.2 Model Evaluation Criteria

So far, the methods to develop models and to ensure their valid application were reviewed. Nevertheless, the evaluation of the models has not been discussed yet. Independently of the model construction, there are several indicators that can reveal the performance of the models and that will serve to control the training but also to assess their validity. Some of these indicators are presented below [Beal, 2005, Guerard, 2013].

The most basic measure of the model response quality is the evaluation of the residuals, i.e. the absolute errors e_i between the targeted property value y for observation i and its estimation by the model $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i \tag{2.6}$$

For the analysis of the performance of the models over all the observed set, the residuals are not the most appropriate indicator. For this purpose, the errors are, first of all, squared - in order to obtain positive figures that enlarge large errors and alleviate small ones- and summed. This is the Sum of Squared Errors noted SSE:

$$SSE = \sum_i (y_i - \hat{y}_i)^2 \tag{2.7}$$

In order to recover an error evaluation in the same dimension than y, the square root of the normalized SSE by the number of observations n is computed:

$$RMSE = \sqrt{\frac{SSE}{n}} \tag{2.8}$$

The $RMSE$ is a measure of the model response accuracy. However, the $RMSE$ depends on the values taken by the observed property and the model responses, it cannot be easily interpreted as a stand-alone value. Comparing two models and their respective $RMSE$ directly shows which of the two is more accurate to describe the given data. Therefore, it is often included in the Machine Learning algorithms as the control criteria of the modeling procedure: the goal is to minimize the $RMSE$ and at each iteration it is evaluated to assess if the learning process improved the model of the previous iteration, and if not, it would stop the model building process.

Another criteria often found in complement with the $RMSE$ is the determination coefficient

R^2. The determination coefficient serves to evaluate to what extent the model properly fits the observations, hence the term 'goodness-of-fit' often used to refer to R^2.

To compute R^2, the Total Sum of Squares (SST) is needed:

$$SST = SSE + SSR = \sum_i (y_i - \bar{y}_i)^2 \tag{2.9}$$

where SSR is the Sum of Squares of the Regression

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2 \tag{2.10}$$

and $\bar{y}$ is the estimate of the mean of y_i

$$\bar{y} = \frac{1}{n} \cdot \sum_i y_i \tag{2.11}$$

Finally

$$R^2 = 1 - \frac{SSE}{SST} \tag{2.12}$$

It is noteworthy that R^2 is the square of the correlation coefficient. The determination coefficient indicates how correlated two data sets are, i.e the observed data and the model's responses. Unlike the $RSME$, the R^2 is straightforward to interpret: as its values range between 0 for no correlation and 1 for total correlation, it can reveal the performance of the model for itself and not relatively to another model. A combination of $RMSE$ and R^2 is more informative on the model's performance than $RMSE$ alone.

It is also important to note that R^2 increases with, and thus favors, the inclusion of parameters to the model. Therefore, if used alone, it tends to drive over-fitting issues in learning methods that are prone to over-fitting.

To avoid this phenomena, the $RMSE$ and R^2 can be adjusted by taking into account the residual degree of freedom.

$$RMSE_{adj} = \sqrt{\frac{SSE}{n-p}} \tag{2.13}$$

and

$$R^2_{adj} = 1 - R^2 \cdot \frac{n-1}{n-p-1} \tag{2.14}$$

where p is the number of parameters included in the model.

It is also possible to modify the error function with a penalty term that virtually increases the error with the increase of included parameters. The error function could then be the SSE with an additional term corresponding to the sum of squared coefficients [Bishop, 2006].

Moreover, other measurements, such as the Akaike Information Criterion (AIC) [Akaike, 1974] or the Bayesian Information Criterion BIC [Schwarz, 1978], provide indication on the goodness-of-fit while taking into account the number of parameters p in the model:

$$AIC = n \cdot ln\left(\frac{SSE}{n}\right) + 2p \tag{2.15}$$

and

$$BIC = n \cdot ln\left(\frac{SSE}{n}\right) + ln(n) \cdot p \tag{2.16}$$

For small data sets, the AIC may also be corrected and becomes

$$AIC_c = n \cdot ln\left(\frac{SSE}{n}\right) + 2p + \frac{2p(p+1)}{n-p-1} \tag{2.17}$$

The AIC, AIC_c or BIC may help selecting the model among several possibilities that is most likely to be the "true" model. These criteria are relative and the model with the smallest information criterion would be selected.

For summarizing the errors over all observed data in a figure that is more readable than the residuals, it is often helpful to compute the Average Relative Deviation ARD expressed in percentage [%]:

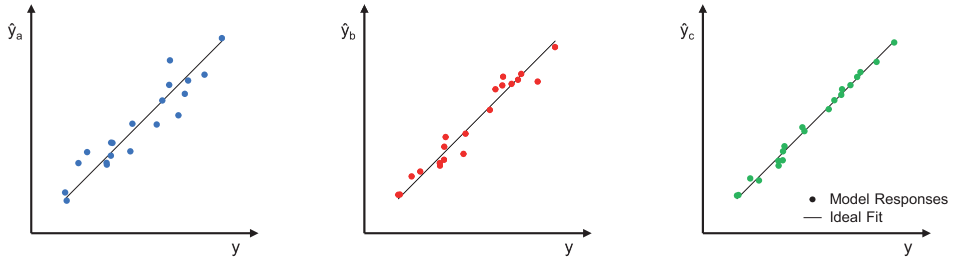
$$ARD = \frac{100}{n} \cdot \sum_i \frac{\| y_i - \hat{y}_i \|}{y_i} \tag{2.18}$$

To illustrate this procedure, randomly simulated data served to create a variable y, and three nested linear regression models A, B and C were built and evaluated. Three models are proposed in order to discuss the evaluation criteria exposed here as many are comparative values and do not inform of the absolute quality of the model.

The nested models are based on similar parameters, and incremented with new ones: model A comprises 3 parameters, model B has 7 and model C has 8. All parameters of model A are in

B and C, and all parameters of model B are in C. The true model used to simulate y is actually build with 10 parameters.

The results are shown in Figure 2.15 which shows the graphical representation of the models' responses $\hat{y}_i$ for models A, B and C vs the targeted values y_i. The evaluations of the three models based on all the criteria presented above are gathered in Table 2.1.



(a) Model A **(b)** Model B **(c)** Model C

Figure 2.15 – Examples of Regression Models

Table 2.1 – Comparative Evaluation of Regression Models

Target Variable y			
n	20		
$\bar{y}$	20.3		
SST	387.4		
Models	A	B	C
p	3	7	8
SSE	44.0	18.8	3.1
SSR	343.4	368.7	384.4
$SSE + SSR$	387.4	387.4	387.4
R^2	0.89	0.95	0.99
R^2_{adj}	0.87	0.92	0.99
$RMSE$	1.48	0.97	0.39
$RMSE_{adj}$	1.66	1.25	0.53
AIC	21.8	12.7	-21.4
AIC_c	23.5	31.4	14.6
BIC	24.8	19.7	-13.5
ARD	6.0	3.8	1.7

The summary in Table 2.1 shows several elements discussed above:

- $RMSE$, AIC, or BIC when taken individually are insufficient to evaluate a model;
- when comparing models, $RMSE$, AIC, and BIC all point at model C as the most appropriate model;

- R^2 and ARD also allow to select model C as the most appropriate model;
- All these criteria favor model C over model B, and model B over model A; yet, when using the "small dataset" correction AIC_c, it appears that model A becomes favorable to model B, as it comprises less parameters. However model C is still the best model as it allows very highly accurate results;
- visual inspection of the graphics in Figure 2.15 confirms this discussion, as the responses of model C are very closely dispersed around the ideal fitting line, whereas model A and B, despite showing good results are not as accurate as model C.

Nonetheless, the ARD of the three models are lower than 10%, and would therefore all be acceptable. The indicators seen above can be used to evaluate the errors between the observed values and the model responses for the training set. This determines whether the model construction leads to a satisfactory result from a descriptive point of view.

Ultimately, the validation set should be considered, and the model's predictive power, or its generalization evaluated. The model that will be most appropriate to apply to out-of-the-sample data will be selected.

In general, the indicators will be marked X_{Tr} or X_{Val} for training and validation respectively.

2.5.3 Statistics for Experimental Data Analysis

So far, focus has been put on the data treatment and the modeling procedure. The data itself has not been discussed yet. However, as it is the starting material of data mining studies, it is appropriate to consider its influence over the outcome.

In this project, two hazardous properties of chemicals, thermal stability and explosive properties, are studied by the two types of modeling methods, QSPR and GCM. The input data is hence comprised of the chemicals' properties and structures.

For GCM applications, groups and their frequency are directly derived from the molecular structures and are not subject to variations that would result in errors. Similarly, the structural descriptors destined for QSPR applications are based on theoretical calculations using a software specially designed for this purpose, and the fluctuations may only arise from the existence of several methods of calculations as mentioned in Sections 1.2.2 and 1.2.3 of the previous chapter. Therefore, the group contributions and the structural descriptors will be considered as true values free from disturbances.

On the other hand, the properties values are mainly experimentally measured values. Hence, the measurement of a "true value" is subject to disturbances that are caused by random and systematic errors. These errors need to be quantified in order to ensure the reliability of the measurements.

Random errors influence the measurement in various ways so that they cannot be explained straightforwardly. They may be the combination of several factors as the noise of an electronic apparatus, or the imprecision of the operator's manipulations, or an accidental fluctuation that has not been taken into account. They can lead to over and underestimations of the true value. Eventually, if the measurements are repeated a sufficient number of times, the random errors should average out, and the mean value correspond to the true value. If it is not the case, it implies that the errors are not purely random and could be due to a phenomena occurring repeatedly through several measurements.

Systematic errors could be due to an imperfect measuring instrument or unavoidable influences of the surroundings. For instance, the influence of the temperature over several types of instruments and experiences is known to be a typical systematic error. Systematic errors cannot be eliminated, however, they can be estimated and the measurements corrected to take these elements into account. This can be done by performing calibration experience.

The total uncertainty corresponds to the added effects of random errors and systematic errors. Even though systematic errors may be corrected, the corrective measures may be imperfect and thus reduce the error without removing it completely.

The value and uncertainty of a measurement are determined through the conduct of a series of observations respecting the conditions of repeatability. That is that a series of experiences are conducted following the same procedure, by the same operator, on the same instrument, under the same conditions at the same location and over a short period of time [Joint Committee for Guides in Metrology (JCGM), 2008]. Under these conditions, the measurements are comparable and hence allow for the statistical treatment to determine the value and the errors.

In the case of a statistical determination of the value y of a measure, where the experience is repeated n times, and the true value of y is estimated by the average value $\bar{y}$ of all observations:

$$\bar{y} = \frac{1}{n} \cdot \sum_i y_i \tag{2.19}$$

The random variations influence the value of y and give rise to a dispersion around $\bar{y}$ that is estimated with the experimental standard deviation σ^2 expressed as:

$$\sigma_y^2 = \frac{1}{n-1} \cdot \sum_i (y_i - \bar{y})^2 \tag{2.20}$$

The variance of the mean gives an estimate of the accuracy of the measurement and can be determined from the experimental standard deviation as:

$$\sigma_{\bar{y}}^2 = \frac{\sigma_y^2}{n} \tag{2.21}$$

The uncertainty over the measurement is the square root of the variance of the mean:

$$\sigma_{\bar{y}} = \sqrt{\sigma_{\bar{y}}^2} \tag{2.22}$$

The final result should be given the form of

$$y = \bar{y} \pm \sigma_{\bar{y}} \tag{2.23}$$

and the value of $\bar{y}$ should be given with as many significant digits as significant relative to $\sigma_{\bar{y}}$.

Moreover, here, y has been treated as property that is directly measured and whose uncertainty depends on its measurements. In the cases where y depends on other variables a, b and c for instance as in the following equation

$$y = f(a, b, c) \tag{2.24}$$

and that a, b and c are measured to determine y, then the uncertainties over a, b and c must be determined and propagated to y.

$$\begin{aligned} a &= \bar{a} \pm \sigma_{\bar{a}} \\ b &= \bar{b} \pm \sigma_{\bar{b}} \\ c &= \bar{c} \pm \sigma_{\bar{c}} \end{aligned} \tag{2.25}$$

The "best" estimate value of y becomes y_B

$$y_B = f(\bar{a}, \bar{b}, \bar{c},) \tag{2.26}$$

and the variance of y_B is estimated as

$$\sigma_{y_B} = \sqrt{\left[\left(\frac{\partial y}{\partial a}\right)_B \cdot \sigma_{\bar{a}}\right]^2 + \left[\left(\frac{\partial y}{\partial b}\right)_B \cdot \sigma_{\bar{b}}\right]^2 + \left[\left(\frac{\partial y}{\partial c}\right)_B \cdot \sigma_{\bar{c}}\right]^2} \tag{2.27}$$

This procedure of statistical determination of the true value and uncertainty of measurement is referred to as "Type A" evaluation of standard uncertainty. However, it is neither mandatory nor always practically feasible. In many cases, for practical reasons, the uncertainty is not determined from series of observations but rather from previous measurements, from calibrations or tests over standard materials, or knowledge and experience [Joint Committee for Guides in Metrology (JCGM), 2008]. This is referred to as "Type B" evaluation of uncertainty.

Besides, when the calibration experiments are performed, it may be that the errors are small relative to the required accuracy of the measurement and are thus negligible. Theoretically, this implies that if not reported, the uncertainty is significantly smaller than the accuracy of the measurement.

Finally, the "true value" of the measured quantity y is "true" under the experimental conditions set for the repeated procedure. In order to test the value under different conditions, reproducibility tests may be performed. When evaluating the reproducibility of a measurement, all conditions may be varied: the principle of measurement, the instrument, the procedure, the standard, the operator, or the location and time. Usually, these parameters can be varied unilaterally in order to evaluate their individual influences; nonetheless in practice, when the principle of measurement is changed, it often requires a different device, or if the experimental conditions are varied, it would be necessary to use a standard that is destined for these particular working conditions.

2.6 Conclusion

This chapter presents a far from exhaustive review of Machine Learning techniques and tools. The most popular techniques and those that will be used in this work, were detailed here. Some definitions were given to avoid any confusion with the not-so-intuitive terminology.

Then, the most efficient manners to reduce dimension and to select among features, filtering, wrapping and Principal Component Analysis (PCA), were covered. The dimensionality problem is very often faced: narrow sets of observations with large number of attributes create a dataset that is extensive to search to find the best model. Moreover, the search can also be hindered by the possible presence of equivalent combinations. Therefore, this feature selection is often a crucial preliminary step to optimize the model construction phase.

Depending on the problem to solve, there are two types of model that can be developed:

classification or regression models. For classification problems, we have seen some rather simple yet efficient procedures such as the decision tree, or k-means clustering. K-means clustering is an unsupervised learning method, and thus it can be applied to determine unknown hidden classes of a dataset. Hierarchical clustering and decision trees require that the classes are known and fed to the algorithm, but on the other hand, they deliver graphical representation of the dataset classification by its features that can be highly valuable.

Regarding regression problems, the simplest solution is the Multiple Linear Regression (MLR). MLR owes its popularity to its relatively simple development and application. Indeed, the linear combination of attributes are accessible and clearly highlight the influence of parameters. Besides, to predict a new value is performed through simple computation of a weighted sum. Artificial Neural Nets can be applied to develop more complex relationships if the correlation between the property and the features requires it.

In the last part, we have discussed some rules and 'good practices' such as the parsimony principle. Several ways to divide the dataset into a training and validation part were discussed. The method of choice will be determined depending on the considered set size. In all cases there should always be a proper validation of the developed models. Then, evaluation indicators of the model quality were defined. They assess the model quality, its goodness-of-fit and predictive power, and we have seen that several measures of the model errors are complementary. Finally, the methods to evaluate the input data reliability were reviewed.

In the light of this discussion, the procedure introduced in Figure 1.1, Chapter 1, is now updated and elaborated in Figure 2.16. Following the creation of the dataset from the "Structures" and "Property" spaces, it is then split into the training and the validation set.

The training set will serve to train the model. For regression models, the stepwise procedure will be applied. When clustering is necessary, k-means is applied, and when classification is necessary, decision or hierarchical trees will be favored. These methods are selected for their relative simplicity, availability and flexibility. Indeed, these elements are available on Matlab software [MathWorks, 2014], which allows parametric manipulations and avoids the "black box routines" offered in other software packages.

The models are then applied to the validation set and evaluated. Cross-validation is employed whenever the dataset size is limited. If available, external validation is also applied to evaluate the models produced.

The evaluation criteria employed are those presented above, however, the reports will only present the "stand alone" terms (for instance RMSE is only shown if comparing various models).

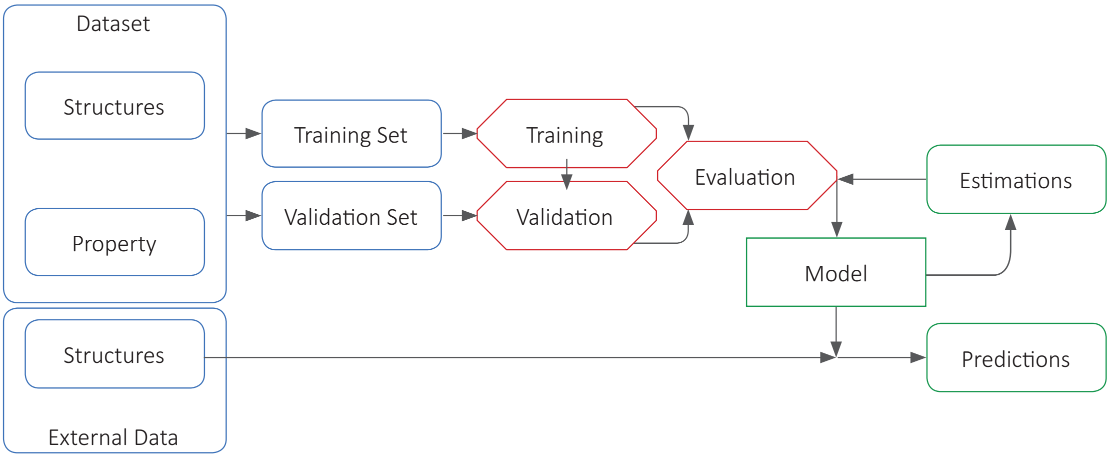


Figure 2.16 – Updated Procedure for Development of Structure-Based Predictive Models

II

Experiments, Data Preprocessing and Extraction

3 Minimum Ignition Energies

Minimum Ignition Energy (MIE) is a vulnerability evaluation of flammable and explosive compounds, and will be studied for development of predictive models. This chapter serves to introduce this variable and the experimental procedure of determination. But first, the context in which it applies is exposed in order to better understand the interest and motivations of investigation.

3.1 Definition and Use

If a given process involves a combustible compound, the fire hazards incurred are considerable and should be treated in order to avoid ignition. Moreover, if the combustible is in the form of a fine powder that may be dispersed to form a dust cloud, the ignition is made even easier and the combustion propagation increases with the increase of contact surface area between combustible and oxidant. The same holds with gases and mists being more sensitive and burning more violently than flammable liquids [Carson and Mumford, 2002].

In confined spaces, or under specific concentration conditions of dusts clouds or deposits, the violence of the combustion reaction is such that the burning phenomena is no longer a fire but an explosion: the heat and byproducts released from a fire produce a sudden and important pressure rise. That is the definition of a chemical explosion, as given by Eckhoff [2003] :

> *"An exothermic chemical process that, when occurring at constant volume, gives rise to a sudden and significant pressure rise"*

It is important to note here that only the chemical explosions are considered and the physical causes of explosions, overpressure, overheat, or Boiling Liquid Expanding Vapor Explosions (BLEVE) will be disregarded.

Explosions are the results of the simultaneous fulfillment of six elements:

- presence of a combustible,
- formation of an explosive vapor or dust cloud,
- mixture concentration is within the explosive limits,
- presence of an oxidant, usually oxygen contained in air,
- an efficient ignition source
- and perhaps confinement - which may be considered as an aggravating factor rather than a requirement.

Confinement would accentuate the pressure accumulation during the explosion, and would lead to an increased severity of the explosion. Moreover, projectiles may be added to the chemicals and flame front propagation. However, open air explosions are possible, only the pressure wave will propagate more easily at the initiation of the reaction and thus the pressure differences are reduced compared to the confined setting.

To tackle the fire and explosion risks, preventive measures are directed at avoiding the encounter of all these elements. Preferred preventive strategies would favor organizational measures to eliminate the explosions hazards, and in the cases where the risk remains, technical measures should be implemented to reduce it.

To eliminate the risk, either one of the elements listed above should be removed:

- substitution of the combustible by another compound
- maintain mixture concentration outside the explosive limits,
- prevent the formation combustible cloud,
- remove oxidant,
- in last resort, remove ignition sources.

Considering that the hazardous combustible is necessary and could not be substituted, the other preventive measures are investigated. In order to practically implement these strategies, several physical characteristics are needed.

For a flammable mixture to ignite, concentration should be comprised within flammability or explosive limits, i.e. Lower and Upper Explosive Limits (LEL/UEL), and if the mixture is sufficiently dilute, the probability the mixture would ignite is eliminated. Then, with gaseous or vaporous compounds, working under inert gas or ventilation could be sufficient to maintain concentrations outside the explosive ranges.

It is also possible to control the working temperature to ensure safe conditions. The flash point temperature corresponds to the lowest temperature at which the vapor pressure of flammable liquid is such that a vapor/air mixture could reach flammable concentrations and would ignite if exposed to an ignition source. Thus, it could serve to establish a safe working temperature range below which the liquid is unlikely to form a flammable vapor. This may be sufficient to assess the probability of ignition of flammable liquids and hence, the flash point is a very widely used criteria. Yet, it is important to note that the flash point is measured under atmospheric pressure, and working under lower pressures would favor the vaporization.

When handling dusts, working under an inert atmosphere would remove the presence of the oxidant and remove the explosions risks. However, for practical and economical reasons, this strategy is rarely feasible. Low concentrations outside the LEL/UEL range may be ensured by introducing non-combustible compounds into the mixture. It is also possible to prevent dusts to form flammable clouds by modifying the particle size: besides the contact surface area argument mentioned previously, the higher the particle size, the lower the dust dispersion, and shorter the suspension time as the particle settle faster [Eckhoff, 2003]. Moreover, coal dust particles ignite at higher concentrations when the particle size increases[Man and Harris, 2014]. However, that does not alter the flammable nature of the substance and would not prevent combustion from occurring.

Therefore, after considering the replacement of the combustible, the control of its concentration, or prevention of the combustible cloud formation, the removal of the oxidant should be considered. An inert atmosphere with nitrogen or argon could be applied to reduce the oxygen content of the atmosphere, however this measure may be efficient for enclosed volumes but practically difficult at larger scales.

Finally, the ignition sources should be removed where the explosion risk remains. To fight against ignition sources, one must look at how they are created. Their causes are diverse and could be "direct" (i.e. open flame or hot surface), mechanically or electrically induced, result of an exothermic reaction, or due to an electrostatic discharge [Rogers et al., 2003].

In some cases, self-sustained combustion can even be observed in the absence of an ignition source other than a heat source or high temperatures. This is referred to as auto-ignition and is characterized by the lowest temperature at which this phenomenon could occur (Auto-Ignition Temperature AIT). Hence, working conditions must be set as to avoid at all cost that such mixtures are exposed to temperatures close to these limits.

Electrostatic discharges are the most complex to prevent as charge accumulation can occur during normal process and operations (i.e liquid flowing through pipes or powder charged in bags) [Eckhoff, 2003].

There are five types of electrostatic discharges, with different origins and resulting in different energy ranges [Stoessel, 2014]. As these energies could be sufficient to ignite some flammable compounds and not others, it is necessary to estimate the degree of vulnerability of compounds in presence with the energies that could potentially be released. While gases and vapors are much more sensitive to ignition than dusts, concentration controlling systems could be sufficient preventive measures but as mentioned above, it is not necessarily the case when dusts are involved. This raises an important criteria: the sensitivity towards ignition that will determine the types of ignition source to prevent against.

The Minimal Ignition Energy (MIE) serves to evaluate the sensitivity of an explosive dust to an electrical spark ignition. It is a measurement of the minimal amount of energy necessary for a flammable gas, vapor, or dust cloud to ignite. MIE measurements are performed as tube explosion tests conducted in modified Hartmann tubes, and the detailed procedure is presented in Section 3.2. The MIE is a required explosion characteristic under the EU Directive 99/92/EC on minimum requirements for the safety and health protection of workers potentially at risk from explosive atmosphere, referred to as ATEX 137 [Council of the European Union, 1999]. By comparison to known values, it serves to identify the types of electrostatic sparks that could ignite the compound.

The diagram represented in Figure 3.1 gives approximate energy values that electrostatic discharge can emit [Suter, 2008] . If a compound with a MIE lower than the maximal energy that the discharge can generate comes in contact with the spark, ignition could occur and give rise to a hazardous situation.

On this diagram, the green, yellow and red lines give three indicative thresholds: compounds with very small MIEs would graphically be represented beneath the red limit, which corresponds to extremely sensitive compounds, most often gases and vapors; between the red and yellow limits is a region of highly sensitive compounds as hybrid mixtures (mixture of gases and dusts in air) or highly flammable powders that could be ignited with energies less than 10 mJ; between the yellow and green lines, the energy necessary to ignite these compounds would be higher, between 100 mJ to 1000 mJ, which corresponds to normal combustible powders, with medium sensitivities; finally above the green line, a compound with a MIE of 1000 mJ or more would be considered not sensitive to electrostatic ignition and should be subjected to further investigation to determine the related risks of accidental combustion or explosion if these compounds were in contact with different ignition sources or under different conditions (contact with a hot surface, flammability of layer depositions, etc.).

Corona and brush discharges occur when the air in vicinity the edge of a conductive material becomes ionized and conductive. Corona discharges are the least energetic electrostatic discharges, and are rarely considered as hazardous, though they could ignite compounds with MIE in the 0.01 mJ to 0.1 mJ region which is the case for certain gases or vapors.

Brush discharges are more energetic as they can release up to 1 mJ to 3 mJ. Propagating brush discharges (referred to as “Prop. Brush” in Figure 3.1) can occur due to the high speed transport of liquids or powders in contact with an insulating material and are much more energetic and can release energies up to several joules. Sparks result from the ionization of non-conductive material between two conductive points, as for instance from the human body to another conductive element, and the released energies depend on the interacting bodies as shown in the diagram.

Finally, cone discharges are due to the charging of a powder when displaced or stacked and a “bulk cone” forms. As particles fall down the bulk in the container, they are charged due to the friction at the bulk surface. The energy that can accumulate depends on the path the particles follow and along which they are charged, which in turn depends on the diameter of the container (D=0.5 m to 3 m on the diagram). Moreover, it also varies with particle size: the larger particles will give rise to higher energies than the finer particles [Glor, 2003]. The two blue curves correspond to this behavior for particles with median size d=1 mm or d=3 mm.

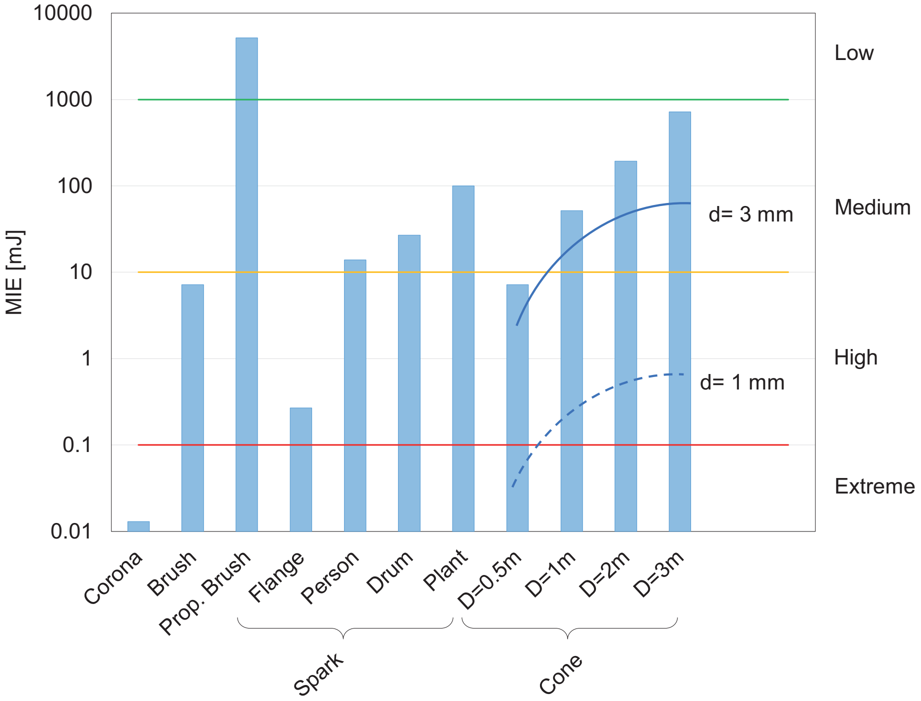


Figure 3.1 – MIE and Electrostatic Discharges - with ignition sensitivity thresholds, and cone discharges energies for different particles sizes

The previous discussion highlights the importance and the potential use of the MIE. For instance, considering the charging of a fine powder into a non-grounded container, sparks could be generated, the powder is likely to be in suspension while flowing down into the container and the explosion risks are high, and none of these elements can be taken away. To avoid any issues, the container sizing and in particular its diameter should then be fixed in function of the powder’s MIE, an earth grounding solution to or within the container should be designed, or if possible, provide an inert atmosphere to remove the oxidant. Besides the contained, the human body may also cause sparks of approximately 10 mJ, and when handling sensitive compounds, depending on their MIE, it could be necessary to provide earth bonding personal equipment for the operators.

MIE is a required information by the ATEX EU Directives and features among the necessary flammability and explosion characteristics of the compounds during the risk analysis that serves to identify potential explosive atmospheres which is the first step of the implementation of ATEX legislation. This allows to proceed with the second step by defining the zone classification which also depends on how regularly the explosive atmosphere might be present. The zones definitions according to the ATEX Directive are presented in table 3.1. The following steps of the procedure deal with the consequent choice and design of equipment, materials, and the assessment of where and which appropriate protective and preventive measures to implement (i.e. earth bonding, or personal antistatic precautions) [Janes et al., 2011].

Table 3.1 – Zones Definitions of Explosive Atmospheres

Explosive Atmosphere Duration and Frequency	Zone Gas/ Vapors	Zone Dusts
Continuously present, or for long periods, or frequently	0	20
Likely to occasionally occur in normal operations	1	21
Not likely to occur in normal operations, or only for a short period, or accidentally	2	22

3.2 Experimental Measurement

The assessment of ignition energy has been a concern for the mining industry since the early 20^{th} century due to the flammability of coal and coal dusts. The US Bureau of Mines and the UK Safety of Mines Research Establishment have been conducting systematic ignition testings since the 1910's. Over the century, the tests were improved and developed in parallel by different national institutes, and applied to other types of dusts from the agricultural and industrial worlds. Eckhoff [2003] reviews some significant developments brought to the ignition tests by scientists all over the world through the 1960's to the 1980's. Finally, in 1994, the International Electrotechnical Commission (IEC) set an international standard on the measurement of minimum ignition energy of dust/air mixtures [International Electrotechnical Commission, 1994] in order to ease the standardization of safe design of electrical equipment destined for use in explosive atmospheres. The ASTM International (formerly American Society for Testing and Materials) has also set an international standard accepted in 2003 [ASTM International, 2007]. Overall, the modern procedure to measure MIE is the following.

A dust sample is dispersed within an explosion vessel with pressurized air, and two electrodes connected to a circuit produce an electrical spark of known energy. If the spark induces observable flame propagation, the spark's energy is reduced by half until no flame propagation is observed for ten consecutive tests. The highest energy for which no flame propagation occurs is recorded. As the sample suspension concentration in air influences the ignition energy, a range of dust sample weights are tested successively to determine the most readily ignitable mixture, for which the lowest ignition energy is considered as the MIE.

The experiments are conducted in a modified Hartmann tube with a capacitor spark generator or with a commercially available device, Mike 3, which is schematically represented in Figure 3.2.

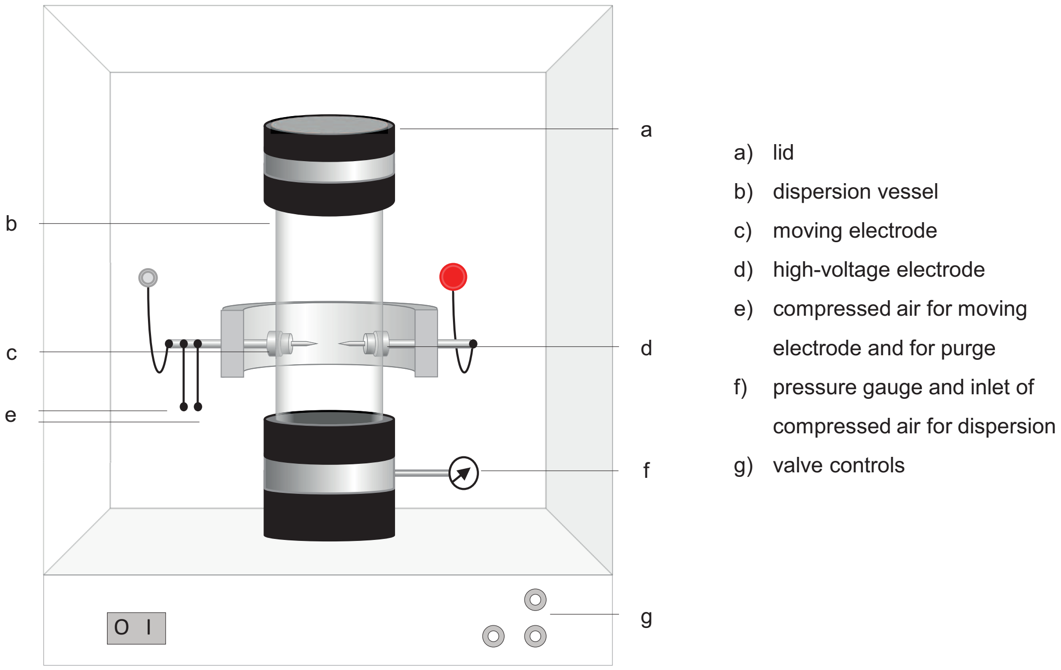

a
b
c
d
e
f
g
O I
a) lid
b) dispersion vessel
c) moving electrode
d) high-voltage electrode
e) compressed air for moving electrode and for purge
f) pressure gauge and inlet of compressed air for dispersion
g) valve controls

Figure 3.2 – Schema of a MIKE 3 Apparatus

The MIE is in reality a value comprised between E_1 the highest energy at which ignition failed ten consecutive times and E_2 the lowest energy at which ignition did occur [Cesana and Siwek, 2010]:

$$E_1 < MIE < E_2 \tag{3.1}$$

For a more accurate assessment of the actual sensitivity, the statistic energy is introduced and can be calculated as follows:

$$log(E_S) = log(E_2) - \frac{I[E_2].(log(E_2) - log(E_1))}{(NI + I)[E_2] + 1} \tag{3.2}$$

where $I[E_2]$ is the number of tests at energy E_2 where ignition occurred, and $(NI+I)[E_2]$ is the total number of tests at energy E_2 with and without ignition observed. An example taken from work of Cesana and Siwek [2010] is represented in Figure 3.3 and shows how the calculation of E_S is performed: once the preliminary tests determined E_1 and E_2, the probability of ignition is determined according to the test results from different sample weights tested at E_1 and E_2. In this example, the probability of ignition is 2 out of 5, and calculated statistic energy is E_S=21 mJ.

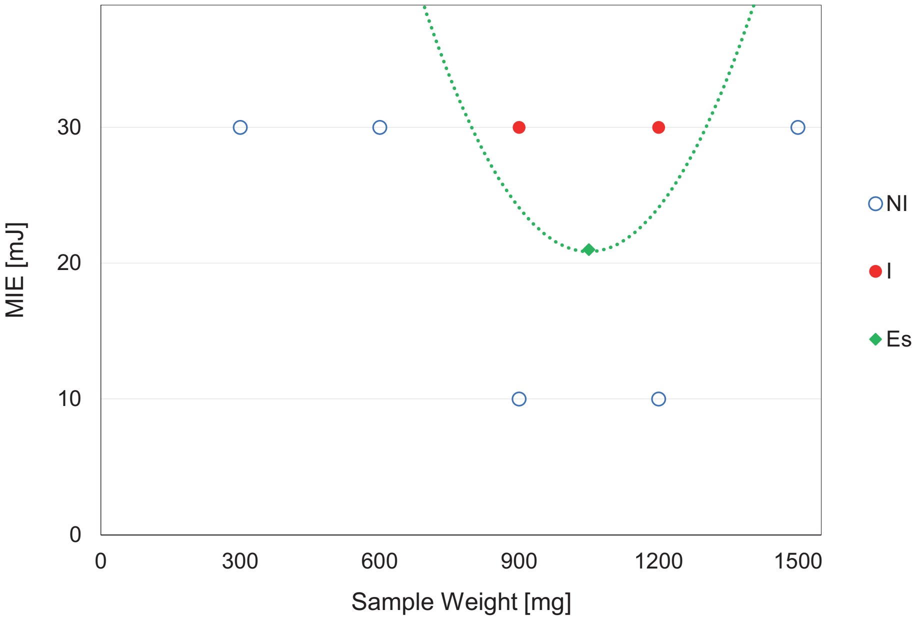


Figure 3.3 – Statistic Energy: Calculation Example

The experimental procedure to obtain this result is tedious: one has to conduct all the necessary tests prior to narrowing the MIE between E_1 and E_2, then between 50 and 70 tests are required as for indicating NI for a given sample weight and energy, one has to repeat the test ten consecutive times. Therefore, usually, it is E_1, the highest energy at which ignition failed ten consecutive times, that is reported to be the MIE.

Some studies propose to reduce the amount of necessary trials to determine MIE more accurately, by means of a slightly different probabilistic approach: with the same information, E_1, E_2, number of tests with ignition observed or not, but with a reduced number of attempts, a normal or a log-normal law partition function serves to define a probability of ignition as a function of energy [Bernard et al., 2010]. The authors consider this estimation method to be rather accurate, and allows to define several energy levels as $E_{0.05}$ for a 5% probability of ignition or $\mathrm{E}_{0.95}$ for 95% probability of ignition according to the risk level one is ready to take or the applications.

The MIE determination for gases and vapors is performed differently: the mixture is introduced in a 2 L vessel in which concentration is controlled by temperature variations, and ignition is induced with a calibrated spark generator. Energies of 1 mJ or 0.4 mJ can be applied with 5% confidence and for lower energies, calibrations with products of known MIE are used as comparative points. The procedure is tedious as well and requires numerous tests to obtain accurate results. Hence, predictive models for the MIE could be beneficial to reduce the efforts required by the experimental determination.

3.3 Influencing Parameters

Explosive and flammability characteristics of gases and dusts exhibit numerous similarities and are therefore treated by harmonized regulations. The dependance to flammability limits, similar burning mechanisms and detonation phenomena, well defined minimum ignition energies or temperatures are few examples [Eckhoff, 2003]. The influence of the concentration, which is valid for gases, vapors and dusts clouds, lies on the necessity of low inter-particles distances for the ignition propagation within the explosive cloud.

However, Eckhoff [2003] stresses two fundamental differences: firstly, vapor or gases only propagate a flame when the mixture formed with air lies in the flammability range, while settled or layer deposits of dusts could propagate a flame due to the presence of air in the inter-particles spaces; secondly, the dusts do not create ignitable clouds under all circumstances.

Several parameters have influence on the dusts cloud formation and thus, on the ignition sensitivity.

The particle size and the powder density influence the dispersibility and the settling velocity of suspended dusts. Very fine powders tend to form agglomerates, and are therefore less likely to disperse in a gas phase. On the other hand, coarse particle sizes are also less likely to disperse, settle faster and are less sensitive to ignition sources. Dusts with particle sizes higher than 500 µm do not explode [Bartknecht, 1989]. Similarly, powders of higher densities also tend to settle faster due to higher particle weights.

Usually, dusts with particle size in the 1 µm to 100 µm range would exhibit the highest sensitivities, and therefore, it recommended to measure the explosibility characteristics of powders after sieving the powder to obtain particle of less than 63 µm [Bartknecht, 1989].

The moisture content may also increase the tendency to agglomerate and decrease the ignition sensitivity [Eckhoff, 2003].

Finally, technical aspects of the measurement can affect the MIE measurement. Experiments conducted on the Hartmann tube and the Mike 3 apparatus have shown that the dispersion method and the delay time elapsed between dusts dispersion and spark generation create different turbulence regimes of the dusts cloud. Thus, real concentrations at the initiation time near the spark region differ from the concentration estimation based on the dusts sample weight. Therefore, the ignition could be provoked or not and give rise to different MIE measurements [Janes et al., 2008].

3.4 Data Collection and Treatment

For this study, the MIE values collection were not experimentally measured but gathered from several reference sources: works by Babrauskas [2003], Hertzberg et al. [1992] and Grossel [1988], authors of various handbooks on ignition, dust explosion and the safe handling of powders, report numerous data of MIE. Tables A.1, A.2, and A.3 in the Appendix A summarize the MIE data used and the sources that originally reported them.

More than 130 data were found in literature, corresponding to gases, vapors and dusts. A first selection was conducted in order to separate the data according to physical state at 20°C (the melting and boiling points are also reported in appended tables) and ensure they are considered apart. Nevertheless for modeling purposes, they are later intentionally merged as explained in Chapter 5.

All molecular structures were created and processed with the Codessa Pro Software [Petrukhin et al., 2001] in order to generate the molecular descriptors necessary for the modeling. In total, and after removal of the descriptors with lowest variances or that were not defined over the complete dataset, 357 descriptors were recorded for 132 compounds.

3.5 Conclusion

We have seen that if a risk analysis highlights the presence of an explosive compound or the probability of formation of an explosive atmosphere, many preventive measures could be implemented. However, in some cases, it could be necessary to remove the ignition sources and in order to target in priority those that can actually trigger the sensitive mixture, one should assess its sensitivity. That is where the MIE is of great importance, and therefore is required by the ATEX Directive.

The experimental procedure to determine the MIE was presented, and the influencing parameters were reviewed. Considering the time and efforts required for a single MIE measurement, it could be highly advantageous if predictive models could be proposed to aid the experimental determination.

Therefore, a dataset was collected and prepared for the development of molecular based predictive modeling. The results are presented in Chapter 5.

4 Differential Scanning Calorimetry

Heat has remarkable effects on matter. It drives most matter transformations, either chemical, physical or mechanical. Heat is a form of energy that may affect the internal energy of molecules. Hence, it influences many of their microscopic and macroscopic properties. A body subjected to a heat input, can either absorb this energy and increase its temperature or release this energy in another form, i.e. work. Heat transfers through solids or fluids by conduction, through fluids by convection or by radiation even through vacuum. As heat itself is not a measurable variable, heat exchanges and heat variations symptoms, as the temperature changes, can be observed to determine heat variations.

As heat plays an important role in most matter transformations, it is omnipresent in human activities and examples range from cooking pasta to launching spatial probes or producing daily used objects. For instance, phase transitions occur with internal energy changes. That is why processing plastic materials requires to heat them to obtain a ductile and malleable body in order to easily perform the desired transformation, and later to cool them down to preserve the newly formed body from unintended deformations.

In chemical reactions especially, heat is of crucial importance. Chemical reactions consist in breaking and making new bonds, accompanied by energy consumption (endoenergetic) or release (exoenergetic) expressed as heat exchange. Heat governs the rates of the transformations, and therefore applying heat to a seemingly inert body can accelerate the transformation. In all cases, controlling the operating conditions (pressure and temperature) of chemical production is a mean of controlling the rate of the transformation.

For productivity purposes, operating conditions are usually set to favor high yield and throughput of the desired product, which often involves high temperatures or pressures. Nevertheless, when the reaction is exoenergetic, the conditions must be designed to evacuate the heat produced during the reaction. Heat accumulation within a reactor, or any containing vessel, may be hazardous as it will have several consequences: pressure and temperature rise, acceleration of the ongoing reaction, degradation of the present products by triggering of potential side-reactions. A thermal runaway takes place when the heat accumulation causes the increase of the reaction rate and the heat production rate, which in turn contributes even further to the temperature rise. The apparatus needs to be designed to face these conditions, so that the consequences of heat accumulation could be managed; otherwise risks of severe consequences are incurred.

In order to design the adequate equipment and to apply optimal conditions for safe operations, it is necessary to properly assess the thermal behavior and the calorimetric potential of the handled reaction and compounds. A safety analysis may rely on several experimental measurements.

Calorimetric analysis are the experimental measurements of heat flow rate during a chemical or physical transformation. Either the heat flow rate is directly measured or indirectly through measurement of the temperature evolution, for instance. In reaction calorimetry for example, the production operating conditions are replicated as closely as possible, at a smaller scale (mL to L scale), and the heat production occurring during the reaction is deduced from comprehensive energy balance over the reactor. The Accelerating Rate Calorimeter (ARC) reveals how a sample mass behaves in adiabatic mode (when the accumulated heat is not evacuated) and can reveal important information regarding how the sample reacts in case of cooling failure [Stoessel, 2008]

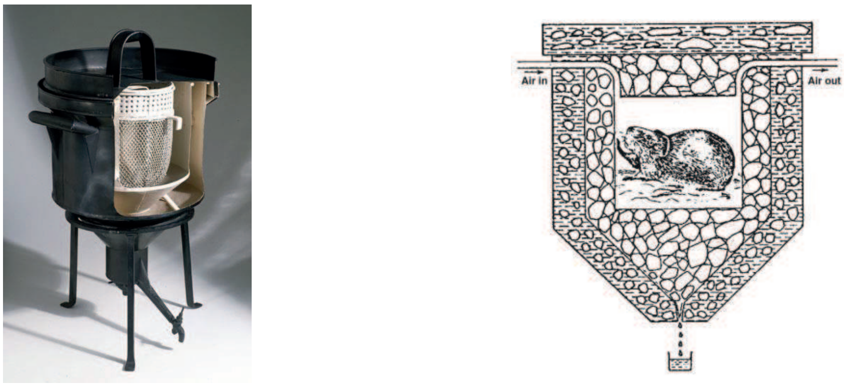
Thermal analysis techniques, which follow the evolution of physical and chemical properties as function of temperature, may also be employed to gather information on a sample thermodynamic behavior. These methods include thermogravimetry for instance, that follows mass variations, or Evolved Gas Analysis (EGA), that monitors gaseous products resulting from thermal decomposition.

Differential Scanning Calorimetry (DSC) is at the crossroad of calorimetry and thermal analysis. DSC is a particular analysis in which a sample and a reference are subjected to a temperature pre-programmed profile and the heat exchange at the sample is measured. These measurements can reveal important information regarding the tested compounds and the reaction they can undergo. Besides heat potential evaluation of sample, the applications of DSC are broad as it can be employed for identification and purity evaluation of compounds, determination of phase diagrams, or kinetic investigations [Höhne et al., 1996], and this comprehensive overview makes it well adapted for safety studies.

This chapter will focus on the functioning principles of DSC in a first section. Then, the possible experiments that can be run and the information they expose will be discussed. In the third section we will detail how DSC experiments will be exploited in this project in particular and how the DSC experiments are pre-treated prior to modeling.

4.1 DSC Principles

The first calorimeter was developed by Lavoisier and Laplace around 1780 (according to Rawlinson [2006], Sarge et al. [2014a]). The experiment they designed involved the burning of lamp oil in a meshed container placed in a double-walled vessel. Both chambers contained ice, and were connected to collecting funnels to evacuate melted ice. The purpose of this setup was to trap the middle layer of ice between the heat producing combustion and the outer ice layer, which protects the middle ice layer from melting due to heat exchange with the surrounding at ambient temperature. Then, the melted ice was collected, weighted and the heat production deduced from the latent heat of fusion of ice (already known since the 1760's thanks to the work of Joseph Black, as reported by Emeis [2004]) now referred to as melting enthalpy. This setup also served to prove and quantify the heat production during metabolic processes with the famous experiments conducted on a guinea pig, reported by Höhne et al. [1996], Holmes [1987], Rawlinson [2006], Sarge et al. [2014a].



(a) Calorimeter of Lavoisier and Laplace (from Science Museum, London [Ice calorimeter])

(b) Guinea Pig experiment (from [Sarge et al., 2014a])

Figure 4.1 – Calorimeter of Lavoisier and Laplace

The heat released by the studied body, be it the oil lamp or the guinea pig, is absorbed by the experimental setup, i.e. the middle ice layer. Besides, the system is insulated from the surrounding environment thanks to the outer ice layer. Therefore, the temperature is maintained constant and the heat produced does not cause any temperature elevation, only the phase transition of the melted ice. This system is considered as heat-compensating as the temperature is held constant by evacuation of the produced heat.

Calorimetric measurements evolved since then in terms of precision and technology and several designs arose from the initial basic calorimeter to allow for various experiments. In Differential Scanning Calorimetry, the sample is not held at a constant temperature, it is subjected to a temperature scan: the temperature is linearly varied between an initial temperature $\mathrm{T_i}$ to a final temperature $\mathrm{T_f}$ at a constant scanning rate β. Usual temperature scan rate values range between $0.5\,\mathrm{K\,min^{-1}}$ and $10\,\mathrm{K\,min^{-1}}$.

Moreover, to be able to assess the heat production/consumption by the sample, it is compared to a reference sample placed in the exact same environment either placed symmetrically in the same furnace or in an identical one (referred to as twin design).

The conduct of typical DSC experiment is as follows:

- the sample is prepared, weighted and placed in a container, known as crucible; depending on the type of crucible, it may be sealed;
- the reference, typically an empty crucible, is also prepared;
- both the sample and the reference are placed in the furnace on their respective holders; lids may be placed above the samples in the furnace to minimize heat losses;
- the desired temperature profile is programmed (either isothermal, heating or cooling, initial and final temperatures and scan rate), information relative to the sample (material, weight, etc) and the experiment are entered in the program and the experiment is started;
- during the experiment, the temperature of both samples is very precisely monitored;
- the temperature records are automatically analyzed and the heat flow rate is directly computed and graphically displayed as a function of time or temperature to render the DSC curve, referred to as a thermogram.

The apparatus shall be regularly calibrated in order to obtain reliable quantitative measurements. An experiment with standard reference material as the sample is conducted. Standard materials are usually metals of known melting enthalpies; this allows to match the measured heat flow with the known heat flow. Similarly, the temperature is also calibrated thanks to melting points of metals that occur at a precise temperature (given that the metal is pure). Multiple calibrations points through several reference materials are often necessary and a minimum of three is recommended [Gmelin and Sarge, 1995]. Besides, calibration determines several apparatus-depending parameters. For instance, the heat transfer from the furnace to the sample is not immediate and depends on the heat capacities of the crucibles and holders that are the main heat conduction path. The calibration experiments should be performed in conditions as close as possible to the experimental conditions: similar crucibles, reference, and temperature profile should be employed in the measurements and the calibrations.

There are two DSC systems that are employed nowadays with different functioning principles: power compensation DSC and heat-flux DSC. In both cases, the procedure detailed previously is applied.The main difference lies in the tracked parameter. In power compensation DSC, the pre-programmed temperature is strictly enforced; if the sample consumes or generates energy, the power supplying system compensates this heat source to maintain the sample's temperature identical to this of the reference. In heat-flux DSC, both the sample and the reference are supplied the same heat flux, and their temperatures allowed to evolve differently. These two systems are presented in the following sections.

4.1.1 Power Compensation DSC

Power compensation DSC was developed in 1964 by Watson et al. [1964] for Perkin-Elmer Corporation, who still commercializes this type of apparatus. It was not the first type of differential analysis as the Boersma Differential Thermal analysis (DTA) already existed then [Boersma, 1955]. Nevertheless, the power compensation DSC is more similar to the historical example seen above as it operates under the heat-compensation principle.

Heat compensating consists in neutralizing the sample heat consumption or production in order to suppress its contribution to temperature variations. The sample and the reference are maintained at the same temperature despite the transformations the sample may experience. For this, the temperatures are closely monitored with sensors placed in the holders. The temperatures are processed by the CPU in a two-fold analysis. On one hand, the average temperature control system ensures that if the sample and reference average temperature differs from the programmed temperature profile, the power supplied to both heaters is varied consequently to come closer to the desired temperature. On the other hand, differential temperature control monitors the sample and reference temperatures: if they differ from each other, the power supplies are individually varied in order to recover equal temperatures. If the sample temperature is lower than the reference, higher power is fed to its heater; otherwise, it is the reference material that receives an increased power input.

During an endothermic transition, the energy consumed by the sample is equal to the extra energy the system fed to the sample relative to the reference; inversely, during exothermic reactions, the energy generated by the sample, is equal to the energy fed to the reference to preserve the thermal balance. Therefore, the system provides a direct measurement of the heat flow from and to the sample as equal to the compensating energy and reported graphically as a function of temperature or time. The system is schematically represented in Figure 4.2.

The heat flow rate fed to both holders $\mathit{\Phi}_{\mathrm{F-SR}}$ is proportional to the temperature difference between the programmed temperature and the average temperature of the sample and reference $\mathit{\Delta}\mathrm{T}_{\mathrm{P-SR}}$:

$$\mathit{\Phi}_{F-SR} = -k_1 \mathit{\Delta} T_{P-SR} \tag{4.1}$$

where k_1 is a proportionality factor set by the controlling unit.

Similarly, the reference and the sample receive individual heat feed (noted $\mathit{\Phi}_{\mathrm{FS}}$ or $\mathit{\Phi}_{\mathrm{FR}}$ for

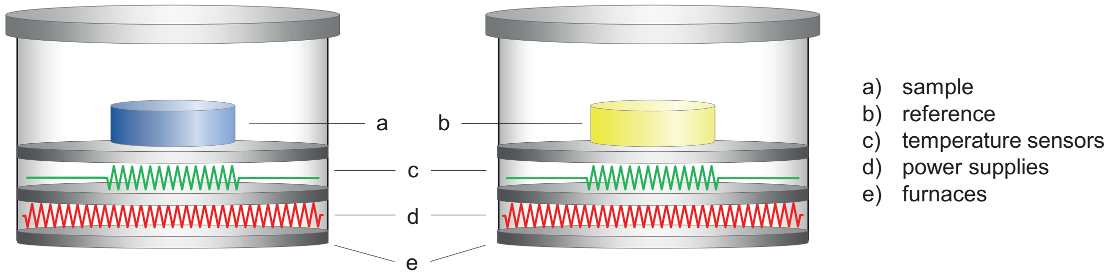


Figure 4.2 – Schematic representation of power-compensated DSC

sample and reference respectively) proportional to the temperature difference between them ΔT_{SR} when it is non-null.

In order to recover the heat flow of interest, i.e. this related to the reaction or transition the sample undergoes Φ_r, the heat supplied to the sample and reference respectively are compared:

$$\begin{aligned}\Phi_r &= \Phi_{FS} - \Phi_{FR} \\ &= -k_2 \Delta T_{SR} + k_3 \Delta T_{SR} \\ &= K \Delta T_{SR} \end{aligned} \tag{4.2}$$

where k_2, k_3 and K are also proportionality factors set by the controlling unit.

Overall, the measured signal of ΔT_{SR} is directly proportional to the heat flow production or consumption by the sample Φ_r, and equal to the heat flow that should be supplied to the sample Φ_{FS} or removed (actually fed the reference Φ_{FR}) in order to maintain the sample and the reference at equal temperatures [Höhne et al., 1996].

Nonetheless, this discussion considers an ideal case and neglects the heat conduction time between the measurement point and the sample. As the temperature is measured beneath the sample holder, the heat conduction path is short and can be neglected in the discussion. However, in practice it is taken into account thanks to calibration measurements to determine the time constant τ which depends on the sample heat capacity C_S and the global heat flow resistance between the measurement point and the sample $\mathrm{R_{MS}}$:

$$\Phi_S = \Phi_M + \tau \cdot \frac{d\Phi_M}{dt} \qquad \text{with} \qquad \tau \approx C_s \cdot R_{MS} \tag{4.3}$$

where Φ_S is heat flow rate from the sample and Φ_M the measured heat flow rate.

4.1.2 Heat-Flux DSC

In heat-flux DSC, the apparatus design and the experience control are slightly different from power-compensated DSC. First of all, both the sample and the reference are placed in a single furnace, and are subjected to the same heating source. Both crucibles are placed on their respective holders connected to the temperature sensors. The holders are often designed as a metallic or ceramic disk where samples are placed and in which the thermal resistors are directly embedded [Sarge et al., 2014b]. This design also ensures symmetrical positioning of both samples in the furnace.

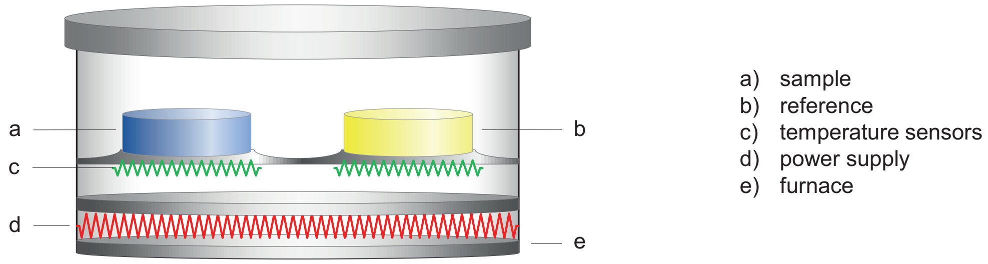


Figure 4.3 – Schematic Representation of Heat-Flux DSC

In the furnace, due to this symmetrical arrangement, if the sample and the reference are at the same temperature, the heat exchanged between the furnace and the sample Φ_FS is the same than between the furnace and the reference Φ_FR.

If a temperature difference is detected between the sample and the reference $\Delta\mathrm{T_{SR}}$, it is necessarily caused by transitions or reactions occurring in the sample. This temperature difference will give rise to a heat flow rate Φ_SR between the sample and the reference, and the new balance is the following:

$$\Phi_r = \Phi_{SR} = \Phi_{FS} - \Phi_{FR} \tag{4.4}$$

Considering the sample and the reference individually, the following heat flow balances are valid:

$$C_S \cdot \frac{\mathrm{d}T_S}{\mathrm{d}t} = \Phi_{FS} + \Phi_r \qquad \text{for the sample} \tag{4.5}$$

$$C_R \cdot \frac{\mathrm{d}T_R}{\mathrm{d}t} = \Phi_{FR} \qquad \text{for the reference} \tag{4.6}$$

Now, considering the heat flow rates in the furnace

$$\begin{aligned} \Phi_{FR} - \Phi_{FS} &= \frac{T_F - T_R}{R_{FR}} - \frac{T_F - T_S}{R_{FS}} \\ &= \frac{\Delta T_{SR}}{R_{th}} \end{aligned} \tag{4.7}$$

where $\mathrm{R_{FS}}$ and $\mathrm{R_{FR}}$ are the heat transfer resistances between the furnace and the sample and the furnace and the reference respectively, and due to the symmetrical design of the furnace and holders $\mathrm{R_{FS}} = \mathrm{R_{FR}} = \mathrm{R_{th}}$ with $\mathrm{R_{th}}$, the global heat transfer resistance of the system[1].

Finally, by substituting Equations 4.5, 4.6 and 4.7 into Equation 4.4, the reaction heat flow rate becomes:

$$\Phi_r = -\frac{\Delta T_{SR}}{R_{th}} - \beta \cdot (C_S - C_R) - C_S \frac{\mathrm{d}\Delta T(t)}{\mathrm{d}t} \tag{4.8}$$

where $\Delta \mathrm{T_{SR}}$ is the temperature difference between the sample and the reference (the measured signal), $\mathrm{C_S}$ and $\mathrm{C_R}$ are the heat capacities of the sample and the reference respectively, β the scan rate [Höhne et al., 1996].

Hence, the reaction heat flow rate is not directly proportional to the measured temperature difference, but time-shifted due to two elements taken into account by the second and third

---

[1] The notation $\mathrm{R_{th}}$ stands for "Thermal Resistance" and is used here only to avoid confusion with R the gas constant, which will be encountered later.

terms of the Equation 4.8: the inherent heat capacities differences between the sample and reference due to the sample nature and weight principally, and the resulting thermal inertia of the sample. As the reference is usually an empty crucible, the sample is necessarily heavier and even if massic heat capacity remains unchanged, a lag appears and has to be taken into account to recover the signal of the heat flow from the reaction mass (τ is the time constant relative to the sample thermal inertia $\tau \approx C_s \cdot R_{th}$ as seen for the power compensation DSC in Section 4.1.1).

Thus, from the measurement of $\Delta \mathrm{T}_{\mathrm{RS}}$ it is possible to recover the reaction heat flow rate Φ_{r}, given that the necessary calibration experiments were conducted in order to determine all the other parameters.

It is important to note that this holds under two approximations: that the heat transfer is mostly conveyed through conduction rather than convection or radiation; and that the thermal behavior is governed by the crucibles and sample holders, i.e. that their thermal resistance and heat capacities are larger than this of the studied substances [Höhne et al., 1996]

4.2 Thermogram Analysis

Independently of the apparatus employed, the results of a DSC experiment are usually the same: the records of sample and reference temperatures and sample heat flow rate as function of time or temperature. Depending on the temperature profile applied, heating, cooling or isothermal, and on the sample thermal behavior, the thermal curve or thermogram, may exhibit several elements.

Figure 4.4 shows an example of typical DSC thermogram recorded under a heating program and the thermal events observed are marked from (a) to (d). The information and characterization analysis that may be drawn from a DSC experiment are detailed below (labels (a) to (d) refer to the elements of Figure 4.4) .

(a) Heat Capacity: the heat flow rate from the sample, when no thermal events are observed besides the heating or cooling phenomena, is proportional to the heating rate β and the heat capacity of the sample $\mathrm{C}_{\mathrm{p,S}}$. As it happens in Figure 4.4, the heat flow signal is not normalized by the sample weight or molar quantity, thus to recover the specific heat capacity, one can simply compute it as:

$$\overline{C_{p,S}} = \frac{1}{m} \cdot \frac{\Phi}{\beta} \tag{4.9}$$

It is important to note that the heat capacity varies during the experiments as it is a temperature dependent factor, hence the measured value is indeed the average of heat capacities over the studied temperature range and marked $\overline{C_{p,S}}$, but also due to the

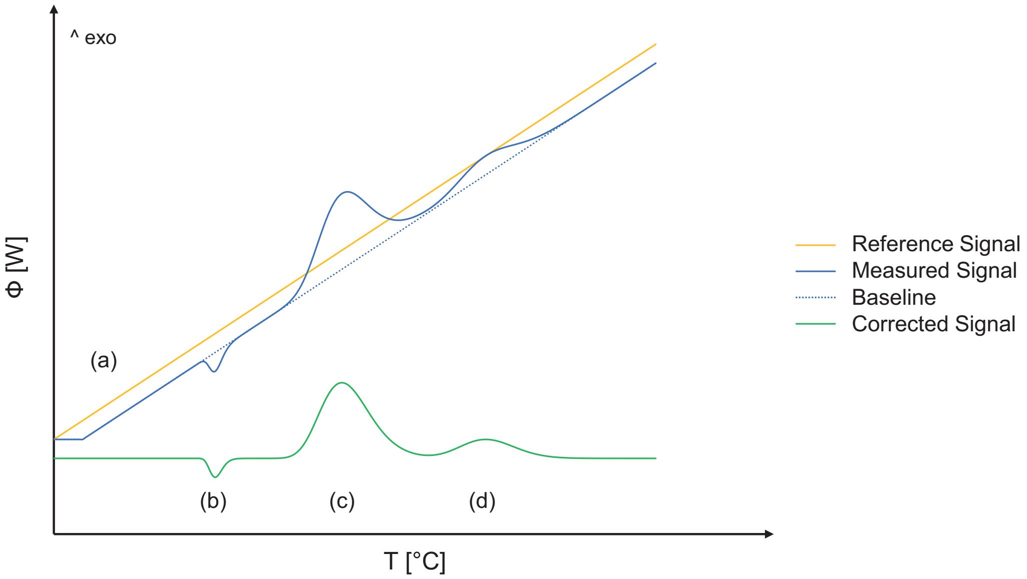


Figure 4.4 – Example of a Typical DSC Thermogram

phase or nature changes of the sample caused by the transformations. The slope or the curvature of the baseline may be affected, and the proportional relationship holds only on narrow temperature spans.

(b) Melting Point: DSC can be applied to accurately determine the phase change of samples, especially the melting temperature and enthalpy of solids. Moreover, the melting point analysis can also reveal the purity of the sample Rawlinson [2006] (see Figure 4.5).

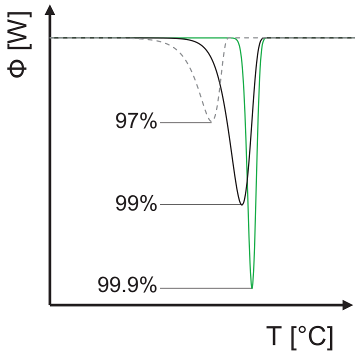


Figure 4.5 – Purity Effect on Melting

At the highest purity, the melting point is characterized by a sharp endothermic peak with linear slope that reaches the minimum at the compound's exact melting temperature. In presence of impurities, the peak tends to broaden and occur at lower temperatures than the actual melting point. The melting enthalpy is the integral of the curve in this temperature range or the area under the melting point peak.

(c) Chemical Reaction: depending on the kind of experiment conducted, the tested sample can either be a pure compound or a mixture. Reactions can either be endothermic or exothermic as represented in Figure 4.4 (c). In the case of a pure compound, such an exothermic peak indicates thermal decomposition, oxidation or polymerization of the sample. For a reactive mixture, the peak could reveal a chemical reaction between the present compounds.

Several elements can be withdrawn of the analysis of the curve. The onset temperature (T_o) may be defined as the temperature at which the reaction is progressing at a significant rate. The difficulty to assess what "a significant rate" is, makes the onset temperature determination nonstandard and subjective to operators and protocols.

There are however two common methods to determine it, either as the temperature at which the curve reaches a given percentage (10% in Figure 4.6) of the maximum or through extrapolation of a tangent [Sarge, 1991]. The tangent is by definition built at the inflection point of the peak, i.e. on the ascending part of the peak, when the slope is maximal. However, the result is not necessarily ideal, as the peak itself is rarely ideal and doesn't exhibit a linear heat flow rate. Therefore, most software packages automatically calculate the tangent and allow the user to adjust it to an "auxiliary" line that fits to a certain extent the "almost linear" part of the peak [Höhne et al., 1996, Sarge, 1991]. Considering that the baseline itself can be determined by various methods [Saito et al., 1986] and that the onset extrapolation parameters are dependent on the sample characteristics [Sarge, 1991], the onset temperature determination can be rather challenging in some cases.

Besides the onset temperature, the reaction enthalpy (ΔH_r) is also calculated from the characteristic peak of a reaction as the area under the curve. Both of these elements are represented in Figure 4.6.

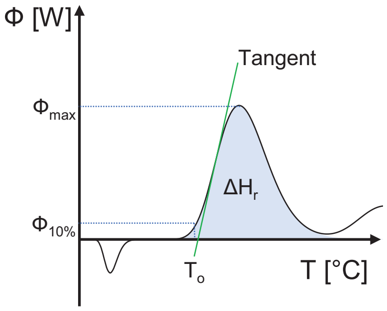


Figure 4.6 – Onset Temperature and Reaction Enthalpy

(d) Secondary Reactions: Figure 4.4 exhibits a first reaction peak on (c) and a second one on (d). In the cases where the tested sample is a mixture, the reaction (c) could be the desired chemical reaction under investigation. Then, the secondary reaction may be a side reaction as for example the decomposition reaction of the first reaction products, and may be unexpected. Indeed, the main reaction could be identified from the usual laboratory experiments, yet, the temperature is usually not raised to elevated temperatures and this decomposition could be missed. Therefore, in DSC tests, the scanned temperature range should go beyond the working temperature range in order to shed light on potentially unnoticed reactions. From a safety point of view, these reactions could be highly important especially if the heat release potential is high or if the onset temperatures are neighboring from the working temperature range. From a quality point of view, it is equally important to set the operating conditions such that the freshly produced chemicals are not directly decomposed in a secondary reaction.

(e) Kinetic Behavior: the kinetics of a reaction are latent information that is not explicit in a thermogram like Figure 4.4. During a DSC experiment, when the temperature is sufficiently high for the reaction to take place at a significant rate (i.e. $T > T_o$), two elements contribute to the variation of the reaction rate r:

$$r = k \cdot C^n = k[C_i(1 - X)]^n \tag{4.10}$$

where k is the reaction rate constant, C the reactant concentration, n the reaction order, and X the conversion of the reactant.

First, with the reactant consumption, the concentration C decreases and its influence on the rate depends on the reaction order n. Second, the reaction rate constant k, is not a constant anymore when the temperature varies (Arrhenius law), which is obviously the case in a DSC test. Therefore, the heat flow rate generated by a first-order reaction may be expressed as follows:

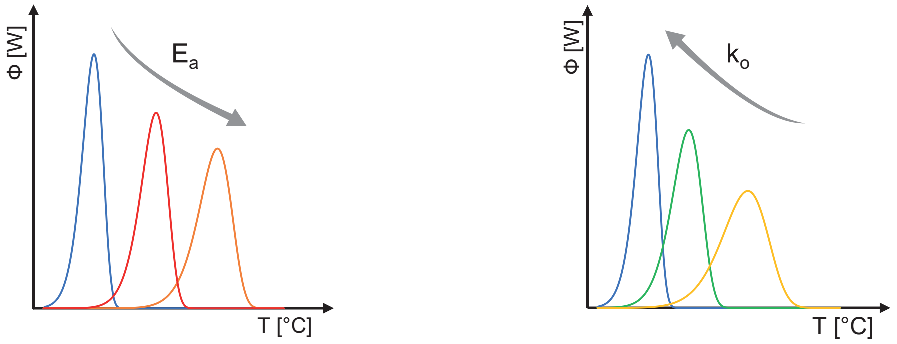
$$\Phi_r = r \cdot V \cdot \Delta H_r = k_o \cdot exp\left(\frac{-E_a}{RT}\right) \cdot (1 - X) \cdot n_i \cdot \Delta H_r \tag{4.11}$$

where ΔH_r is the molar reaction enthalpy, k_o the pre-exponential factor in Arrhenius law, E_a the activation energy, R the gas constant, and n_o the initial molar quantity of reactant (not to be confused with the reaction order n).

The impact of this relationship cannot be observed in one thermogram alone. However, comparing several curves can serve to illustrate this discussion. In Figure 4.7, for a virtual set of reactions of equal heat of transformation ($n_o \cdot \Delta H_r$), the pre-exponential factor k_o and the activation energies E_a are varied and the corresponding heat rates represented. This highlights that reactions with rapid kinetics (i.e higher k_o, lower E_a) exhibit higher and narrower peaks. This indicates that the reaction rate accelerates at a faster pace, and consumes reactants in shorter times. The broader and smaller peaks are symptomatic of reactions with slower reaction rate accelerations, and hence slower heat production.

In the case where the kinetic information of the reaction is unknown, they could be determined from DSC experiments. Several methods based on model-fitting to one single experiment operate by adjusting kinetic parameters, reaction order and reaction mechanisms to measured heat flow rate [Borchardt and Farrington, 1957]. Nonetheless, single measurements are considered insufficient for accurate estimations. For this purpose, several DSC experiments should be performed at different scanning rates (β), and the conversion should be computed from the obtained heat flow rates. Figure 4.8 shows examples of DSC curves obtained for the same sample tested at different scanning rates and the corresponding conversion in function of temperature. Iso-conversional methods [Flynn, 1983, Ozawa, 1965] imply to determine the temperatures at which an equal conversion is achieved at different scan rates, to solve and determine for k_o and E_a. Finally, isothermal techniques are more appropriate for auto-catalytic reactions, as the dynamic measurements could cover the self-accelerating behavior [Sourour and Kamal, 1976] .

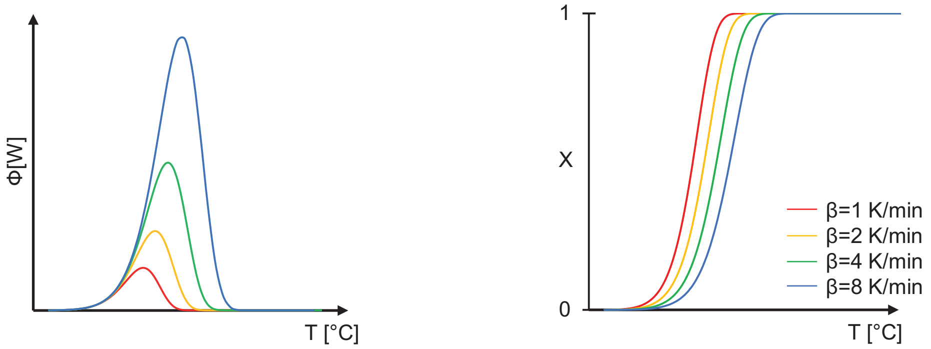
Figure 4.8 illustrates the iso-conversional method. First, DSC records of the sample are conducted at different scan rates, and the result obtained is similar to Figure 4.8(a). From these curves, the conversion X is calculated as the integral of this signal as a function of the temperature, and the result obtained is shown in Figure 4.8 (b). Then, Equation 4.11 can be used to set a system of equations for a given X, where T and Φ_r are known and E_a and k_o can be determined.



(a) Heat flow rate vs temperature for different activation energies E_a

(b) Reaction conversion vs temperature for different pre-exponential factors k_o

Figure 4.7 – Kinetic Behavior on DSC - gray arrows indicate increase of parameter



(a) Heat flow rate vs temperature for different scan rates

(b) Reaction conversion vs temperature for different scan rates

Figure 4.8 – DSC for Kinetic Evaluation

4.3 Data Collection, Treatment and Property Extraction

In the previous section, the procedure to measure DSC thermograms has been detailed and the important information to analyze were reviewed.

For this work, the focus will be on thermal decomposition reactions of pure compounds. The next section will present briefly the experimental conditions and characteristics of the data that will serve for the model building phase.

Moreover, considering the other information contained in DSC curves, it has been decided not to disregard the entire curve and keep only the onset temperature and enthalpy of decomposition reaction, but to preserve the overall thermogram, or most of it, such that if further developments (out of the scope of this project) would be found to complement our work, other parameters such as the reaction kinetics would be recovered eventually.

Therefore, the following part will present how the DSC curves are treated in a manner to abstract them to few parameters that would allow a comprehensive reconstruction.

4.3.1 Data Collection

For this work, the data collection has been conducted in three phases. In the first part, the DSC records of 20 nitroaromatic compounds were performed in heat-flux DSC apparatus by Mettler-Toledo. Later, a set of 20 chemicals were tested by heat-flux and power-compensation DSC. Each sample has been tested several times and these experiments are destined to evaluate the experimental error of the measurements on both apparatus types. Moreover this set has been designed to be structurally diverse and composed of compounds belonging to several chemical families were selected: peroxides, azo-compounds, nitrites and nitrates.

Finally, a collection of about 900 DSC records were acquired from an industrial collaboration[2].

All experimentally measured DSC experiments were recorded at the following conditions:

• Temperature range scanned:	30 °C to 400 °C
• Heating rate:	4 K min^{-1}
• Sample weight:	2 mg to 5 mg
• Enclosing atmosphere:	Air, RTP
• Crucible:	Gold-plated, sealed, high pressure resistant, 20 µL

The data collection from the industrial partner presented few differences. The crucibles were enclosed under inert atmosphere, i.e. argon gas, and the sample weights could go up to 10 mg.

Due to the absence (or low concentration) of oxygen, it is possible to affirm the thermal events observed on the DSC records are pure compound decomposition and not oxidation reactions. Figure 4.9 presents an example of compound (i.e. pentanenitrile) that reacts when enclosed under air at room temperature and pressure (RTP), whereas no particular peak is observed when enclosed under argon. Finally, a test is conducted under 5 bar of oxygen, and the reaction initially observed under air exhibits higher energy release and heat flow rate [Baati, 2011]. This experiment allows identifying this reaction as an oxidation. Nonetheless, under air, the phenomena is so limited that it is practically negligible.

This discussion shows that when the crucibles are enclosed under air, oxidation cannot be excluded as a possible reaction of the sample, nevertheless it is rather limited and presents only small peaks that can be neglected. Therefore, the argon atmosphere is preferable, but the experiments conducted under air are considered as practically equivalent and this parameter will not be regarded as a distinctive parameter for further discussion or manipulations.

---

[2] Novartis AG Safety Laboratory made available for this project their database of thermal analysis of commercially available products which included several hundred entries

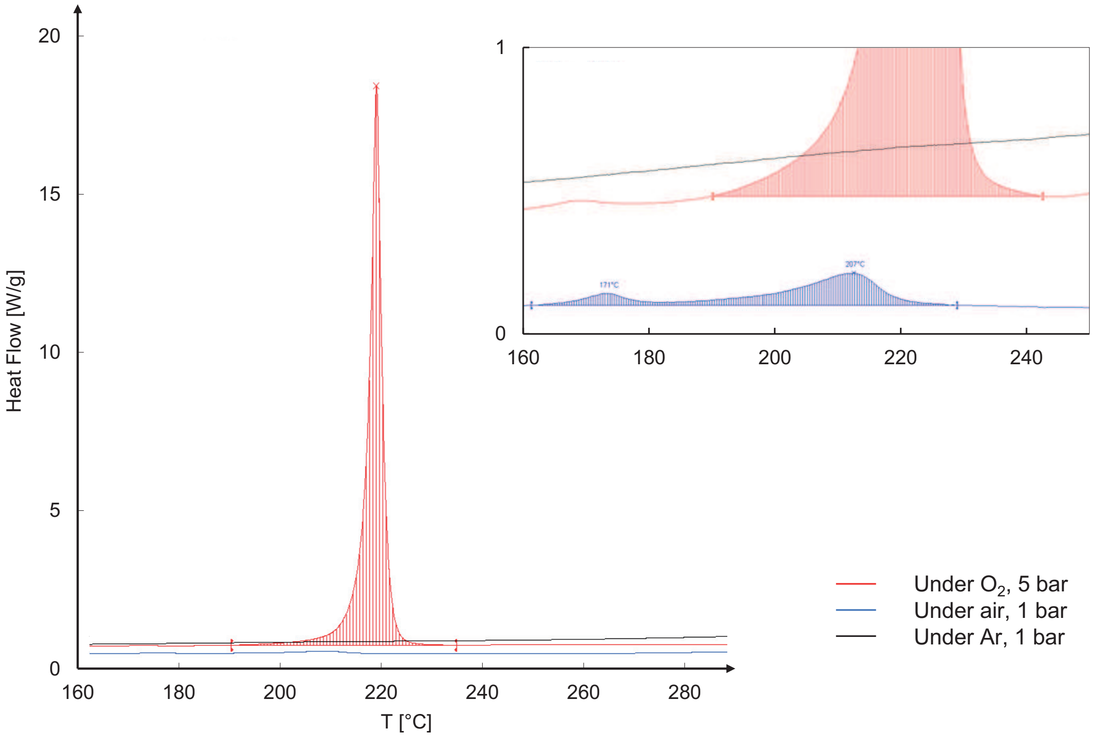


Figure 4.9 – Effect of Enclosing Atmosphere on DSC
Inset shows an enlarged view

Similarly, the sample weight ranges differ from the experimental set and the acquired set. Figure 4.10 shows a compound (i.e. nonanenitrile) tested twice, once with a sample weight of $m = 4.38$ mg and once with $m = 11.0$ mg [Baati, 2011]. As the heat flow rate is normalized by sample weight, it appears that the peak is smaller for higher sample weights. This is due to the fact that higher sample loads imply higher filling of the crucibles, hence the heat exchange area is no longer the crucible base solely, but also at crucibles sides. Therefore, the measurement is biased and potentially not all heat flow from the sample is detected at the measuring points. Empirically, it has been determined that the optimal weight range would be between 2 mg to 5 mg.

Finally, the DSC database acquired from the industry was initially on hard support. In order to use the thermograms, several treatments were necessary to digitize the information and recover the numerical values. This treatment (scanning, image processing to recover curves and axis, scale conversion, and finally digitization) may have introduced errors to the actual data.

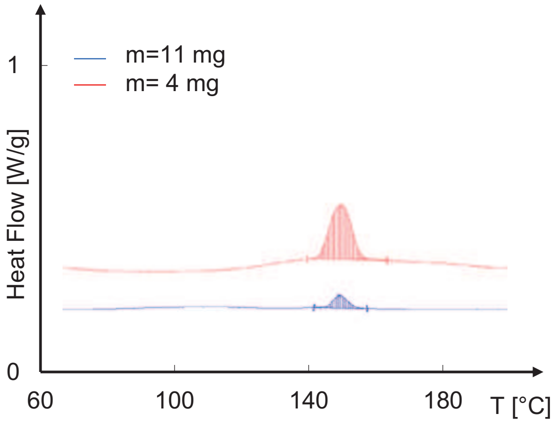


Figure 4.10 – Effect of Sample Weight on DSC

However, in view of the important number of thermograms available, it was favored to optimize the data treatment and to minimize error introduction rather than experimentally reproduced the tests. Besides, a selection was conducted to discard all thermograms that could not be integrated into the database due to various reasons: all samples that were mixtures rather than pure compounds, as well as polymers as their structure cannot be represented for the modeling phase, or inert (no observable thermal behavior of interest) were discarded. It is important to note that not all stable compounds were discarded; in order to have an overview of all types of thermal behavior, some were kept for the study, but most were removed due to their high abundance.

Then, all DSC data from the three collections (nitroaromatic, miscellaneous set for experimental error evaluation and large set from industry) were merged into a unique database that serves for the modeling work presented in Chapter 6. This resulted in a dataset of 414 DSC thermograms.

All molecular structures were created and processed with the Codessa Pro Software [Petrukhin et al., 2001] in order to generate the molecular descriptors necessary for the modeling by QSPR and by ICAS software [Gani, 1999] to generate the Marrero-Gani groups necessary for the modeling by GCM.

4.3.2 Thermogram Parametrization

As previously discussed, the information contained in DSC curves exceeds onset temperature and enthalpy of decomposition reaction. Therefore, here, the DSC curves are abstracted to few parameters and later reconstructed to recover most of the initial information. In principle, a model-fitting method to a known curve with a limited numbers of parameters may be applied and the parameters determined by identification.

In chromatography, the graphical results obtained are comparable to DSC measurements. The detection of products at the exit of the column as function of time presents a sequence of peaks that have to be analyzed to determine the quantities and retention times on the column. By analogy, in DSC the quantities to be evaluated are the reaction energies, the retention times are the temperature of occurrence and the signal is the heat flow rate as function of temperature instead of detected concentration in function of time.

After studying the signal processing methods used in chromatography, two fitting-methods seem reasonably appropriate for our purpose. The DSC curve could be fitted either to a Gaussian model (Equation 4.12) or a Fraser-Suzuki model (Equation 4.13):

$$f(T) = \Phi_{max} \cdot exp\left(\frac{-(T - T_{max})^2}{2\sigma^2}\right) \tag{4.12}$$

$$f(T) = \Phi_{max} \cdot exp\left[-\frac{1}{2a^2} \cdot ln^2\left(1 + \frac{a(T - T_{max})}{\sigma}\right)\right] \tag{4.13}$$

where Φ_{max} is the maximum heat flow rate measured, $\mathrm{T}_{\mathrm{max}}$ the temperature at which the maximum is measured, σ the peak standard deviation, and a in the Fraser-Suzuki model is an asymmetric factor [Felinger, 1998].

In DSC, the peaks are highly comparable to Gaussian curves, and this model could well describe most cases. However, in the cases where the kinetics of the reaction are of zeroth-order for instance, the ascending part of peak would be Gaussian-like, while the descending part would be rather abrupt, due to the entire consumption of the reactant and the immediate interruption of the reaction. This would exhibit a highly asymmetric peak, and would also be the case in other complex reactions mechanisms or non-integral orders. Therefore, the Fraser-Suzuki model would be necessary for a unified description of all cases that could be encountered.

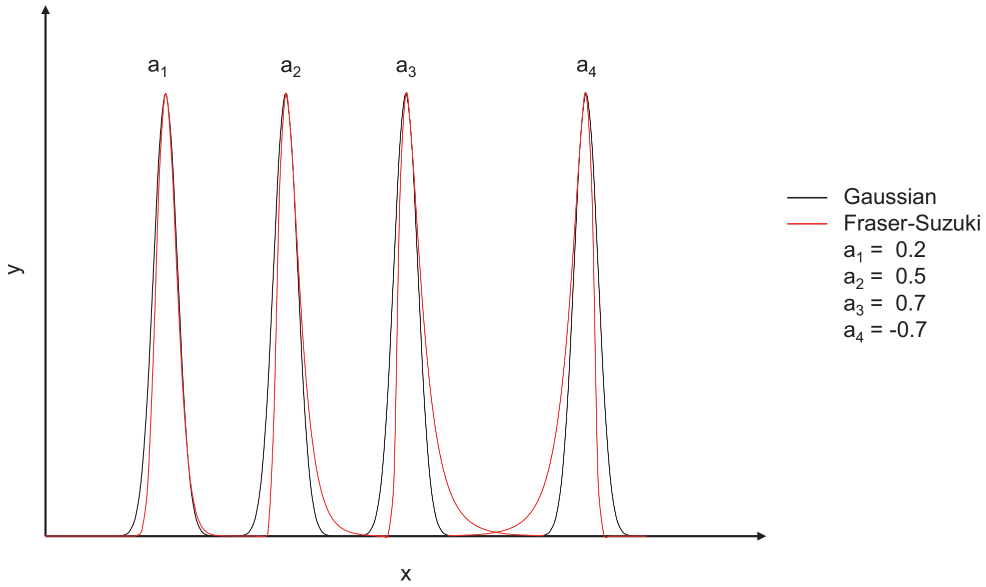


Figure 4.11 – Comparison of Gaussian and Fraser-Suzuki Models

Figure 4.11 shows how the Fraser-Suzuki model compares relative the equivalent Gaussian model, when the asymmetric factor a is varied.

4.3.3 Property Extraction and Data Treatment

In order to properly fit the DSC curves with Fraser-Suzuki, the AKTS software was employed [AKTS SA, 2000]. Not only does it include a fitting mode to identify the Fraser-Suzuki parameters, it is specifically designed to treat DSC curves, hence the process of baseline correction, the conversion of the curve from time scale to temperature scale or from heat flow to weight normalized heat flows and the integration of energies are also enabled, given that the sample weight and the scan rate are fed to the program.

The DSC curves are abstracted to five key properties, i.e. the peak maximum height, the peak maximum temperature position, the width, and an asymmetric factor and the area under the curve, as shown in Figure 4.12. These properties will later be modeled and estimated individually (see Chapter 6). When the simulated values are computed, they are reassembled into a Fraser-Suzuki equation, and the simulation of the entire DSC curve can be recovered thanks to this back-processing.

The area under the curve, which is directly linked to the reaction enthalpy, could be estimated from the curve simulations rather than modeled as a stand-alone parameter. This would represent one less variable to model, and would be a significant reduction in time and computation. However, in order to avoid potential error propagation from mis-estimations of the other parameters onto the area, it was chosen to be studied apart here.

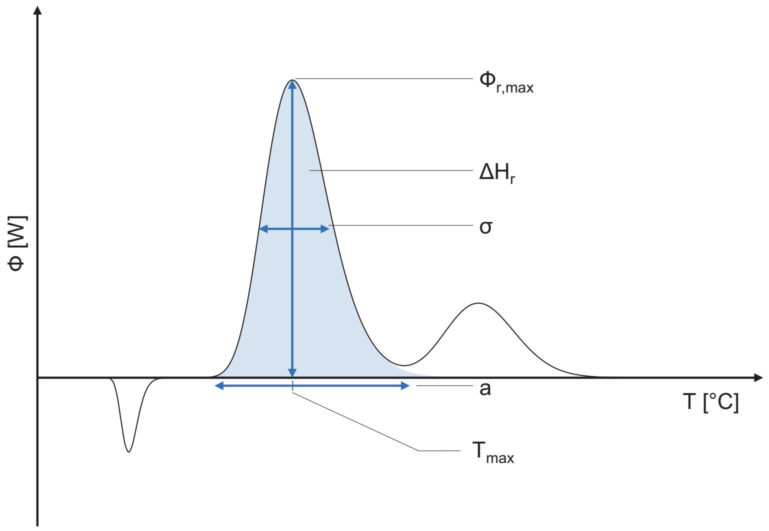


Figure 4.12 – Extraction of DSC Key Properties

For some thermograms, depending on the thermal behavior exhibited, there could be several thermal events to analyze. As the focus is on the decomposition reactions, endothermic peaks are disregarded as they are most often related to the melting of the compound if it is in solid phase at the beginning of the experiment. Moreover, if several reaction peaks are observed, only the "main peak" is held for the study. However, the assessment of which peak is considered principal could either be based on energy release or temperature of occurrence.

This procedure was not automatized and a case by case evaluation was performed to determine whether the main peak is the peak with highest energy release or the peak that appears at lower temperatures by relative comparison of the multiple peaks.

The data collection gathered as explained in Section 4.3.1 is treated with the present procedure. Finally, the DSC database has the following composition:

- more than 400 structurally diverse and thermally reactive chemicals
- DSC thermograms recorded with a heat-flow apparatus at $\beta = 4\,\mathrm{K\,min^{-1}}$
- the numerical descriptors for QSPR modeling
- the Marrero-Gani groups for GCM modeling
- 5 key properties extracted from each thermogram's main peak.

4.4 Conclusion

In this chapter, the general principles of Differential Scanning Calorimetry were reviewed. Then, the two types of apparatus, power compensation and heat flow DSC, and their particular functioning principles were detailed.

As the obtained results from both methods are sensibly the same, the elements to analyze from a resulting thermogram were discussed independently from the measurement method. Section 4.2 highlights the various parameters that can be identified or computed from DSC experiments, as for instance the heat capacities, the melting point, which can be used for compound identification, and the possible reactions the sample may undergo.

Regarding the reactions, DSC can expose the occurrence temperature and the energy release. Moreover the kinetic behavior can also be determined from DSC experiments. For this purpose, some of the major techniques were mentioned and briefly described.

Then, the data that will serve for this study in particular were gathered from three different sets: two sets were experimentally recorded (40 DSC) and another larger set acquired from a collaboration (≈400 DSC). The major differences between the data sets such as the weight range or the enclosing atmosphere were discussed and their effect assessed to ensure that these data were comparable and that they could be used within a unique database.

Finally, the DSC curves are abstracted to one main peak characterized by five key properties, i.e. the peak maximum height, the peak maximum temperature position, the width, and an asymmetric factor. These properties can be modeled and estimated individually and when estimates are reassembled into a Fraser-Suzuki equation, the simulation of the entire DSC curve can be recovered.

III

Applications

5 MIE Modeling

5.1 Literature Review

As noted in Chapter 3, several tests and characteristics of compounds are used to evaluate the probability of the explosion of a dust/air or a gas/air mixture: the auto-ignition temperature (AIT), the flammability or explosive limits (LEL/UEL) or the flash point (FP), to name a few. Besides the probability, severity needs to be estimated as well in order to grasp the related risks. The severity can be evaluated with the Explosion Constant ($\mathrm{K_{st}}$, or $\mathrm{K_G}$ for gases) and the Maximum Explosion Overpressure ($\mathrm{P_{max}}$) . Explosion constants represent the pressure rise a sample of the considered product would cause in a 1 m^3 spherical volume and the Maximum Explosion Overpressure corresponds to the difference between the pressure at ignition and the highest pressure recorded during the explosion.

With the rise of predictive models and their applications broadening to different fields, the fire and explosions characterization field was also investigated through some of these properties and several models were found in the literature, based either on GCM or QSPR. Table 5.1 summarizes some examples of models encountered in the literature for explosive properties. This table shows for each model which one of the molecular structure based methods was used, GCM or QSPR, mentioning if the molecular structure was the only input or if the model is based on other physical or chemical properties (in column "Add. Param.?"). It appears in Table 5.1 that several explosive properties were successfully modeled, however no MIE models were found.

The most similar property modeled is probably the electric spark sensitivity $\mathrm{E_{ES}}$, which was investigated in several studies. Zeman et al. [2006] define the electric spark sensitivity as the "electrostatic discharge energy required for 50 % initiation probability" and present experimental measurements for several detonating secondary explosives, mainly polynitro compounds, conducted on two laboratory-made instruments. The experimental procedure exposes the tested sample to the electric spark as small sample in a cylinder of 5 mm height and 5 mm diameter rather than dispersed. Works by Zeman et al. [2006] served as the basis for several studies that correlate $\mathrm{E_{ES}}$ to molecular structure using mostly their experimental data.

Table 5.1 – Comparative Summary of Literature Models for Explosive Characteristics

Reference	Method	Model Type	Dataset	Performance	Add. Param.?
Flash Point					
Albahri [2003]	GCM	polynomial	300	$R^2 = 0.99$	
Katritzky et al. [2007]	QSPR	linear	758	$R^2 = 0.92$	ΔH_f
Valenzuela et al. [2011]	GCM	linear	48	$AAD < 5K$	Const. and ΔH_{vap}
Rowley and Wilding [2010]	GCM	polynomial	+1000	$ARD < 10\%$	
Pan et al. [2010]	GCM	linear	314	$R^2 = 0.98$	
Keshavarz and Ghanbarzadeh [2011]	GCM	linear	173	$R^2 = 0.97$	
Auto-Ignition Temperature					
Egolf and Jurs [1992]	QSPR	linear	312	$0.94 < R^2 < 0.98$	
Suzuki [1994]	QSPR	linear	250	$R^2 = 0.91$	T_b, T_{cr} and p_{cr}
Mitchell and Jurs [1997]	QSPR	linear/ANN	327	$0.68 < R^2 < 0.87$	
Albahri [2003]	GCM	polynomial	500	$R^2 = 0.92$	
Pan et al. [2009]	QSPR	linear	446	$0.85 < R^2 < 0.89$	
Flammability Limits					
Albahri [2003]	GCM	linear	475	$R^2 = 0.93$	
Gharagheizi [2009]	QSPR	linear	865	$R^2 = 0.92$	
Pan et al. [2009]	QSPR	linear/ANN		$R^2 = 0.79$	
Lazzús [2011]	GCM	neural nets	418	$R^2 = 0.98$	
Bagheri et al. [2012]	QSPR	linear	+1500	$R^2 = 0.91$	
Explosion Constant and Maximum Explosion Overpressure					
Reyes et al. [2011]	QPSR	linear	35	$0.91 < R^2 < 0.96$	d_p
Electric Spark Sensitivity					
Keshavarz et al. [2009a]	GCM	linear	21	$ARD = 21\%$	
Fayet et al. [2010]	QSPR	linear	26	$R^2 = 0.90$	
Zhi et al. [2010]	QSPR	linear	30	$R^2 = 0.97$	
Wang et al. [2011]	QSPR	GFA/ linear	39	$R^2 = 0.92$	

In 2009, Keshavarz et al. [2009a] proposed a model based on descriptors of the chemical composition and structure that is able to estimate the value of the detonation velocity at maximum nominal density D’, which is directly correlated to $\mathrm{E_{ES}}$ according to their work. The descriptors used as parameters were the following: the number of constitutional atoms, $\mathrm{C_aH_bN_cO_d}$ and $\mathrm{n_{NR}}$ the number of nitrogen double bonds, $\mathrm{—N{=}N}$ and $\mathrm{n_N}$ the number of nitro groups.

The reported models were:

$$D' = 7.68 - 0.198a - 0.111b + 0.294c + 0.0742d - 0.635n_{NR} - 0.735n_N \tag{5.1}$$

$$E_{ES} = -0.4326D'^2 + 37.21 \tag{5.2}$$

The authors did not report an evaluation in terms of determination coefficient or relative errors. However, they presented the experimental and predicted values of their dataset, thus, the average relative deviation could be recomputed to evaluate their results and found to be about 21%. A study from the French national institute for industrial environment and risks (INERIS), relying on the same initial study by Zeman et al. and classical QSPR methodology proposed the following model for the electric spark sensitivity of 26 nitroaromatic compounds:

$$E_{ES} = 29.6n_{single} + 63.3n_{C,max} + 168.4Q_{C,max} - 27.8V_{C,min} + 99.4 \tag{5.3}$$

where $\mathrm{n_{single}}$ is the relative number of single bonds, and $\mathrm{n_{C,max}}$, $\mathrm{Q_{C,min}}$ and $\mathrm{V_{C,min}}$ are respectively the maximum nucleophilic reactivity index, the minimum partial charge and the minimum valence for a carbon atom [Fayet et al., 2010]. This model also results in responses with average relative deviations of nearly 20%.

Later, Zhi et al. [2010] developed the following model:

$$E_{ES} = (-1)^{n_1}\omega_1 Q_{nitro} - n_1 n_2 \omega_2 E_{LUMO} + \omega_3 \tag{5.4}$$

with $\mathrm{n_1}$ the number of aromatic rings, $\mathrm{n_2}$ the number of substituents other than nitro groups, $\mathrm{Q_{nitro}}$ the minimal charge on a nitro group, and $\mathrm{E_{LUMO}}$ the energy of the lowest unoccupied molecular orbital. The correlation coefficient of this model is rather high $R^2 = 0.97$, but on the other hand the studied set of compounds is rather narrow (19 data for the training set, 2 of which have been excluded from the model, and 11 held for the testing set).

Finally, Wang et al. [2011] used an approach more similar to this of Fayet et al. and also obtained satisfactory results as well: their model, built from the study of 39 nitroarenes is composed of 8 parameters and gives $R^2 = 0.924$ and $R^2_{cv} = 0.873$.

Table 5.2 shows the experimental $\mathrm{E_{ES}}$ values of 1,3,-dihydrox-2,4,6-trinitrobenzene, and the predictions found by the three latest models [Fayet et al., 2010, Wang et al., 2011, Zhi et al., 2010]. This comparison highlights the fact despite the differences in the model development and the obtained equations, all three studies propose simulated values for $\mathrm{E_{ES}}$that are relatively similar and accurate. This points out the potential existence of several correlations that may be found, while the establishment of one model as the most appropriate may be challenging.

Table 5.2 – Comparison of $\mathrm{E_{ES}}$ from Literature Models

1,3,-dihydrox-2,4,6-trinitrobenzene				
	$\mathrm{E_{ES}}$ (exp) mJ	$\mathrm{E_{ES}}$ (sim) mJ	Dev mJ	RD %
Fayet et al. [2010]	12.3	11.4	-0.87	-7
Zhi et al. [2010]	12.3	9.7	-2.6	-21
Wang et al. [2011]	12.3	11.4	-0.91	-7

The motivation for simulations of MIE is driven by several elements. Firstly, it is an important and necessary safety parameter when considering the handling of energetic materials. Secondly, the experimental procedure detailed in Chapter 3 requires several repetitive steps: varying sample weights and spark energies, and iterating the ten consecutive trials. These repetitions accumulate the time and material costs. Besides, the results are given as threshold values or ranges, from a finite and discontinuous set of values. Finally, from the discussion above, and to our knowledge, there are no models to predict MIE from the molecular structure.

Therefore, the goal of this chapter is to develop and present predictive models for MIE values only, without any particular prior processing of experimental data, and without taking into account other influencing properties such as the temperature of auto-ignition, nor the minimum ignition temperature, the concentration relative to flammability limits or particle size. From this perspective, only the structural influence is analyzed, hence the differences between dusts, vapors or gas are virtually erased.

Firstly, the collected dataset presented in Tables A.1, A.2 and A.3 will be studied as a whole to produce global models, in order to test if one correlation can be built directly from molecular structure independently of all other considerations (physical state, dispersion, etc).

Secondly, local models will be developed for three subsets of the initial ensemble, mainly to separate the artificial merge we imposed between dusts, vapors and gases.

Finally, MIE values are measured in thresholds, hence the true MIE values are comprised between delimited ranges, and therefore, classification could be more appropriate than regression. Moreover, a classification method will also show whether or not there is a better separation than the physical state criteria to propose local models.

5.2 Resulting Models

5.2.1 Global Models

The modeling of the MIE data was conducted as follows. Once the values are acquired and 318 QSPR descriptors are generated as detailed in Chapter 3, the dataset is ready to be analyzed.

The two first steps serve for feature selection. The first one consists in developing a repetitive loop that allows to divide the dataset into a training and a validation set, with 90-10% proportions, then applying the General Linear method to assign coefficients to all parameters, and then evaluating the model on both training and validation set. The loop is repeated 100 times. The coefficients adjustments vary for every iteration depending on which data are on the training set, on the other hand the subsets of parameters for which non-null coefficients are determined does not vary significantly. This wrapper allows rapidly reducing the feature set size approximately by half.

The second step calls for a stepwise regression model. The parameters are included successively, only if they contribute to improving the model's performance. Here, the process is made iterative as well, and several models are produced in order to determine the necessary parameters to obtain a model's performance higher than $R^2 = 0.95$. The stepwise regression algorithm requires a threshold p-value of an F-statistical test, above which the descriptors are considered statistically irrelevant and are not included in the models. The iterative loop put in place here increases the p-value in order to develop different models including additional parameters, and achieving higher correlations to the targeted observation values.

It is as considering that the entire data set potentially contains 100% of the information available, and this estimates which subset minimizes data loss due to dimension reduction. Less than 100 descriptors are selected for the modeling phase.

Finally, after these two consecutive reductions of the descriptors space, the dataset is processed through a third and last loop. Once again the division into training and validation sets is performed to randomly take out 10% of the observations to later validate the model. The stepwise regression is applied again, but this time the p-value threshold is fixed. Several training - validation combinations are run, and one model is built for each of these random combinations. The 10 best ones in terms of fitting of the validation set are selected to be analyzed more closely. Figure 5.1 shows the responses of one particular model which is

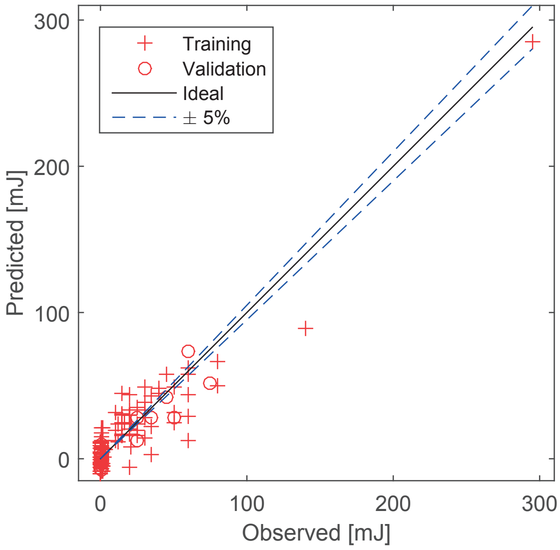


Figure 5.1 – Global Model Responses for MIE: Predicted vs Observed Values

representative of the obtained models.

In Figure 5.1, and on the other similar graphics, the model's estimations and predictions are plotted against the actual values of the observed compounds. Data in the lower right half of the chart are underestimated, while data in the upper right corner are over-estimated by the model. Ideally, the model should present good predictions of the observed values and data lie close to the ideal fit represented by the black line and indicated by the blue dashed lines that delimit a ± 5% range.

To recover the entire model, the parameters and their related coefficients are reported in Tables A.5 and the responses in Appendix A.8, in Appendix A. The global model presented here is evaluated and its performance is summarized in Table 5.3. This evaluation shows overall a good performance of the model, that reflects well what can be observed on Figure 5.1. Both training set and validation set are well estimated, and the determination coefficients R^2 and R^2_{cv} are relatively high. Nevertheless, average relative deviations (noted ARD_{Tr} and ARD_{Val} in the table for training and validation sets respectively) show that the estimates are not accurate and the mean errors are higher than 1000%. Indeed, on the Figure 5.1 it appears clearly that a large subset of the observations are in the extremely low range of MIE values. The relative errors occurring when predicting these very small values are rather large, which explains the obtained ARD. On the other hand, the average absolute errors (AAE) for both training and validation set are of about 8 mJ to 9 mJ, which would be a reasonable margin for data with high MIE but meaningless for observations with $MIE < 1mJ$. Therefore this raises the question as to whether it is appropriate to study the dataset entirely when it is unevenly distributed across the value range as half the set is concentrated within a region that represents about 1% of the range of observations. In the following section, the dataset will be subdivided in order to be studied with respect to these observations and local models will be developed and discussed.

Table 5.3 – Global Model Evaluation Summary

Evaluation	Global Model
R^2	0.860
R^2_{cv}	0.850
# Parameters	16
Training Set	117
Validation Set	15
ARD_{Tr}	1327 %
ARD_{Val}	1210%
AAE_{Tr}	8.7 mJ
AAE_{Val}	8.1 mJ

5.2.2 Local Models

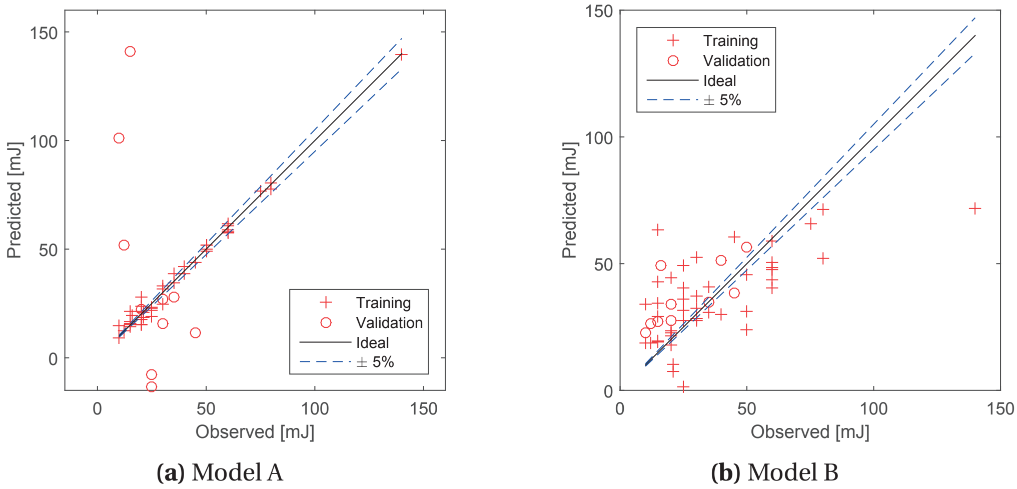
Dusts

As mentioned previously, the dataset comprises disparate values over the represented range of MIE values. More than half the values are lower than MIE=10 mJ, and all dusts observations present an MIE higher than 10 mJ. Thus, the first studied subset investigated is the dusts as it seems the most clearly defined subset.

The data analysis conducted is the same than previously detailed for the development of the global model. The results are illustrated with the three examples, which will be referred to as model A, B and C.

Models A and B are typical illustrations of the main difficulty encountered when building and selecting models. As one can see in table 5.4, model A reaches high correlation efficiency with the training set as $R^2 = 0.986$. However, the model fails to perform as well with the validation set as the $R^2_{cv} = 0.365$ and ARD_{Val} is higher than 250%. This is supported by the Figure 5.2 that shows clearly how all training data lie within the region close to the ideal fit, whereas the validation data are rather poorly predicted, and are dispersed away from fidelity. This is a strong indication of an over-fitting issue: the model is adjusted to the training set so well that it becomes unadapted for data outside the set, therefore the bad estimations for the validation set.

On the other hand, model B presents a better visual aspect as the validation data are more closely gathered around the ideal fit. Nonetheless, the evaluation reveals a rather weak model that does not properly fit even the training set. Another important parameter to take into account in this comparison is the number of parameters included in model A and B. They are composed respectively of 28 and 7 structural descriptors.



(a) Model A **(b)** Model B

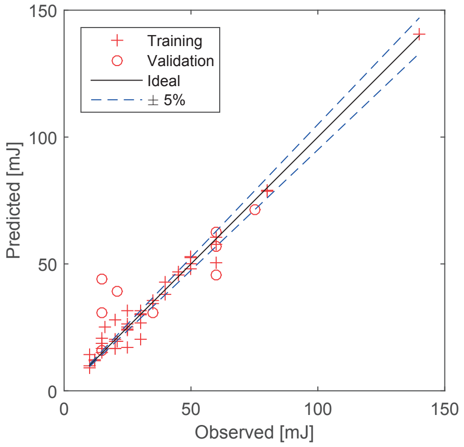
Figure 5.2 – Local Models Responses for MIE of Dusts

Table 5.4 – Comparative Evaluation Summary of Dusts Models for MIE

Evaluation	MIE: Local Models for Dusts	
	Model A	Model B
R^2	0.986	0.398
R^2_{cv}	0.365	0.502
# Parameters	28	7
Training Set	43	43
Validation Set	10	10
ARD_{Tr}	9.8%	53.6 %
ARD_{Val}	253%	70.3 %

Ockham's razor principle [Bishop, 2006, Witten et al., 2011] applies in these cases: in general, the fewer parameters, the better the model, both in terms of simplicity and generalization. However, the inclusion of more parameters usually permits to refine fitting and obtain higher correlations; therefore the ideal number of parameters within the model should be a compromise between the complexity, performance and generalization of the model. A rule of thumb also advices not to adjust more than 1 parameter per five observation entries. In this case where about 50 observations are studied, an efficient model would describe the data with 10 parameters. Unfortunately, no ideal model could be obtained to answer all these requirements, nonetheless, model C hereafter could be considered as a good compromise to describe MIE of the present subset.

Model C evaluation is presented in Figure 5.3 with the graphical representation of the model's reponses. Overall, it is highly similar to model A concerning the training set, where both



Evaluation	Model C
R^2	0.974
R^2_{cv}	0.543
# Parameters	27
Training Set	43
Validation Set	10
ARD_{Tr}	11.2 %
ARD_{Val}	60.4%

Figure 5.3 – Best Local Model for MIE of Dusts

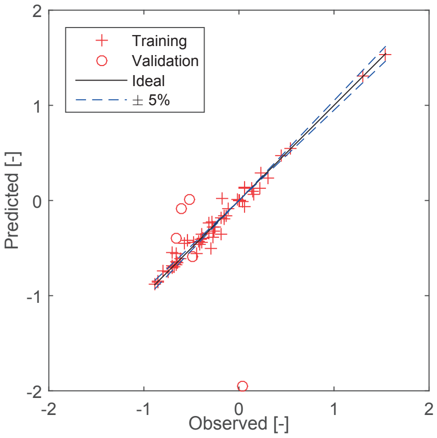
achieved good fitting results. Yet, model C overpasses model A when it comes to validation. From the validation data, an outlier is predicted with a negative MIE. If a cut-off value were to correct negative responses and replace them by zeros, the average deviations of the model would be diminished by 6% points. The descriptors involved in this model are given in Appendix A, Table A.6.

Liquids

As noted previously, about 50% of all observations fall in the region of $0 < MIE < 1$ mJ, and therefore the dataset was divided in order to study local subsets separately. After taking out the dusts and gases from the dataset, the MIE data are comprised in the range 0 mJ to 3.5 mJ with 75% exhibiting ignition energies lower than 1 mJ (with the exception of trichloroethylene (#78) with an $MIE = 295$ mJ, which was held out of the set due to very low sensitivity compared to the other observations).

In order to increase the difference between the values comprised in the 0 mJ to 1 mJ region and ease the modeling procedure, the MIE values are transformed into their corresponding logarithm. After this operation, one observation was discarded: 2,2,3-trimethylbutane (#22) which exhibits an $MIE = 1$ mJ, which became 0 in logarithmic scale, and dirturbed model evaluations due to virtually infinite errors.

Besides this pretreatment, models were built following the same procedure detailed in Section 5.2.1. In the third step of the procedure, it is important to fix the threshold p-value to determine which parameters to add or remove from the model. Here, this value was particularly delicate to adjust. When set to p_{enter}=0.250 no relevant models were obtained. In general, about 3 to 10 parameters were combined, however, the training set was not well described and for all of them $R^2 < 0.5$. If the threshold p-value was fixed to p_{enter}=0.255, the algorithm could not rank the most relevant parameters properly, and this produced over-parametrized models with up



Evaluation	Liquids
R^2	0.954
R^2_{cv}	0.871
# Parameters	35
Training Set	54
Validation Set	5
ARD_{Tr}	29.7 %
ARD_{Val}	1238.5%
- outlier	61.7 %

Figure 5.4 – Local Model for MIE of Liquids

to 60 parameters.

The following example presented in Figure 5.4 is one of the best correlations obtained. The correlations to the training and validation sets appear to be correct, and if one removes the outlier from the evaluation, the relative deviations are relatively low. The removed outlier from the evaluation of ARD is n-propyl chloride (#66) with $MIE = 1.08\,\text{mJ}$ and $log(MIE) = 0.0034$ and therefore the relative error is very large, resembling the issue with observation #22. Despite all efforts to limit the descriptors within the model, this one includes 35 parameters, which makes it over-parametrized. Besides, this model is built from the logarithmic values of the MIE, and when transformed back to recover the actual values, errors are also scaled up, resulting once again in $ARD_{Val} > 100\%$. Overall, this model cannot be considered as satisfactory from any criteria used for evaluation.

Liquids and Gases

Following the modeling of dusts and then liquids, gases were also studied separately, however, all models developed presented either very low correlations to training and validation sets, or extreme cases of over-fitting as shown in Figure 5.5: the models were adjusting too well to the training set, and were then unable to generalize and apply properly to the validation set. In this case, the correlation to the training set is so high that the measured determination coefficient is $R^2 = 0.997$ while the deviations on the validation set reach $ARD_{Val} = 900\,\%$.

As the 19 observations in gas phase comprised in the dataset follow a similar distribution to those in liquid phase, they were studied in a combined subset and an example of the obtained results is presented in Figure 5.6. Results are given for the logarithm of the MIE values. This model could be considered better than the models presented previously for liquids only, as it is based on a larger set and comprises fewer parameters. This potential to generalize better is confirmed by the validation set evaluation : R^2_{cv} is higher in this case and the deviations

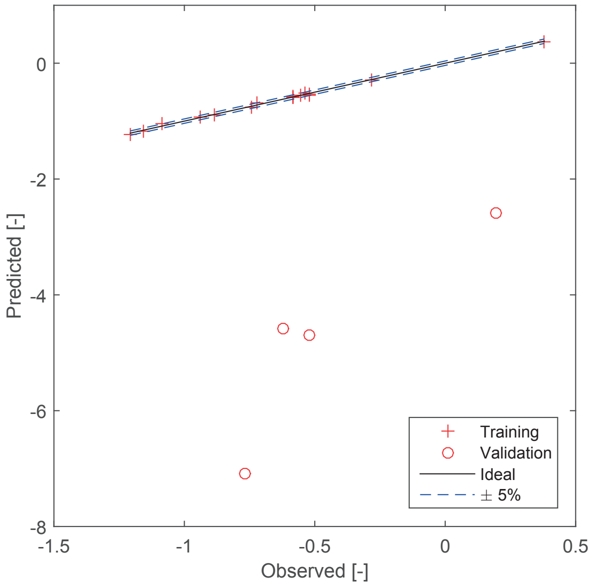
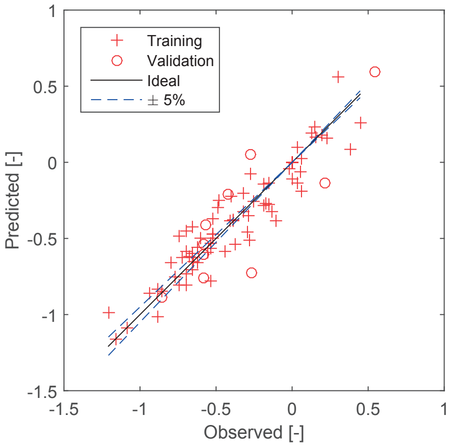


Figure 5.5 – Local Models for MIE of Gases: Overfitting Example - both axes are logarithmic values [-]

ARD_{Val} are lower. Overall, the predictive power of this model is arguably improved from the previous one. Moreover, when recovering the MIE values by back transformation of the logarithmic values, the impacted errors are less important than previously and the relative deviations are actually decreased to $ARD_{Tr} = 22\%$ and $ARD_{Val} = 41\%$.



Evaluation	log(MIE)
R^2	0.873
R^2_{cv}	0.711
# Parameters	24
Training Set	68
Validation Set	10
ARD_{Tr}	48.5 %
ARD_{Val}	58.7%

Figure 5.6 – Local Model for MIE of Liquids and Gases

5.2.3 Sensitivity Classification

Following the partition in various subsets according to chemicals physical state, data analysis from the previous sections reflected the known fact that the vapors and liquids are more sensitive than dusts, and present MIE values in the lowest range of the entire studied span. It also highlighted that besides the differences in sensitivity, the different reactions mechanisms and explosive behaviors could be anchored in structural differences as no global model could be successfully built while higher results are achieved after data separation.

A more pragmatic approach would encourage performing a categorical classification of MIE values. Indeed, if one applies a predictive model to some compounds of unknown MIE, and the result points out a possibly sensitive or highly sensitive compound, one would not take the risk to rely only on the predictions, and would use them to perform a guided testing phase with expectations regarding the energy levels at which ignition is most probable. Thus, it seems quite appropriate to investigate the possibility to categorize the data rather than predicting exact values - with mitigate accuracies.

Accordingly, categories of sensitivity to ignition were to be delimited for the dataset.

In the Ignition Handbook of Babrauskas [Babrauskas, 2003], a British Standard classification is reported as in Table 5.5 with the corresponding recommended precautions (note: the remark concerning class 3 is not a recommendation, but was stated as it is in the original reference).

Table 5.5 – Ignition Sensitivity Categories of Powder Suspensions

MIE mJ	Class	Recommendations
1	1. Extreme sensitivity	The presence of explosible mixture should be avoided. Handling operations should minimize possibility of powder dispersion. All possible steps should be taken to ease the dissipation of charge and to avoid charge operations.
10	2. High sensitivity	Consider restrictions on the use of high resistivity non-conductors when ignition energy is at or below this level
25	3. Medium sensitivity	The majority of ignition incidents occur when ignition energy is at or below this level
100	4. Low sensitivity	Consider earthing personnel when ignition energy is at or below this level
500	5. Very low sensitivity	Earth plant when ignition energy is at or below this level

The dataset was then distributed within these categories: 64 compounds having $MIE \leq 1\,\mathrm{mJ}$ were assigned to class 1, 18 with $1 < MIE \leq 10\,\mathrm{mJ}$ were assigned to class 2 and 25 with

$10 < MIE \leq 25\,\mathrm{mJ}$ were assigned to class 3. Finally, the remaining 25 with $MIE > 25\,\mathrm{mJ}$ were assigned to class 4 and class 5 is not represented here.

A classification tree algorithm was applied in order to determine the structural descriptors that could allow to recover this partition into the different MIE categories. Several series of cross-validations were conducted to determine the ideal tree construction. At each iteration, 15 observations were hold out for validation, whereas the 117 others were used to train the tree. In order to prevent the algorithm from developing complex trees with narrow categories and numerous branches, the leaf size parameter was fixed, so that the tree construction would always ensure that subcategories contain a minimum number of observations.

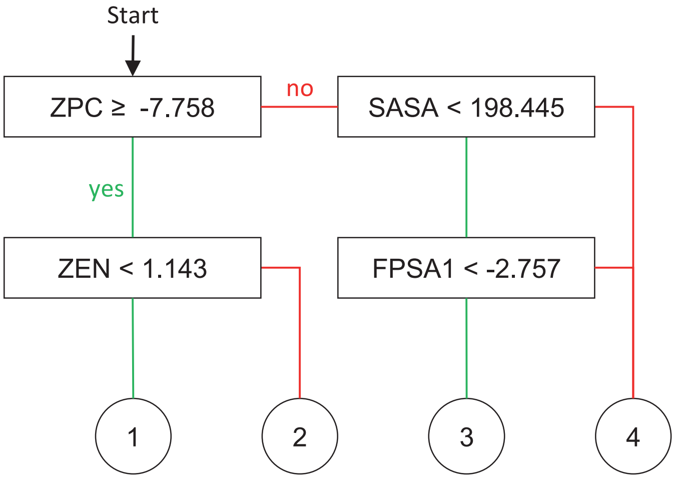


Figure 5.7 – MIE Classification Tree

An example of obtained tree is shown in Figure 5.7. This tree, comprises 4 nodes, i.e only 4 criteria to classify the data into the 4 classes.

The splits at each node are explicitly presented in Appendix A, Figure A.1. One can see that at the first tree node, the descriptor ZPC serves to delimit the regions of classes 1 and 2 on one hand, and classes 3 and 4 on the other. The second node at which classes 1 and 2 are separated lies on descriptor ZEN. Finally, the tree determines class 3 and 4 depending on SASA and FPSA1 values. The descriptors are given in Table 5.6.

All these parameters were involved in the models developed in the previous sections as for instance Model C constructed for the dusts. For the tree construction, the constraint was on the minimal population per leaf, which was set to be at least 8. This was determined through numerical simulations to identify the values that realize a satisfactory compromise between good separation and tree complexity. Indeed, this constraint prevents the tree construction from developing numerous branches and giving rise to a complex and intricate category system.

Table 5.6 – Classification Tree Parameters

Splits	Threshold	Parameters	Name
1-2 / 3-4	-7.75	ZPC	Zefirov's Partial Charges for atom #0000003(C)
1/2	-1.14	ZEN	Sanderson's atomic electronegativities for atom #0000007(O)
3-4/4	1.98×10^2	SASA	Solvent accessible surface for atom #0000008(C)
3/4	-2.76	FPSA1	Fractional PPSA (PPSA-1/TMSA) (Zefirov PC)

The main drawbacks from developing a simple classification are that some splits are faulty as can be seen in Figure A.1. For instance, at the first node, three observations belonging to class 2 are already misclassified and sorted with classes 3 and 4. The second node has a high failure ratio concerning class 2 as more data have been misclassified than not. Fortunately, they are assigned to a more critical class and it performs much better with class 1 as only 3 data fall on the wrong side of the separator. The third node separates well some of class 4 from class 3. Finally, the fourth node has a lower performance as it appropriately sorted out 36 out of 48 observations that reach this point.

Globally, the classification tree presents good results as its performances are 80% correct classification for the training set (94 out of 117), 60% for the validation set (9 out of 15) and overall 78% correct classification.

Some improvements would be possible with higher branching degree, as the inclusion of more parameters could allow to better refine the flaws noted above. Besides, it could be beneficial to weight the error function, in order to favor "conservative" errors. Indeed, it would be preferable to have a mis-classification into a higher sensitivity class than the opposite. All attempts to do so did not exhibit enhanced classification performances than the tree discussed here.

As mentioned earlier, it could be sufficient to bring the simulation to this point. Nevertheless, in-class modeling was conducted in order to assess if the classification brings an improvement to the models. For this purpose, stepwise regressions were conducted on the populations of each class, and the results are gathered in Table 5.7. The models parameters are gathered in Appendix A, Table A.7.

Considering the correlation coefficients obtained for the models built after classification, the results do not show significant improvement. On the contrary, it would seem that classes 1 and 3 present lower correlations than the previously obtained ones. Moreover, the correlations of the models' response to validation sets for all classes are rather low. Nonetheless, the major improvements lie in the number of parameters required that are kept to between 5 and 10 parameters, and the precision of the fittings and the predictions that are significantly enhanced. While the previous models presented deviations ranging between 40 and 60% for the training and higher than 100% for the validation set, here the models present average

$ARD_{Tr} = 15\%$ and $ARD_{Val} = 34\%$.

Class 1 was studied after logarithmic transformation as it gathers the lowest MIE values corresponding to the most sensitive observations. When recalculating the MIE values from their simulated logarithmic correspondents, the deviations are modified, simply due to the fact that the MIE values are in absolute smaller than their logarithms, hence the smaller relative errors for training set and larger for validation set, respectively.

Table 5.7 – MIE Post-Classification Models

	Class 1		Class 2	Class 3	Class 4
	LOG	MIE	MIE	MIE	MIE
R^2	0.669	0.589	0.998	0.734	0.938
R^2_{cv}	0.032	0.008	0.037	0.033	0.468
# Parameters	10		5	5	8
Training Set	59		13	20	20
Validation Set	5		5	5	5
ARD_{Tr}	40	29	8	9	10
ARD_{Val}	22	42	37	27	29

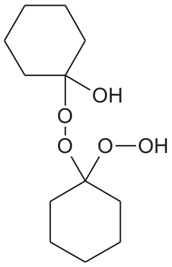
To illustrate the model application, simulation of the MIE value of cyclohexanone peroxide (#98) with the model of Class 3 is detailed here. Table 5.8 represents the structure of this compound and the descriptors included in the corresponding model.

If the descriptors values are replaced in the MIE equation for Class 3, the MIE of cyclohexanone peroxide may be computed as:

$$MIE = 42.5 + 27.1 \cdot HDCA - 2 + 0.472 \cdot CPSA + 2.17E - 2 \cdot SASA_N - 0.987 \cdot^2 IC - 883 \cdot I_C \tag{5.5}$$

and the result obtained is $MIE_{(\#98)} = 22.6\,\text{mJ}$, which represents an accurate estimation of the actual value $MIE_{(\#98)} = 21\,\text{mJ}$, with a relative deviation of $RD = 8\,\%$. This result shows a great enhancement compared to the dusts model, which predicts $MIE_{(\#98)} = 30.6\,\text{mJ}$.

Table 5.8 – Cyclohexanone Peroxide Structure

Structure	Value	Descriptors	Name
	3.50×10^{-2}	$HDCA-2$	HA dependent HDCA-2 (Zefirov PC)
	9.14×10^{-3}	I_C	Moments of inertia C
	-8.77×10^{1}	$SASA_N$	Solvent accessible surface for atom #0000001(N)
	1.46×10^{1}	^2IC	Average Structural Information content (order 2)
	7.46	$CPSA$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000012(H)

5.3 Model Interpretation

The physical interpretation of models involving approximately 30 structural descriptors is almost impossible and would not bring coherent sense. Nonetheless, it is possible to extract some elements of discussion. For instance with model C developed with dusts MIE data, it is possible to determine which of the 27 parameters of the model bring the most significant contribution to the overall equation either by determining those with the highest correlations to the responses or those with the highest weights in the model [Guha, 2008, Polishchuk et al., 2013]. Another manner to point out the highest contributors to the models is to permute their values and see how this influences the predictive power [Polishchuk et al., 2013]. This is also a manner to check for chance correlations. Indeed, if randomized data give similar correlations than the initial model, it simply invalidates it.

These three manipulations were performed. It was possible to determine that among the 27 descriptors in model C, only 2 do not bring a significant contribution to the model: after permutation of the values of descriptors #23 (DPSA3) and #27 (Bond orders for $\mathrm{N-O}$, see Table A.6, Appendix A), it appears that the model overall predictive power (for both training and validation data), initially $R^2 = 0.885$, was only slightly affected by this, as it decreased to $R^2 = 0.764$ and $R^2 = 0.854$ respectively. While for all other parameters, permutations brought the correlation coefficient to $0.001 \le R^2 \le 0.429$. The highest loss of information can be directly identified with the highest correlations drops, and this lead to marking highest contributing descriptors. Overall, the three methods revealed some descriptors with higher importance:

- #3: ^1BIC - Bonding Information content (order 1)
- #7: $LOGZEN_C$ - Natural logarithm of Sanderson's atomic electronegativities for C-atom
- #10: ZPC_C - Zefirov's Partial Charges for C- atom
- #14: $ETS_{B,C}$ - Electrotopological state of atom for C- atom
- #17, #19 and #22: $SASA_x$ - Solvent accessible surface for several atoms and

- #18, #20: $CPSA_x$ Charge density on solvent accessible surface (Zefirov's PC) for several atoms

The electrotopological state of a given atom "encodes the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule" [Hall and Kier, 1995]. It is thus a function of the electronic, topological and valence state of the atoms and is highly dependent of its electronegativity. The partial charges are also dependent on the electronegativities which will determine the charge distribution among the structure and the surface area. The solvent accessible surface areas are derived from the van der Waals areas of atoms and geometrical configuration to take into account the parts that are "buried" inside the molecule's saddles or angles. Finally, the bonding information content is a topological index that captures the molecule branching degree. All together, these descriptors are mostly based on the atoms electronegativties on one hand, and the molecule overall geometry on the other hand. The electronegativty-related descriptors are fairly correlated to each other and one could imagine that their contributions to the model are redundant. However, when one of them is missing the predictive ability of the model is significantly affected. Their contributions are somehow synergetic and inter-correlated; thus they make further interpretations much more complex.

It is important to remark once again that the node-descriptors in the classification tree are the same that we encountered in model C, namely ZEN (Sanderson's atomic electronegativity for O-atom), two derivations of ZPC (Zefirov's Partial Charges for C-atom and Fractional Positively charged surface area) and Solvent accessible surface for C-atom. This also confirms that these categories of information are not chance correlations, however they are insufficient to entirely describe the behavior towards ignition energies, and only permit classification. A selection of a handful of parameter fail to offer more accurate estimations.

5.4 Conclusion

For this section, MIE data of more than 130 molecules were gathered from several reference sources and treated with Codessa Pro software to derive the necessary numerical descriptors for QSPR modeling. In a first stage, all available data were studied as whole set without discrimination based on neither of physical state nor the MIE order of magnitude. This allowed to develop a global model, general for the entire ensemble. Unfortunately, this model, comprising 16 parameters, could only estimate the tendencies within the set as the general trend is graphically recovered (see Figure 5.1). The response values computed, though, were very weakly representative of the target values as the average deviation exceeded 1000%. This was accounted on the unbalanced data distribution among the evaluated range, as 50% of the observations are concentrated on 1% of the value span.

Following this, the second step consisted in the development of three local models for three subsets corresponding to the data partition in function of the physical state. Among these three correlations, the only one that was considered satisfactory is the model C built for the observations in solid state. This model relies on 27 parameters, which makes it rather complex, nonetheless it fits well to the training data and gives strong indications for most of the validation data.

As to compensate for the failed modeling of the gases and vapors sets, a fourth local set was developed for a combination of these two sets. This could be justified not only by the fact that they have similar behavior towards ignition [Eckhoff, 2003] but also by their similar distribution in the studied dataset. This model exhibited much better performance than these of gases and vapors separately, and even though it was built using the logarithmic transform of the MIE values, it did not fail to recover the values after the reverse treatment. The average deviations were about 20 to 40% which is still fairly large for predictions of a sensitive property.

Finally, the classification tree was developed to answer a more realistic approach and increase the usage potential. As we consider here a sensitive property, for which experimental measurements are very tedious and requires tens of trials, it could be more helpful to use predictive models to correctly assign a molecule of unknown MIE to a category rather than give a poorly reliable estimate value. The developed tree was reasonably simple and efficient: four nodes to define four categories of sensitivity, with overall 78% correct assignment. Moreover, the classification according to MIE values also improved considerably the modeling. Indeed, the correlations developed within classes were less complex and showed improved accuracy relatively to all previous ones.

The mitigate results obtained here probably reached their limitations. The initial hypothesis that MIE could be correlated to molecular structure only without consideration of state, concentration, or particle size regarding dusts, cannot be confirmed here. It is probably erroneous and should be rejected, or subjected to a another analysis with a more comprehensive dataset.

It could have been interesting to investigate modeling of MIE values as function of the structure and particle size for instance, in a comparable manner to the models proposed by Reyes et al. [2011] for the explosion constant K_{st} and maximum explosion overpressure P_{max}. This would also be in closer agreement with observations of Eckhoff [2003] or Bartknecht [1989] that the particle size is a highly influencing factor.

Unfortunately, the collected dataset does not include such supplementary information that could have been taken into account to develop structure-property-property models: indeed, the references reported the MIE data with mentions of the concentration being the most readily ignitable concentration or the dust particle size being comprised within a given range; however these incomplete data could not be used for quantitative models.

This project has two major positive outcomes:

- First of all the classification tree is an interesting tool: it cannot be translated into a rule of thumb as the descriptors require computation, yet, with the descriptors available it is a strong indication regarding a compound sensitivity to ignition. The classification itself represents a primary determination of the energy level at which to start the experimental procedure.

- The modeling procedure developed here, combining a classification model and regression models, allows to screen the data without prior knowledge nor supplementary data such as the concentration or particle size. Moreover, it allows developing local models more accurately than global models while it broadly applies to all data available here.

6 DSC Models

In this chapter, a literature review will point at the various GCM and QSPR models that were developed specifically for DSC derived data. This review allows identifying a suitable starting point in the study of Nitro compounds, as several studies were successful in predicting some of their properties. Nonetheless, with a broader dataset available, the investigated range is enlarged. Here, all developed models for DSC thermograms are presented and discussed.

6.1 Literature Review

As shown in Chapter 1, molecular structure based models focused on hazardous characteristics of chemicals since their early development stages. Toxicity but also thermal properties were among the first applications of GCM and QSPR.

6.1.1 Group Contribution Models of Thermal Properties

Following the development of Group Contribution Methods by Lydersen [1955] and Benson and Buss [1958], several researches proposed correlations between various hazardous characteristics of compounds and their chemical structure.

In 1968, Benson and co-workers broadened the group contribution framework for the estimation of several thermodynamic properties, and in 1974 the ASTM Chemical Thermodynamic and Energy Evaluation Program (referred to as CHETAH) was released [Seaton et al., 1974]. The program's calculations are based on the Benson groups contributions and give thermochemical data from chemical structure only, as long as the molecule can be described by Benson's system. Among the criteria that can be estimated, six serve to assess the thermal hazard related to the substance.

An updated version of the program was reviewed by Shanley and Melhem [1995]. The authors noted several deficiencies in the software hazard evaluation method and faulty results. The limitation to molecules that can be described in the Benson group-additivity system, the inability to distinguish isomers and the reliance on cut-off values were noted as the major drawbacks of the method. To improve the CHETAH software, they developed a procedure to identify the most stable possible decomposition products and calculate the corresponding reaction enthalpy [Melhem and Shanley, 1996].

Murphy et al. [2003] completed this procedure with a methodology that allows calculating the maximum temperature reached during rapid decomposition. For this purpose, they used the CHETAH software to calculate the standard heat of formation, the method of Melhem and Shanley [1996] to identify the most stable decomposition products to find the corresponding reaction heat and finally estimations of the heat capacities as function of temperature to calculate the adiabatic temperature rise (noted CART).

Considering the time and computational expenses of this procedure, Hada and Harrison [2007] proposed to distinguish between CART and MART, defined as the maximum adiabatic rise of temperature. To determine MART, they applied a highly similar procedure, while sparing significant computing time as they bypass the equilibrium-based determination of the most stable products and select the reaction route that maximizes the temperature rise. Despite the theoretical flaws, MART calculations are conservative from a safety perspective as they most probably over-estimate the actual ΔT_{ad} of decomposition. This tends to miss classification of non-hazardous compound to hazardous, which represents an error on the safe side.

In 1991, two studies analyzed thermal stability of compounds with similar approaches. A research from the Japanese Research Institute of Industrial Safety presented an analysis of the DSC thermograms of 820 chemicals [Ando et al., 1991]. The compounds are classed according to their characteristic functional groups and for each category, mean values for the onset temperature and the reaction enthalpy and a percentage of exothermic samples in the considered category are presented.

However, compounds bearing more than two functional groups were listed in several categories, thus blurring the statistics as they have been taken into account repeatedly. Moreover, non-exothermic compounds were also maintained within classes, diminishing significantly in some cases the mean ΔH_r values. Therefore, using this classification in order to estimate the characteristics of a compound's decomposition might results in erroneous evaluations. Besides, presenting a standard deviation for the mean values calculated and assigned to each category would have reflected the dispersion within class.

The same year, Grewer presented a study conducted with a similar approach to this of Ando et al. [Grewer, 1991] . The author analyzes the decomposition energies relatively to the chemical structure of a data collection found in the literature, experimentally measured on DSC and ARC systems, or computed from known formation enthalpies (using CHETAH when possible). He also computes the average values for each category. However, the results for the onset temperatures are noted to vary greatly with the different substituents a molecule may bear and their position (especially with aromatic compounds) and therefore, the average onset temperature values cannot be applied for estimations. For instance, the average decomposition energy of an organic peroxide is noted to be about $\Delta H_r = -300 \pm 50 \mathrm{kJ\,mol^{-1}}$, or the opening of a double bond gives in average $\Delta H_r = -80 \pm 19 \mathrm{kJ\,mol^{-1}}$. With nitro compounds, the author noted the dependance of the energy generation with the number of nitro groups, hence, he suggests that the decomposition energies are $\Delta H_r = -400\ \mathrm{kJ\,mol^{-1}}$ per NO_2 group.

Unfortunately these two studies cannot be compared due to the use of different units: while Grewer reports his values in [$\mathrm{kJ\,mol^{-1}}$], Ando et al. uses [$\mathrm{cal\,g^{-1}}$].

Grewer also investigated T_{D24}, the temperature at which the adiabatic induction time to maximum rate is 24 hours and suggest it is correlated to several functional groups [Grewer et al., 1999]. These correlations are drawn based on nitro compounds with secondary substituents.

Nitro compounds are frequently used in the chemical industry and are known to decompose with a high energy release; thus they present a serious risk in the industry [Grewer et al., 1999]. Therefore, the amount of published data regarding their thermal stability is larger than for other chemical families, which in turn, also attracts further interest, especially applications of predictive modeling which is rather data consuming.

For instance a study on nitro compounds incompatibility published DSC data of various chemicals including a set of 24 nitro compounds [Duh et al., 1997]. The authors qualitatively discussed the effect of several substituents on the stability of the compounds, showed an effect of the ortho- meta- or para- position of multiple NO_2 and of the number of NO_2.

Their data were later employed for the development of predictive models by other research groups. For instance, the decomposition enthalpies of 19 nitroaromatic compounds were correlated to n_{NO_2}, the number of NO_2 in the molecule, with a mono-parametric relationship (see Equation 6.1), and the average relative deviation of $ARD = 5\%$ shows good agreement between measured and predicted data [Saraf et al., 2003].

$$\Delta H_r = -75 \cdot n_{NO_2} \qquad \text{in kcal mol}^{-1} \tag{6.1}$$

The activation energies of thermal decompositions of nitramines [Keshavarz, 2009a], and nitroparaffins [Keshavarz, 2009b], and the onset temperature of decompositions of polynitroaromatic [Keshavarz et al., 2009b] were correlated to the molecular structures through relationships comparable to GCM. The used parameters were mainly the elemental composition of the compounds (atoms counts), however some other parameters were included such as the oxygen balance or a binary parameter that takes into account the presence or not of carbons atoms bridging two aromatic groups. Therefore, these models do not correspond exactly to GCM nor QSPR as seen in Chapter 1, however the structural descriptors do not require particular computing and are developed for simple modeling purposes.

The oxygen balance is defined as follows:

$$OB = \frac{-1600(2n_C + n_H)}{2 - n_O} \cdot \frac{1}{MW} \tag{6.2}$$

where $\mathrm{n_C}$ is the number of carbon atoms, $\mathrm{n_H}$ the number of H atoms, $\mathrm{n_O}$ the number of oxygen atoms and MW the molecular weight [Lothrop and Handrick, 1948].

More recently, the DSC data of 198 ionic liquids were modeled by GCM [Lazzús, 2012]. The contributing groups that allow to represent all molecules of the dataset were identified as 11 cation substructures, 15 substituents to these substructures, and 31 anion substructures. Their contributions to the thermal decomposition onset temperature were assessed from a training set of 120 observations and were tested on the remaining 78 observations. The author reports satisfactory results with average relative deviations of $\mathrm{ARD_{Tr}}$= 4.34 % and $\mathrm{ARD_{Val}}$= 4.18 % for the training and validation sets respectively.

6.1.2 Quantitative Structure-Property Relationships

In the same study mentioned above, Saraf et al. [2003] propose a GCM model for decomposition enthalpies of 19 nitroaromatics, but also a QSPR model of the onset temperature of these reactions [Saraf et al., 2003]. Their model depends on 3 parameters only, namely the highest positive charge, HPC, the electrons delocalizability index, $\mathrm{S_r}$, and the dipole moment, μ. The obtained simulations of $\mathrm{T_o}$ show significant agreement with experimental data and present an ARD= 6 %.

Despite the small deviations between experimental and predicted data, the correlation coefficient was rather low and no validation tests were performed, as noted by Fayet et al. [2010] Therefore, Fayet et al. reproduced the models of Saraf et al. by analyzing the same set of nitro compounds (experimental data from [Duh et al., 1997]), and propose different models, that are arguably more robust. The major improvements are in the use of a validation set to test the models and higher correlation coefficients ($R^2 = 0.91$ and $R^2_{cv} = 0.84$) .

They apply a procedure to develop several predictive models from the QSPR method for decomposition enthalpies of nitro compounds [Fayet et al., 2009, 2010, 2011], electric spark sensitivity [Fayet et al., 2010] (discussed in Chapter 5) and for impact sensitivity [Fayet et al., 2012].

The descriptors used are generated with Codessa software [Petrukhin et al., 2001] or from the Density Functional Theory (DFT) method. They select the Best Multi-Linear Model (BMLR) according to the correlation coefficients yielded by the smallest number of parameters following a so-called "breaking point rule": the correlation coefficient is analyzed in function of the increasing number of parameters for nested models and if the R^2 increase drops, the parameter addition is stopped. Moreover, they test several random training-validation divisions and select the division that performs the best correlations. This allows to avoid the dependance of the performance evaluation on the training or validation sets, to ensure that if there are outliers in the validation sets, they would not affect the models' performances for instance.

For a study conducted on 77 nitro compounds (experimental data from [Ando et al., 1991]), they obtained lower performances than for their previous study on 22 nitro-aromatic compounds ($R^2 = 0.77$, and $R^2_{cv} = 0.70$) and therefore suggest a decision tree to categorize data into two classes with $\Delta H_r > 500\,\mathrm{kJ\,mol^{-1}}$ or $\Delta H_r \leq 500\,\mathrm{kJ\,mol^{-1}}$.

Organic peroxides constitute another chemical family that has been often studied due to relatively high thermal reactivity [Ben Talouba et al., 2011, Li and Koseki, 2005, Melhem and Shanley, 1996].

Lu and coworkers looked into prediction of thermal stability of organic peroxides, characterized by the onset temperature of decomposition and the heat release [Lu et al., 2011]. To build their models they used two statistical methods: Multi-Linear Regression (MLR) and Partial Least Squares (PLS). It was observed that the MLR method proposes models with fewer parameters than PLS, but the cross-validation tests reveal poor predictive power (low or even negative correlation coefficients).

Therefore, the PLS method is favored, and in order to limit the number of parameters, the authors suggest performing a sensitivity analysis which would highlight the parameters that have the greatest influence. Then, a 'breaking point rule' similar to the one proposed by Fayet et al. [2010] could be applied to suppress some parameters and favor a model presenting a good compromise between correlative and predictive power and the number of parameters.

In 2014, another study proposes predictive models for the onset temperature of decompositions reactions of both nitro compounds and organic peroxides [Zhang et al., 2014]. The data sets they use have both been investigated in the past (data from [Ando et al., 1991, Lu et al., 2011]) but they apply the genetic algorithm (GA) to identify the best descriptors combinations to build the predictive models.

The genetic algorithm among other non-linear techniques, has been more and more employed in recent years for development of predictive models of safety related data [Gharagheizi, 2009, Mallakpour et al., 2014, Pan et al., 2009, 2010, 2011] as well as the Artificial Neural Network algorithm (ANN) [Jun et al., 2006, Lazzús, 2011, Nefati et al., 1996]. As they propose models of impact sensitivities, AIT and other explosive characteristics closer to MIE than to DSC data, most of these models were discussed in Chapter 5.

A similar approach has also been applied to study the decomposition temperature of chiral polymers [Mallakpour et al., 2014]. The genetic algorithm is applied to screen among the available descriptors, in order to select the most relevant ones for further non-linear modeling with Support vector machine (SVM). With training and validation sets of 38 and 12 observations respectively, they select five QSPR descriptors to develop the SVM model. They obtain high correlations between experimental and predicted data as $R^2 = 0.995$ and $R^2_{cv} = 0.992$, however the structure of the model is rather complex and computationally expensive.

Klos et al. developed MLR and ANN models for thermal stabilities of 66 derivatives of phenylcarbamic acid gathered from DSC experiments. Their work exhibits the singularity of also considering the temperature position of the peak maximum [Klos et al., 2008]. However, the models show low performances, even with the non-linear modeling.

Table 6.1 summarizes the predictive models built for DSC data with both GCM and QSPR methods.

Overall, this discussion allowed highlighting the different issues encountered by different research groups during their work.

Mainly, the data availability seems to be a recurring limitation. The experimental data of the thermal property to study has to be uniform in order to allow proper correlations. This imposes comparable experimental conditions and should preferably come from a single reference or experimental acquisition. This seems to be the main reason that some published data sets serve for several studies.

Hence, the data sets are usually of restricted sizes (many studies from Table 6.1 considered sets comprising about 20 observations) and usually focus on a specific chemical family so that the compounds have similar features (i.e. nitro compounds mostly, in Table 6.1 only three studies focus on other categories).

The calculated molecular descriptors must be available and relevant for each compound of the considered set. Then, when developing predictive models, it is important to keep in mind how these models could be applied later on. Thus, if a model is developed based on a narrow set of data or if the used descriptors are only available for a limited kind of chemicals, then it cannot be applied to chemicals external to this set, or only with a low reliability. Thus, the group contribution method is nonetheless interesting, but for future purposes, the QSPR methodology is preferable as it avoids limiting the method to the considered groups only and does not rely either on the availability of other physico-chemical properties.

Another important element that appears from the comparison in Table 6.1, is that, for larger sets, the fitting and predictive performances are much lower than studies performed on narrower sets. Indeed, chemicals from similar subgroups will have comparable behavior and thus can fit into a model, but this model may be unable to apply to several classes. This has been noted when discussing the decision tree built to distinguish high and low decompositions energies [Fayet et al., 2011].

When facing a similar issue while developing predictive models for impact sensitivities of 161 nitro compounds [Fayet et al., 2012], the authors propose four distinct models : three local models for nitroaromatic, nitroaliphatic and nitramines and a global model valid for these three sub-groups.

Table 6.1 – Comparative Summary of Literature Models for DSC-Derived Data

Property	Reference	Method	Model Type	Dataset	Performance
Onset Temperature					
	Saraf et al. [2003]	QSPR	MLR	19 nitro compounds	$ARD = 5\%$
	Klos et al. [2008]	QSPR	MLR	66 phenylcarbamic acids	$0.27 < R^2 < 0.77$
			ANN	66 phenylcarbamic acids	$0.34 < R^2 < 0.86$
	Keshavarz et al. [2009a]	GCM	MLR	12 poly-nitro aromatic	$R^2 = 0.95$
	Lu et al. [2011]	QSPR	MLR	16 peroxides	$R^2 = 0.957$ and $R^2_{cv} = 0.108$
			PLS		$R^2 = 0.978$ and $R^2_{cv} = 0.859$
	Lazzús [2012]	GCM	MLR	178 ionic liquids	$ARD_{Tr} = 4.3\%$ and $ARD_{Val} = 4.2\%$
	Zhang et al. [2014]	QSPR	GA and MLR	63 nitro	$R^2 = 0.738$ and $R^2_{cv} = 0.753$
	Mallakpour et al. [2014]	QSPR	GA and SVM	50 polymers	$R^2 = 0.995$ and $R^2_{cv} = 0.992$
Decomposition Enthalpy					
	Saraf et al. [2003]	GCM	MLR	19 nitro compounds	$ARD = 6\%$
	Fayet et al. [2010]	QSPR	MLR	22 nitro aromatic	$R^2 = 0.91$ and$R^2_{cv} = 0.84$
					$R^2 = 0.98$ and$R^2_{cv} = 0.97$
	Fayet et al. [2011]	QSPR	MLR	77 nitro compounds	$R^2 = 0.77$and $R^2_{cv} = 0.70$
	Lu et al. [2011]	QSPR	MLR	16 peroxides	$R^2 = 0.96$ and $R^2_{cv} = 0.811$
			PLS		$R^2 = 0.956$ and $R^2_{cv} = 0.828$

Finally, from the DSC thermograms only the onset temperature T_o and the decomposition enthalpy ΔH_r were modeled by the publications discussed above.

6.2 Experimental Error and Confidence Interval

As discussed in Chapter 2, Section 2.5.3, the reliability of the experimental data should be assessed prior to modeling. For this purpose, a repeatability and reproducibility study was conducted on a set of 20 chemicals. These compounds were analyzed by Heat-Flux DSC and Power-compensation DSC following the procedure detailed in Chapter 4, Section 4.3.1.

Usually, the repeatability is assessed by analyzing one reference sample [Joint Committee for Guides in Metrology (JCGM), 2008]. Metals used for calibrations could be used as their thermal characteristics are known. However, in order to analyze the reliability of the DSC measurements over a large temperature range, with exothermic reactions, it has been chosen to analyze the results of the tests replicates of all the 20 compounds tested.

It is important to note that the data analysis was conducted on key characteristics of treated DSC thermograms, and not on raw data. The baseline corrections, curve integration, Fraser-Suzuki fittings and property extraction could introduce additional errors. However these steps are necessary and are systematically conducted for all thermogram analysis, therefore they could be considered as part of the measurement. Moreover, the repeatability conditions were also respected as the thermograms' treatments were also conducted by a single operator, following a unique protocol over a short period of time.

From the 20 compounds tested, some underwent several thermal events and thus, exhibited two or three peaks. For each thermogram, the best defined peak was selected: if a thermogram exhibited several peaks, the one whose characteristics varied less across the five replicates was held, while the others were disregarded. This selection allows to estimate the repeatability of the measurements under the best conditions, and to avoid taking into account errors that could have been introduced by thermograms manipulations, such as the baseline treatment. These issues were mostly encountered when the thermograms exhibited overlapping peaks: in those situations, the secondary peaks were often influenced by the baseline treatment and the occurrence of the initial thermal events, and therefore were not selected for the repeatability assessment. However, in the context of modeling, the selection of the main peak to be studied was not based on these criteria, but on the relevance of the peak to the thermal behavior of the compound (see Section 4.3.3).

A single-factor analysis of variance (ANOVA) determined that the samples different are significant and do not allow data to be analyzed as a single group. T-tests could indeed prove that some samples had relatively close means and could be assimilated, but the overall set is rather disparate. Moreover F-tests also showed that the replicates of different samples had different variances.

Nonetheless, due to these disparities, it is impossible to assess among the 20 compounds which one is the most representative of the reliability of the measurements. Therefore, the relative standard deviation of the mean for all tested groups were computed and compared. It appeared that besides few exceptions, the relative deviations were rather similar and could allow for the calculations of average relative deviations. So, after removing the five samples with the highest deviations (mainly thermograms with rather small peaks who suffered large deviations due to baseline treatments), the average relative deviations and repeatability coefficients are computed and the results are summarized in Table 6.2.

In Table 6.2, the symbol of the peak width σ has been replaced by FW for two reasons: the first is that the peak width is actually measured through the Full Width at Half-Maximum FWHM which is correlated to σ through the relation: $FWHM = 2\sqrt{2\ln 2}\cdot\sigma$; the second is to avoid confusion with $\sigma_{\bar{x}}$, the standard deviation of the mean of the replicated measurements and $\sigma_{\bar{x},r}$.

Table 6.2 – Repeatability Coefficients

		$\sigma_{\bar{x}}$	$\sigma_{\bar{x},r}$	$\mathrm{RC}_{95\%}$	$\mathrm{RC}_{99\%}$
Onset Temperature	$\mathrm{T_o}$	0.89 °C	0.5%	1.0%	1.5%
Reaction Enthalpy	$\Delta\mathrm{H_r}$	$3870\,\mathrm{J\,mol^{-1}}$	3.8 %	7.6 %	11.4%
Amplitude	Φ_{max}	$8.60\,\mathrm{W\,mol^{-1}}$	2.2 %	4.4 %	6.7%
Max Position	$\mathrm{T_{max}}$	0.47 °C	0.2 %	0.5%	0.7%
Full Width Half Max	FW	0.53 °C	2.7 %	5.4 %	8.1 %
Asymmetry	a	0.05	17.6%	35.1 %	52.7 %

In the literature, the typical error margins reported for DSC measurements are in the range of ±5 % to ±10 % for the heat of reaction [Ando et al., 1991], and about ±5 % for the onset temperature [Saraf et al., 2003]. In the light of our results, it seems that the accuracy of DSC measurements is much better than the expectations, especially regarding the temperature determination. Indeed, the onset temperature $\mathrm{T_o}$, and the maximum of the peak position $\mathrm{T_{max}}$, are measured with an accuracy of ±5 K, and 95 % of replicates would fall within a margin of less than ±1 %.

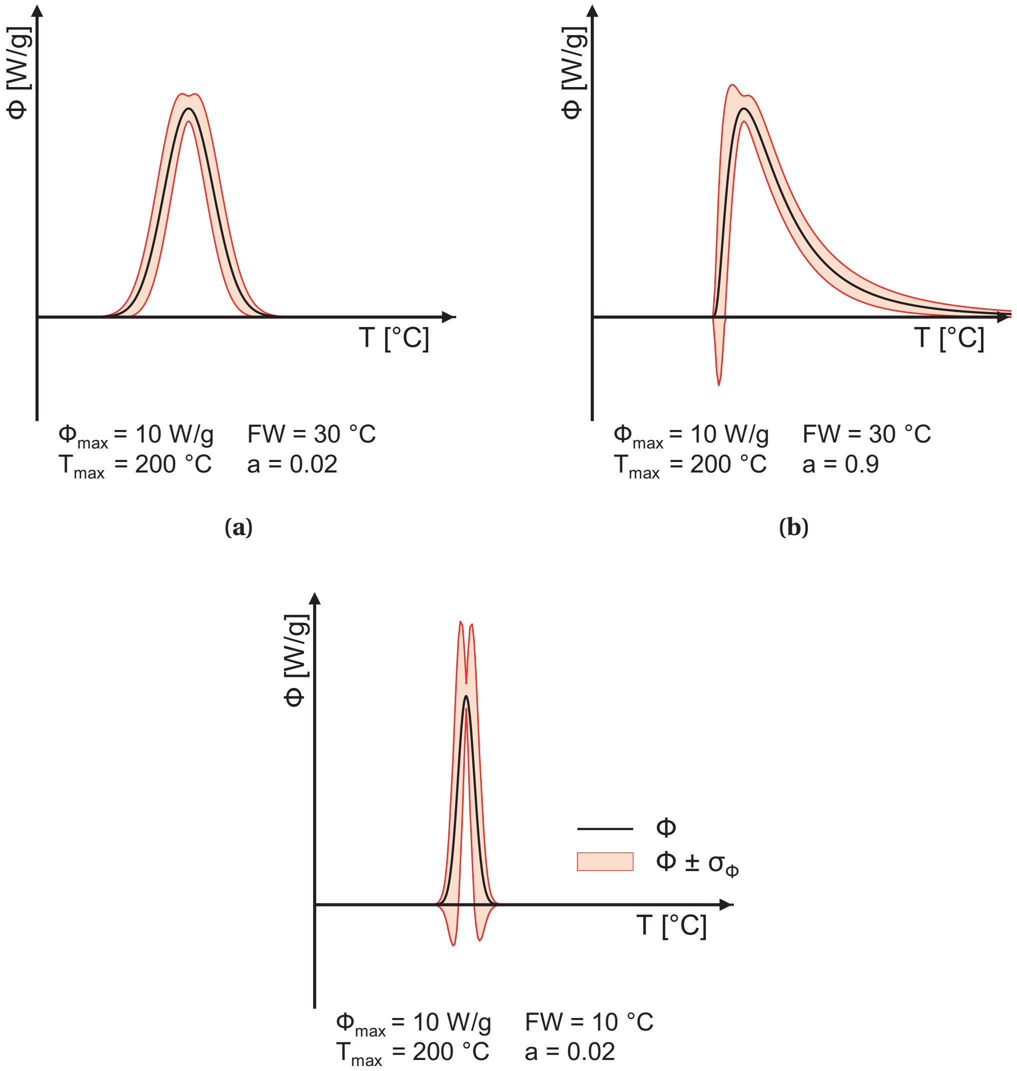
Regarding, the partial area, the influence of the baseline and the measured curve result in larger errors and the 95 % repeatability coefficient is ±7.6 %. The ±10 % margin found in the literature might correspond to the 99 % repeatability coefficient.

The asymmetry is represented by a dimensionless factor that takes small values mostly between ±1.2 in the studied set. Therefore, the relative deviations are much larger than all the other characteristics.

As the DSC curves are fitted by Fraser-Suzuki models, the inference of the error on each individual parameter included in the Fraser-Suzuki model has been assessed in order to estimate the overall impact on the DSC curve. For this purpose, a tolerance zone is constructed around the DSC curves. This tolerance zone is based on the estimation of the overall deviation of the curve σ_{Φ} depending on the deviations for each of the parameters $\sigma_{\bar{x}_i}$.

The results have been computed following the procedure detailed in Chapter 2, Section 2.5.3, and are detailed in Appendix B.1. Their implication is better explained when graphically represented, as shown in Figure 6.1.

For the tolerance zone construction, the relative standard deviations of the means of the different parameters $\sigma_{\bar{x},r}$ have been rounded up to the higher integer to be more inclusive of possible deviations. As the deviations for the temperature are small and a good accuracy is expected, the tolerance zone is computed so that ±2 % are accepted. Similarly, the error margins of Φ_{max} and FW were also increased to accept ±6 %. On the other hand, the deviations of the asymmetric factor used to model the tolerance zone are lower than the actual values , ±20 % instead of ±35 %, but the impact on the tolerance zone is limited. The Tolerance Zone coefficients (abbreviated TZC) are shown in Table 6.3.



(a) **(b)**

(c)

Figure 6.1 – Tolerance Zone

Table 6.3 – Tolerance Zone Coefficients

		$\sigma_{\bar{x},r}$	$\sigma_{\bar{x},TZ}$	TZC
Amplitude	Φ_{max}	2.2%	3%	6%
Max Position	$\mathrm{T}_{\mathrm{max}}$	0.2%	1%	2 %
Full Width Half Max	FW	2.7%	3%	6 %
Asymmetry	a	17.6%	10%	20 %

6.3 Local and Global Models

6.3.1 Objectives

Considering the models found in the literature and the successful results obtained by other research groups regarding the modeling of thermal stability of nitro compounds, it was decided to perform a primary study on nitro compounds as well, in order to develop our own procedure and to challenge the current state of the art in two aspects.

First of all, from all the literature review it appears that only the onset temperature T_o and the reactions enthalpy ΔH_r have been modeled [Fayet et al., 2010, Keshavarz et al., 2009a, Lu et al., 2011, Saraf et al., 2003]. Hence, our first goal is to assess whether the prediction of an entire DSC curve could indeed be performed.

Moreover, a limited number of studies propose models for large sets [Lazzús, 2012] or sets that do not focus on a particular chemical structure. The DSC data available enable us to expand the modeling in three steps:

- among the 400 compounds that constitute the database, several chemical families other than the Nitro group may be investigated separately;
- develop a structurally heterogeneous set, that does not focus on a specific structure in order to verify generalization;
- broaden the generalization even further with the development of "global models" by investigating the overall dataset available.

The distinction between "Global" and "Local" sets comes mainly from the fact that chemical families present intrinsic structural similarities and share common features, hence, they represent a "localized" region of the feature space.

In order to meet these objectives, four studies are conducted separately as summarized by Table 6.4. The following section will present and discuss the obtained results. In the discussion, the models are referred to relatively to the dataset that served for the modeling (for instance NO_2 models are the DSC models developed with a set of nitro compounds only, whereas "Global" models are the models developed with the overall dataset).

6.3.2 Nitro Compounds Study

The best multi-linear model built for each of the DSC key properties was selected based on the performance evaluation. The best models evaluations are summarized in Table 6.5, and the graphical representations of these models are shown in Figure 6.2. In Appendix B.2, Table B.1 shows all the parameters selected for the models and their assigned coefficients.

Table 6.4 – Model Investigation Plan

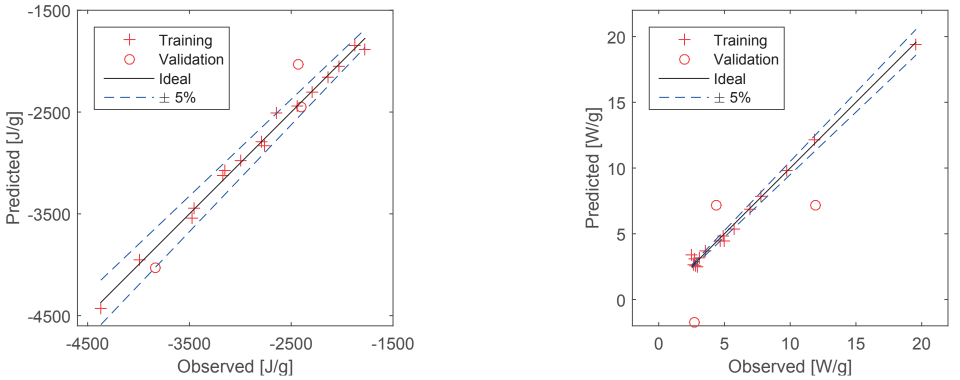
	Initial Set	**Extension**
Local Models	1. NO_2 Set Section 6.3.2	3. Various Chemical Families Section 6.3.4
Global Models	2. Miscellaneous Set Section 6.3.3	4. Overall Dataset Section 6.3.5

All these models very accurately describe the training set, as high determination coefficients were obtained for the training set. The highest deviations are observed for the asymmetry, but as already discussed previously, the asymmetry takes small values, and the relative deviations take large proportions. Indeed the highest relative deviations observed for the training set is recorded for the prediction of a for 2,4-dinitrotoluene: the observed value is $a = -0.06$ while the simulation gives $a = -0.15$ which results in a relative deviation of $RD = 145\%$. The results are detailed in Appendix B.2, Table B.2.

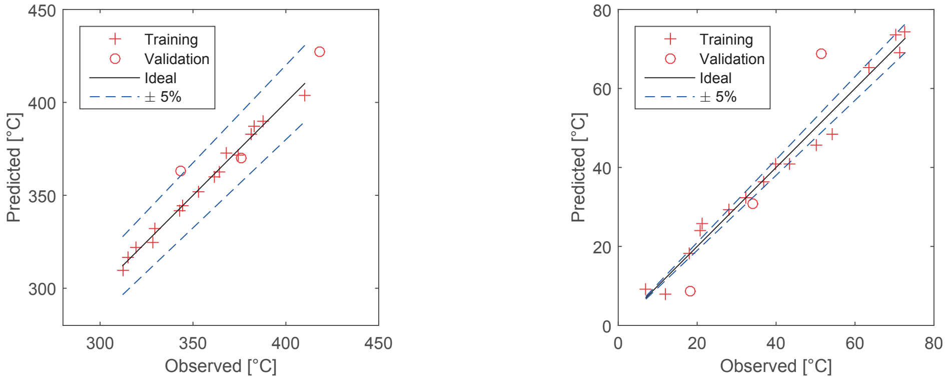
Regarding the validation, the models of peak's amplitude and asymmetry record the most important performance drops as they show lower correlation coefficients on the external validation set and higher average relative deviations for the validation. To ensure these models do not suffer from over-fitting, parameters were removed but this only decreased the goodness-of-fit, without improving the validation correlations.

The peak max position and the reaction enthalpy were very successfully modeled. The training set observations are well described and the models offer predictions of the external validation set with deviations lower than $ARD_{Val} < 10\,\%$. Regarding the peak width model, the deviations for both the training and validation sets are larger than these of $\mathrm{T_o}$ or $\Delta\mathrm{H_r}$, nonetheless the tendencies are well captured and the external set correlation coefficient is $R^2_{ext} = 0.985$.

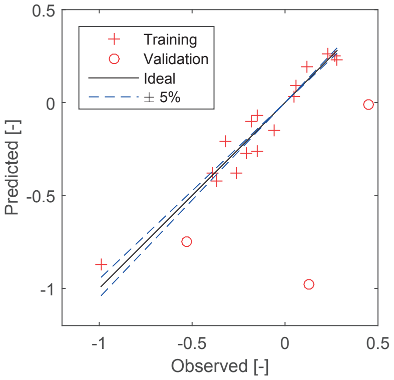
Despite some imprecision, the models were overall satisfactory, and the DSC curve reconstruction could be performed after the modeling of each DSC key property. Few examples are presented in Figure 6.3. Most of the DSC curves of the observations from the training set are very well represented as shown with the two examples of Figure 6.3 (a) and (b). Regarding the validation set, the predictions are not as accurate especially due to the important deviations in the $\Phi_{\max}$ model. For instance, the predicted DSC curve for 1,3-dinitrobenzene (Figure 6.3 (c)) is close to the experimentally measured peak in terms of position, width and asymmetry, however the peak's amplitude is under-estimated and the prediction is below the tolerance zone.



(a) Partial Area ΔH_r **(b)** Amplitude Φ_{max}



(c) Max Position T_{max} **(d)** Full Width FW

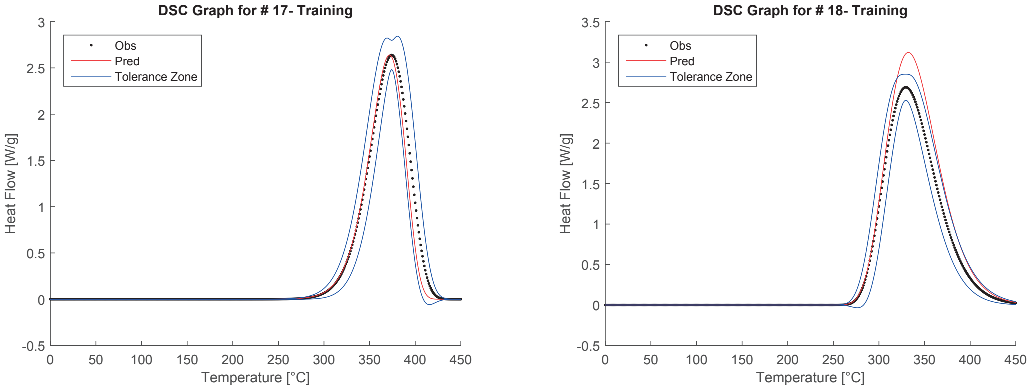


(e) Asymmetry a

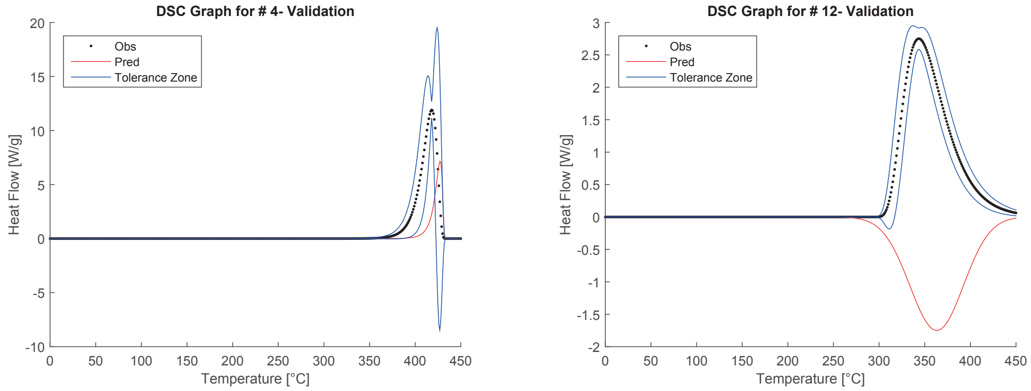
Figure 6.2 – Graphical Representations of the Nitro Compounds Models Responses

Table 6.5 – Nitro Models Evaluation Summary

Evaluation	Partial Area $\Delta \mathrm{H_r}$	Amplitude Φ_{max}	Max Position $\mathrm{T_{max}}$	Full Width FW	Asymmetry a
R^2_{Tr}	0.994	0.994	0.988	0.981	0.942
R^2_{cv}	0.980	0.983	0.967	0.950	0.784
R^2_{ext}	0.950	0.405	0.881	0.985	0.341
R^2	0.977	0.867	0.964	0.940	0.475
ARD_{Tr} [%]	1.7	7.0	0.7	9.3	39.7
ARD_{Val} [%]	7.9	89.2	3.2	32.2	340.6
Parameters	5	6	5	5	5
Dataset Size	19				
Training	16				
Validation	3				



(a) 4-nitrobenzoic acid **(b)** 4-nitrophenol



(c) 1,3-dinitrobenzene **(d)** 3-nitroaniline

Figure 6.3 – Examples of DSC Reconstructions from Nitro Models

The DSC prediction for 3-nitroaniline (Figure 6.3 (d)) shows the most erroneous prediction of

Table 6.6 – Miscellaneous Set Models Evaluation Summary

Evaluation	Partial Area $\Delta \mathrm{H_r}$	Amplitude Φ_{max}	Max Position $\mathrm{T_{max}}$	Full Width FW	Asymmetry a
R^2_{Tr}	0.976	0.973	0.925	0.978	0.924
R^2_{cv}	0.956	0.956	0.860	0.965	0.841
R^2_{ext}	0.986	0.953	0.984	0.887	0.290
R^2	0.932	0.866	0.926	0.920	0.789
ARD_{Tr} [%]	71.7	84.4	6.4	32.2	41.7
ARD_{Val}[%]	81.3	51.8	9.2	68.3	89.3
Parameters	5	6	5	5	5
Dataset Size	25				
Training	22				
Validation	3				

this set, as the amplitude is predicted to be negative while all the studied set only included exothermic reactions. 3-nitroaniline exhibits indeed one of the lowest energy release and smallest peak amplitude, and the model over-estimates this tendency and results in an amplitude prediction out of the studied range.

This first step shows that the DSC reconstruction method is successful and that the first goal is reached. Hence, the investigation may proceed on the application expansion to various sets.

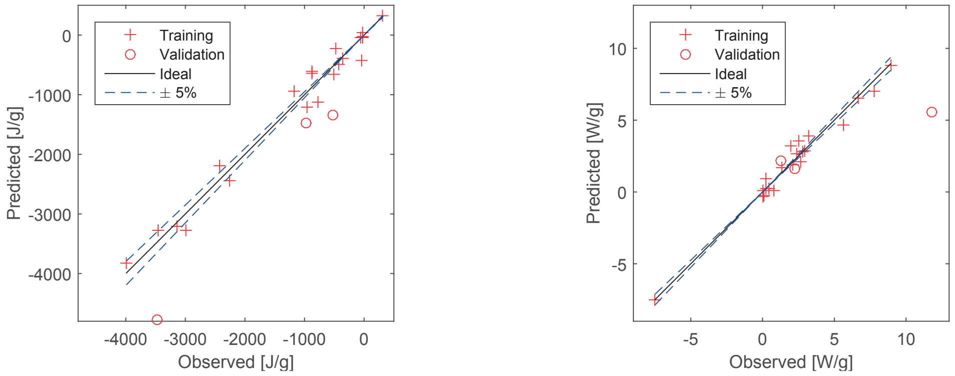
6.3.3 First Generalization

For the generalization study, the 20 DSC records used for the repeatability study were employed. However, if the dataset were constituted from the merger of 19 nitro compounds and 20 miscellaneous chemicals, it would not reflect a chemical diversity as the $\mathrm{NO_2}$ specificities would be over represented. Therefore, in order to obtain a balanced dataset, only five nitro compounds were selected for the fused structurally diverse set that finally comprises 25 observations. This set and its corresponding models will be referred to as "miscellaneous set".

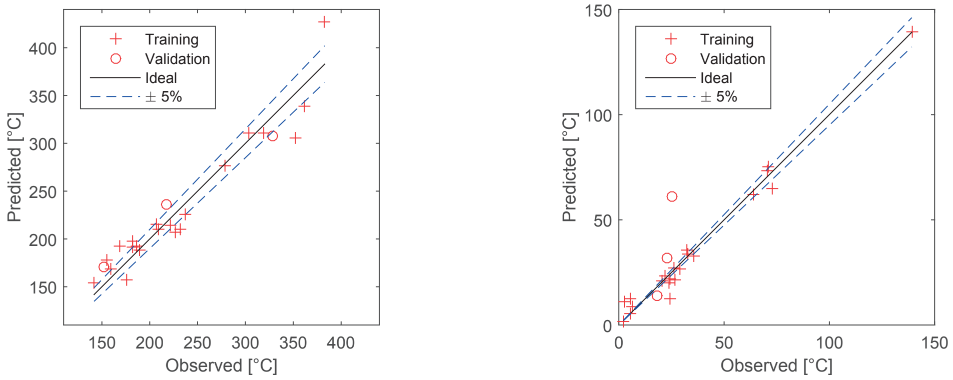
The obtained models are detailed in Appendix B.2, Tables B.3 and B.4, and their evaluation is summarized in Table 6.6.

In this case, the correlation coefficients are very high, for both the training and validation sets. On the other hand, the average relative deviations are globally higher than the $\mathrm{NO_2}$ models, for both sets as well.

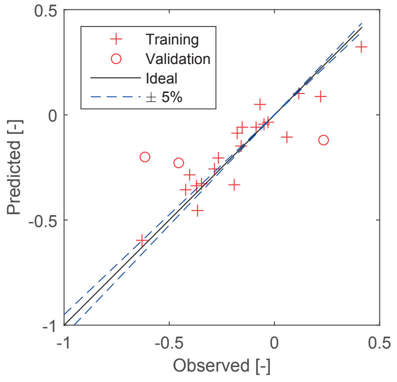
The DSC curves are recovered from the property estimations and predictions and examples are shown in Figure 6.5. As the estimations and observations in the training set are rather well correlated, as seen previously, some DSC simulations fit well or very close to the Tolerance Zone as shown in Figure 6.5 (a) and (b).



(a) Partial Area $\varDelta \mathrm{H_r}$ **(b)** Amplitude $\varPhi_{\mathrm{max}}$



(c) Max Position $\mathrm{T_{max}}$ **(d)** Full Width FW



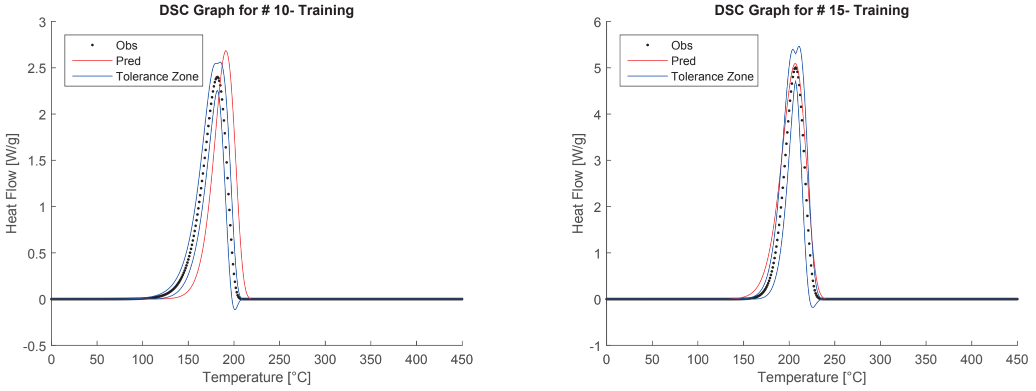
(e) Asymmetry a

Figure 6.4 – Graphical Representations of the Models for the Miscellaneous Set

However, the larger errors also reflect on the DSC predictions. In Figure 6.5 (c) and (d) for instance, the $\varPhi_{\mathrm{max}}$ model exhibits important inaccuracies, and as a result, the predictions either overestimate or underestimate the actual amplitude. Moreover, the $\mathrm{T_{max}}$ and FW models are less accurate than the models built with a single chemical family, which affects the positions and shapes of these peaks.

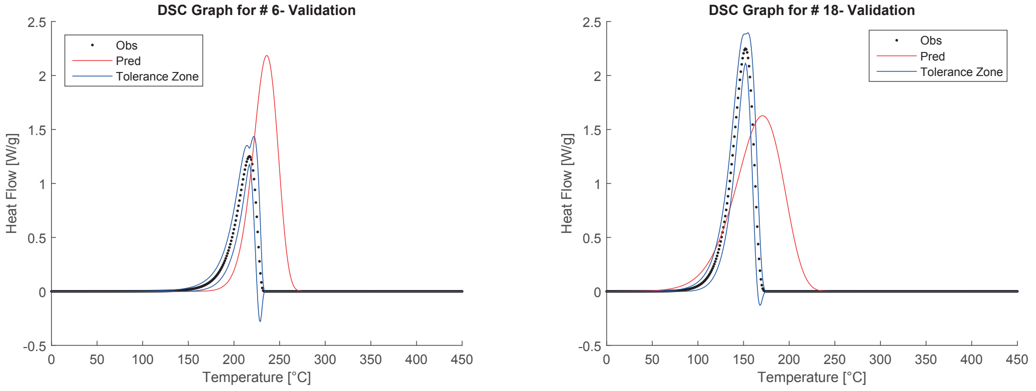
Finally, as for the NO_2 models, two of the smallest positive amplitude were predicted with negative values (#13 heptene and #17 NN-dimethylformamide). However, this can here be explained by the presence in this training set of a compound presenting an endothermic peak (#20 triethylphosphate) and therefore, negative values are not outside of the observed range for this set.

As five nitro compounds were included in both sets studied here, it is possible to compare the results obtained from local models (developed with a set of nitro compounds only) with results of a "global" model (developed with a diverse set, i.e. the miscellaneous set).



(a) di-t-butyl peroxide

(b) isopentyl nitrite



(c) di-(4-Cl-benzyl)azodicarboxylate

(d) t-butyl peroxyacetate

Figure 6.5 – Examples of DSC Reconstructions from Miscellaneous Set Models

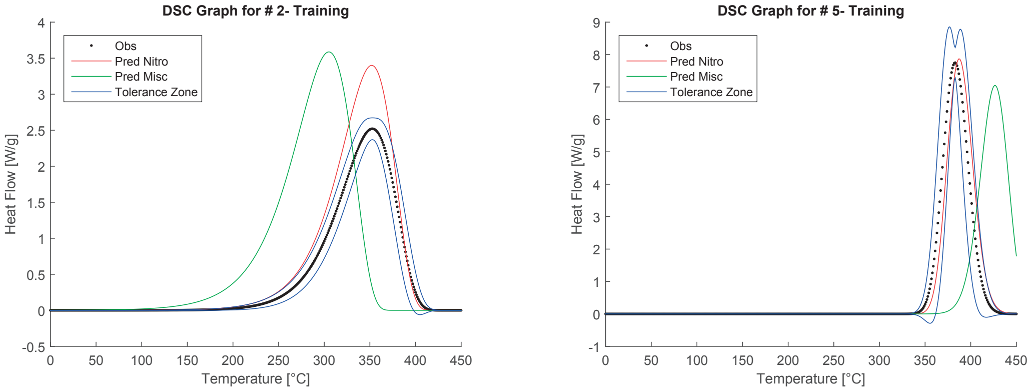
Figure 6.6 gathers the DSC estimations and predictions of the nitro compounds for which both local and global models were developed. Clearly, the local models outperform the global models in all cases shown here. Nevertheless, the global models are not irrelevant: the average relative deviations for this specific group of five observations is of $ARD = 8.0\%$ and $ARD = 11\%$

for the $\mathrm{T}_{\mathrm{max}}$ and FW respectively. For the Φ_{max} the average deviation reaches $ARD = 22\%$ but this is mainly due to high deviations recorded for the prediction of the amplitude of the DSC peak of # 11 3,4-dinitrotoluene (which belongs to validation set of "Miscellaneous set" and training set of "Nitro set"). This is confirmed in Figure 6.6 (d), where the predicted DSC curves for 3,4-dinitrotoluene: the Miscellaneous model underestimates by 53% the actual peak's amplitude while the Nitro model prediction is within the tolerated margins.

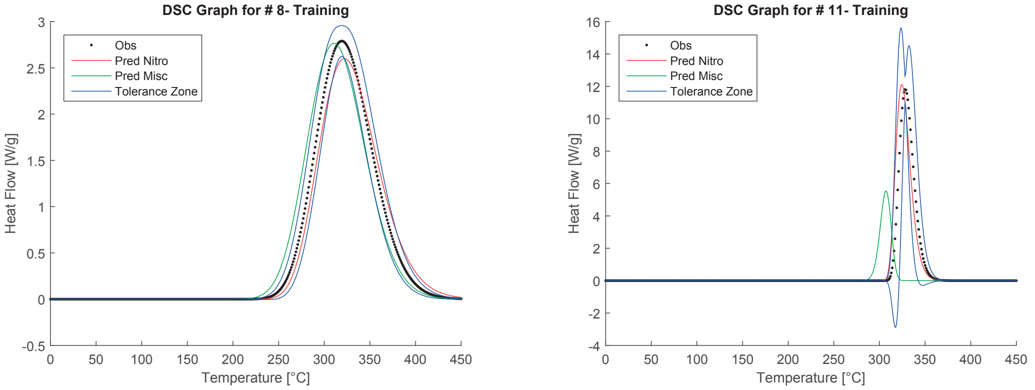
On the other hand, some DSC estimations are resembling much more to the measured peak, as in Figure 6.6 (c), for 2-nitrobenzoic acid, which is a close-to-ideal fit.

Models built on diverse sets seem to be less accurate than local models. The tendencies are well captured and the models generalize well from the training to the validation set, but the overall deviations are larger.

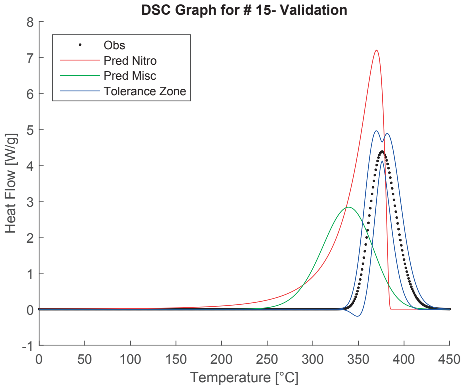
From this first attempt to compare local and global models, it comes out that local models offer higher accuracies, yet the risk of overfitting could hinder their predictive power, while global models perform less accurate estimations and generalize better.



(a) 2-nitroaniline **(b)** 1,4-dinitrobenzene



(c) 2-nitrobenzoic acid **(d)** 3,4-dinitrotoluene



(e) 3-nitrotoluene

Figure 6.6 – Comparisons of DSC Reconstructions

6.3.4 Chemical Families

The second expansion direction from the NO_2 set is to locally investigate other chemical families.

The structures of the compounds comprised in the collected dataset were analyzed and several specific sub-structures were identified for defining the chemical families. The membership of chemical to one or the other family is not exclusive and a single compound could appear in several families.

The identified families , their defining groups, their population size and the average values of the DSC key properties observed for the corresponding dataset are reported in Table C.1, in Appendix C.1. The obtained average values of the the peak positions and decomposition enthalpies are highly comparable to the results exposed by Ando et al. [1991].

In this table, the families are ranked by decreasing decomposition enthalpies. Hence, the nitro compounds are indeed at the top of the rankings as the most energy releasing compounds during their decomposition. Following, are several chemical families defined for the nitrogen-bearing functional groups: nitroso, nitrites, azo compounds, tetrazoles and amines.

It is important to note here that the organic peroxides are not reported on this table, mainly due to the fact that across the database of a few hundred observations, only 6 compounds were of the organic peroxide class, five of which have already been included in the miscellaneous set studied previously.

Out of the 14 families identified, 5 were selected to be modeled. The main criteria for this selection was the energy release of the decomposition reactions as it sets the interest of the family for safety considerations.

The families that served for model construction are Nitroso and Nitrites, Azo and Tetrazoles, Phenylamines, and Ethers, which complete the Top 5 behind the Nitro compounds in the chart of "most exothermic decomposition reactions", and Nitriles.

Nitroso and Nitrites on one hand, and Azo compounds and Tetrazoles on the other hand were considered together in two families as they are structurally comparable, but mostly, in order to obtain sets of sufficient sizes for modeling. The results of the modeling for each family separately are computed in Table C.2 in Appendix C.1. In this table, the evaluation criteria of the models are summarized and, when outliers are identified as highly impacting the relative deviations evaluation, a corrected ARD_c is computed to reflect the models performance on the set after the exclusion of the outliers. Only one outlier is removed if necessary, and only once was it necessary to remove two to recover values reflecting the overall set (for the asymmetry of Azo and Tetrazoles family).

All models parameters, graphic visualization and responses are gathered in Appendix C.1. Not all models are discussed in details here, however, few examples may be highlighted to develop the major outcomes of this modeling study:

- Overfitting is a recurring issue. Excessive parametrization of the models may generate models that highly correlate the training sets but fail to apply properly to the validation sets. In Figure 6.7 (a) is an example of overfitting model, with the representation of the $\Delta \mathrm{H_r}$ model's responses for Nitroso and Nitrates family. The correlation coefficients are rather high, $R^2_{Tr} = 0.958$ and $R^2_{Tr} = 0.929$, and $ARD_{Tr} = 7\%$.

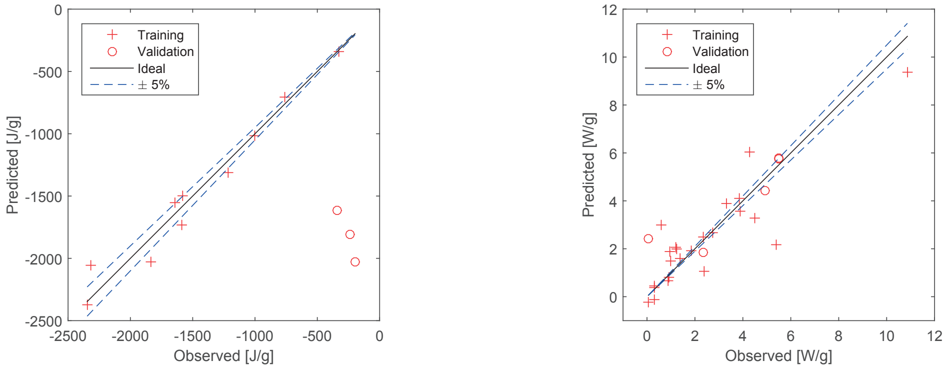
  This would be a successful model if the responses for the validation set were as close to observations as these results show. This is not the case as the model performance suffers an important drop for the validation set as $ARD_{Val} = 653\%$. Yet, the number of parameters is very limited, as only 3 descriptors are included in the model, and the removal of either one of them reduces the model's descriptive power. Therefore, it is not possible to decrease the model parametrization any lower.

  Besides, and this is more critical, the validation set contains three of the five lowest $\Delta \mathrm{H_r}$ values observed on the set. The model responses are indeed much higher than the expected values due to the fact that the model memorizes the training set values and projects the validation set to a higher range of $\Delta \mathrm{H_r}$ than their actual values. This suggests that the division of the dataset into training and validation sets should be revised in order to avoid this sort of distribution for the next models constructions.

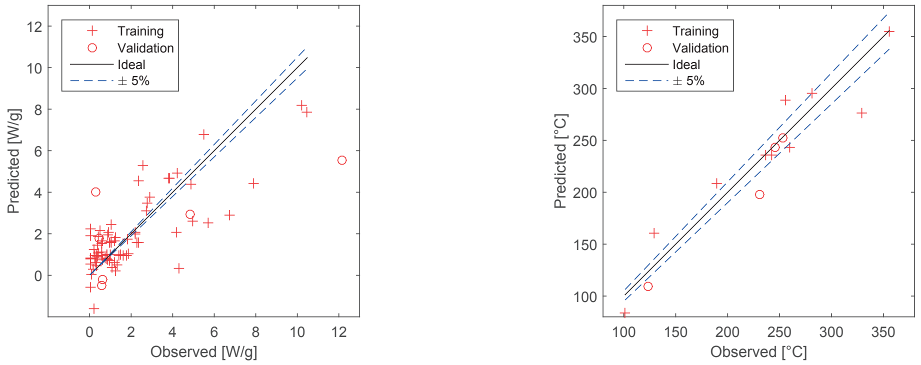
- In order to prevent overfitting, limiting the number of parameters included in the model is efficient. For instance, the model of Φ_{max} of Phenylamines compounds in Figure 6.7 (b) is developed with 5 parameters only and presents relatively low ARD. Nonetheless, this is done at the cost of the correlation coefficients. Indeed, considering the training and validation set correlations coefficients, $R^2_{Tr} = 0.781$ and $R^2_{Tr} = 0.774$, and that the model only includes 5 parameters, it would be possible to include additional descriptors, in order to increase the correlations. Yet, the $ARD_{Val} = 10\%$ and this can be regarded as satisfactory and justifies to stop the model parametrization at this level. It is important to note that here, the ARD_{Val} has been corrected by the exclusion of one outlier (# 26 4-t-butylaniline) for which the relative deviation was very high ($RD > 5000\%$), for the reason that its actual amplitude is $\Phi_{max} = 4.73 \times 10^{-2}\,\mathrm{W\,g^{-1}}$ and the prediction is $\Phi_{max,p} = 2.42\,\mathrm{W\,g^{-1}}$.

- Figure 6.7 (c) presents the model Φ_{max} of Ethers family. The peaks amplitudes of the 78 observations in the Ethers family are highly disparate and non-uniformly distributed on the observed range: the values range between $0\,\mathrm{W\,g^{-1}}$ to $12\,\mathrm{W\,g^{-1}}$, with an average about $1.91\,\mathrm{W\,g^{-1}}$, yet 50% of the observations are in the $0\,\mathrm{W\,g^{-1}}$ to $1\,\mathrm{W\,g^{-1}}$ range. This gives rise to inefficient models that fail both at describing the training set and predicting the validation set ($ARD_{Tr} = 263\%$ and $ARD_{Val} = 343\%$).

- Finally, there were some successfully modeled families as the Azo and Tetrazoles, Nitriles and Phenylamines, which present overall good results for all their properties (good performance on the training set and generalize well to the validation set, high correlations,



(a) ΔH_r of Nitroso **(b)** Φ_{max} of Phenylamines



(c) Φ_{max} of Ethers **(d)** T_{max} of Azo

Figure 6.7 – Examples of Models Responses for Chemical Families Sets

reasonable deviations, restricted number of parameters). For instance, the T_{max} model of Azo and Tetrazole family, shown in Figure 6.7 (d) is an example of successful model.

6.3.5 Global Models

In the previous study, several chemicals of the dataset were not investigated and their properties were not modeled, as they did not belong to the studied families. On the other hand, as the membership to several families is allowed, some data may be duplicated while others are not taken into account.

Similarly to the generalized Nitro study (Section 6.3.3), global models could be studied to see if models not restricted to a single subset of observations could be developed.

The entire database of DSC records was analyzed. For each substance exhibiting an exothermic behavior, one peak is selected, and its five DSC key properties serve as the entries.

A first selection of the QSPR descriptors eliminates those with the lowest variances or with

Table 6.7 – Global Models Evaluation Summary

Evaluation	Partial Area ΔH_r	Amplitude Φ_{max}	Max Position T_{max}	Full Width FW	Asymmetry a
R^2_{Tr}	0.731	0.547	0.275	0.113	0.112
R^2_{Val}	0.313	0.352	0.213	0.112	0.002
ARD_{Tr} [%]	424	659	31	98	189
ARD_{Val} [%]	355	608	36	172	369
Parameters	22	18	10	6	6
Dataset Size	375				
Training	337				
Validation	38				

missing values resulting in a working set of 351 descriptors per structure.

Due to the observation and features dimensions being too close (375 to 351), it was necessary to include a feature selection step prior to modeling. For this purpose a PCA over the descriptors space is performed and the 5 Principal Components (PC) computed. Then only the descriptors that contribute the most to these five PC are held, resulting in a reduced feature matrix of 250 features for 375 observations.

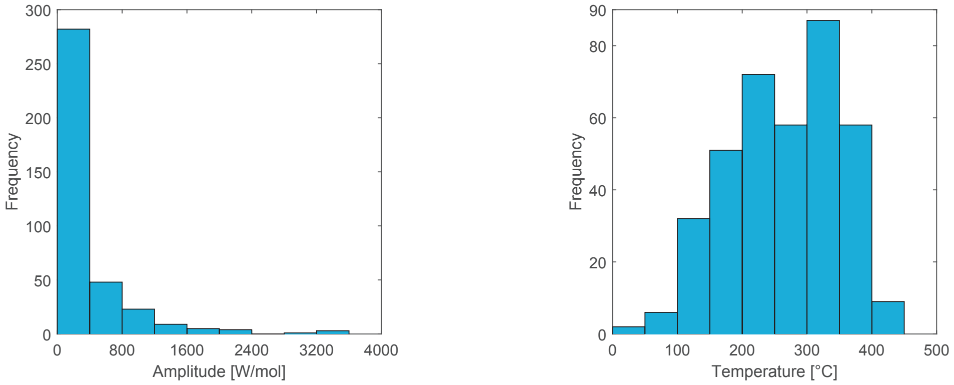
Then, a stepwise procedure is conducted to generate each property's corresponding model. For this attempt, the p-value for the inclusion of parameters was set to $p - enter = 0.05$.

These results are rather poor: the models fail at being descriptive of the training set and predictive of the validation set. The relative deviations are high and the correlations weak. Few outliers were noticed, however, their removal did not bring much improvement on the overall outcome, and these modifications were not reported here.

To test higher parametrization of the models, the p-value for the inclusion of parameters was varied up to $p - enter = 0.20$. This rises the correlation coefficients of the training set, on the other hand, the deviations are increased as well, which indicates clear overfitting. So this option is not maintained.

The distribution of the observed values for each property was analyzed and it revealed unbalance in some cases, which could hinder the correlations of the DSC properties to the molecular structures. Figure 6.8 shows two examples of property values distribution: Φ_{max} and T_{max}. From the analysis of Φ_{max} distribution, it appears that approximately 50% of the observations span in the lowest 4% of the observed range, and over 70% of the observations are concentrated in the lowest 10%. A similar unbalanced distribution is also present in ΔH_r, whereas values of T_{max}, FW and a follow distributions close to normal distribution.

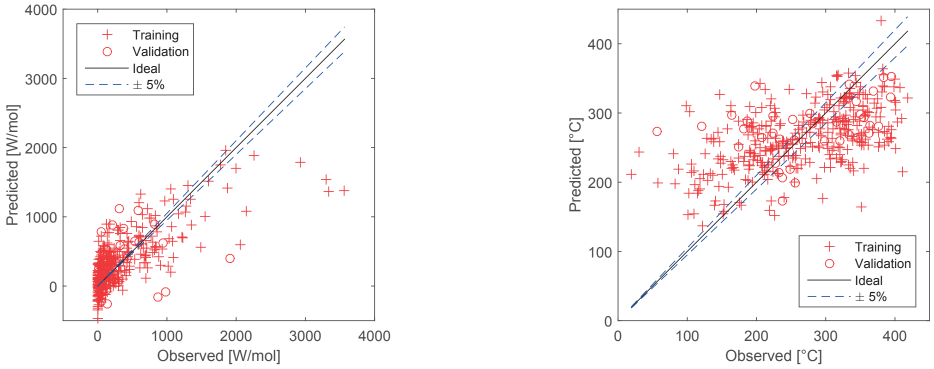
The graphical representations of the models' responses vs the observed values for Φ_{max} and T_{max} are presented in Figure 6.9. The previous remarks concerning the distribution of the



(a) Distribution of $\Phi_{\max}$

(b) Distribution of $\mathrm{T}_{\max}$

Figure 6.8 – Examples of DSC Properties Distribution



(a) Global Model $\Phi_{\max}$

(b) Global Model $\mathrm{T}_{\max}$

Figure 6.9 – Examples of Global Models Responses

values of $\Phi_{\max}$ are clearly visible here: a very important concentration of values are in the region of $\Phi_{\max} = 0$ to $200\,\mathrm{W\,mol^{-1}}$. As a result, the model is biased and predicts underestimated values for the compounds with higher $\Phi_{\max}$. The observations in the lower region of $\Phi_{\max}$ are also predicted erroneously, and give rise to large relative deviations, hence the obtained results showed in Table 6.7.

The property value distribution probably hinders the model formation, however it is certainly not the only element. The $\mathrm{T}_{\max}$ property values exhibit a nearly normal distribution, yet the model obtained is relatively weak. This suggests that the feature space from which the correlations are drawn may be inappropriate to represent the overall dataset, or that it does not hold the right information to represent the properties. It could be that particular features only suit the dataset partially, hence the local models obtained so far, and face their limits when covering the larger ensemble.

It is clear at this point that the considered dataset is not appropriate for a unique global model, that could apply to all chemicals included. The "one size fits all" type of model could not be achieved in a robust and valid way. This would suggest to favor local models and work further into their improvement rather than pursuing the development of a global model.

This assessment poses few questions, that will be addressed in the following parts:

- When building the local subsets from chemical families, how to address the multiple membership issue correctly? Is there a prioritization of chemical families or functional groups?
- Are there appropriate ways to create local subsets different than the chemical families?
- If several local models are available, how to determine which one to employ to predict behavior of a specific compound?

Moreover, the analysis of the previous results calls for further consolidation of the modeling procedure. It has been mentioned previously that when the validation set includes observation data at the edge of the observed range, the models are unlikely to properly predict it. Therefore, the separation procedure of the training and validation sets should be revised in order not to be randomly performed, but to take further considerations to avoid assigning extreme or under-represented cases into the validation set. The feature selection through PCA is a modification that was implemented and, as it was beneficial, it is henceforth applied routinely.

6.4 Systematic Construction of Local Subsets

6.4.1 Modified Modeling Procedure

In order to answer the interrogations that sparked off in the last section, and to improve our current modeling process, different strategies will be applied in parallel. First, in order to determine if local subsets can be developed on different basis than the chemical families, clustering based on the features space and on the labels space will be performed separately. Then, to tackle the chemical family hierarchization, QSPR does not seem appropriate, and therefore GCM will be applied.

To implement these strategies, the procedure is modified as schematically represented in Figure 6.10.

- From the dataset, the training-validation separation, is no longer performed randomly. The dataset distribution is evaluated through the mean and the scattering of the DSC properties. Ten random training sets and the corresponding validation sets are generated, and the selected separation is the one for which the means in the training and in the validation are the most similar. This reduces the risks to form an unrepresentative training set, or the assignment of extreme cases into the validation set.

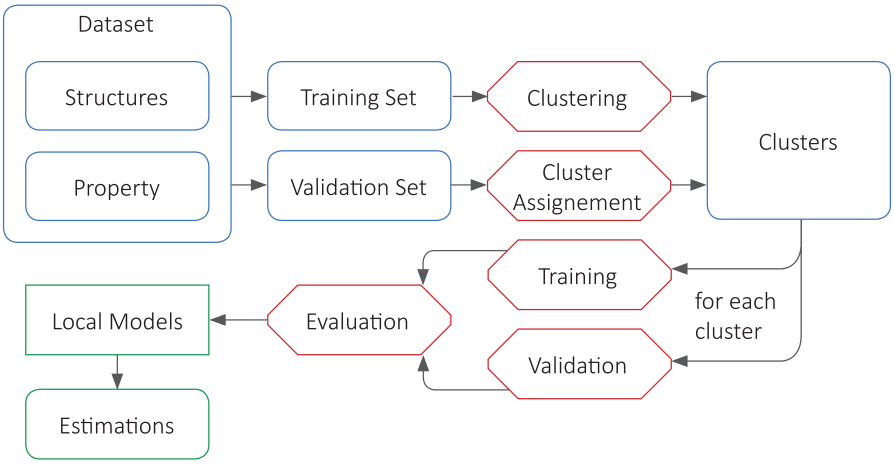


Figure 6.10 – Schematic Representation of the Modified Modeling Procedure

- The training set serves to cluster the data. Two different clustering approaches are applied:

    - k-means clustering on the features space (i.e. the structural descriptors)
    - hierarchical clustering on the labels space (i.e. the DSC properties).

  In both cases, several clusters are created. The observations in the validation set are then assigned to either one of the clusters, so that each cluster's population contains training and validation data.

  It is important to note a major difference here, that is when clustering is conducted on the features space, the introduction of an out-of-the-set molecule is straightforward: from its structure, the distances to all clusters centroids are computed, and it can be assigned to the closest cluster. If the clustering is performed with the DSC properties, classification is required. Indeed, in the context of prediction, or even for the validation set simulations, only the molecule's structure is known and the DSC parameters are to be determined. Therefore, its assignment into one of the clusters must rely on its structural features. Therefore, the procedure step "Cluster Assignment" in Figure 6.10 varies depending on the applied method and includes classification to complement for label space clustering. For this purpose, a decision tree is developed in Section 6.4.3.

- Then, the modeling proceeds within the different clusters to develop local models. For every model, the stepwise procedure is run, and the $p-enter$ value is varied to optimize the models.

- Finally, the estimated and predicted values of the DSC properties serve for the DSC reconstruction with the Fraser-Suzuki equation.

Table 6.8 – Reduced Feature Space

i	Parameter	Name
1	γ	1X GAMMA polarizability (DIP)
2	T^{E}_{all}	Topographic electronic index (all pairs)
3	$HASA2TS$	HASA-2/SQRT(TMSA) (Zefirov PC) (all)
4	J	Balaban index
5	$p_{A,min}$	Min net atomic charge

6.4.2 Features Space Clustering

In this part, the QSPR descriptors of the chemicals constitute the feature space.

Through the PCA procedure detailed in Section 2.2.3, the features of the training set observations are analyzed and the five PC are determined, then a reduced feature space made of the five descriptors that contribute the most to Principal Components is generated. The k centroids of the clusters are generated and adjusted in the five-dimensional reduced features space. The descriptors selected are shown in Table 6.8.

The number of clusters k to build was varied in order to maximize cluster separation. It came out that between 4 and 7 clusters could be built to obtain similarly separated clusters. Therefore, another criteria was applied to set k: the cluster population. When 6 or 7 clusters are built, at least two groups are of restricted sizes (less than 15 observations), which could be problematic for the upcoming modeling. Therefore, k is set to 4.

The validation observations are then projected onto the reduced feature space, the distances to each centroid are computed, and the observations are assigned to the closest cluster.

The model construction is then conducted within the clusters. The evaluation summary for the models is shown in Appendix C.2, Table C.9. As several models are built in every case, a 'leave-many-out' cross-validation was required to allow selecting the best models. Five observations are left out of the training and serve for the evaluation. AIC and BIC are also computed and the "best models" selected are the models that achieve the highest compromise of low ARD_{Val}, low AIC and good fitting correlation coefficient R^2_{Tr}. In the worst cases where no model stands out as the "best", the one with least parameters is held.

To give an overview, Table 6.9 "summarizes the summary" with the average values of all evaluation criteria across the four clusters.

These results reflect rather weak models and hand-picking is necessary to find models that yield ARD values in the range of 20%. The R^2_{Tr} were purposely kept low to avoid highly parametrized overfitting models, yet their performance and their generalization are limited. Only cluster 3 shows high correlations to the training set, but considering its relatively narrow building data set (15 observations), and poor predictive performance, it does not constitute a

Table 6.9 – Clusters Models Evaluation Summary

Evaluation	Partial Area $\Delta \mathrm{H_r}$	Amplitude Φ_{max}	Max Position $\mathrm{T_{max}}$	Full Width FW	Asymmetry a
R^2_{Tr}	0.741	0.716	0.718	0.687	0.744
R^2_{Val}	0.637	0.238	0.297	0.409	0.662
ARD_{Tr} [%]	392	421	18	49	161
ARD_{Val} [%]	194	104	30	74	135

significant contribution.

The underlying cause of these poor results seems to be, here again the unbalanced representation of the properties. Distributions similar to these shown in Figure 6.8(a) are observed here for $\Delta \mathrm{H_r}$ and Φ_{max} in clusters 1, 2 and 4.

As the set imbalance is a recurring issue and limits the modeling even in the case of grouping by structural similarities (clustering on the feature space), it seems appropriate and necessary to implement a clustering on the properties space.

6.4.3 DSC-based Clustering and Classification

For the aforementioned reasons, clusters are developed on the DSC properties space. This part has been the subject of a master thesis conducted in collaboration with the present work [Mage, 2015].

200 DSC curves are processed to render images of the studied peaks and sequenced into vectors. Hierarchical clustering is applied on the DSC image space. The clustering objectives, as in the previous section, are to maximize inter-cluster dissimilarities while maintaining cluster population above 15 observations. Here, the ideal clusters number was $k = 7$. The ideal settings were determined after several comparative analysis that will not be detailed here [Mage, 2015].

Every cluster is represented by the most likely DSC peak: all properties within the cluster are averaged and used to create a representative "typical" curve. These curves are shown in Appendix C.3, Figure C.6.

Following the partition of the DSC thermograms into 7 clusters, classification is required. Indeed, for a new molecule of unknown thermal behavior and for which the DSC thermogram is to be predicted, its membership to either one of the seven clusters should rely on its molecular structure.

For this purpose, a decision tree was built with the use of the Marrero-Gani GC^+ framework to represent the structures [Hukkerikar et al., 2012]. The tree is presented in Figure 6.11.

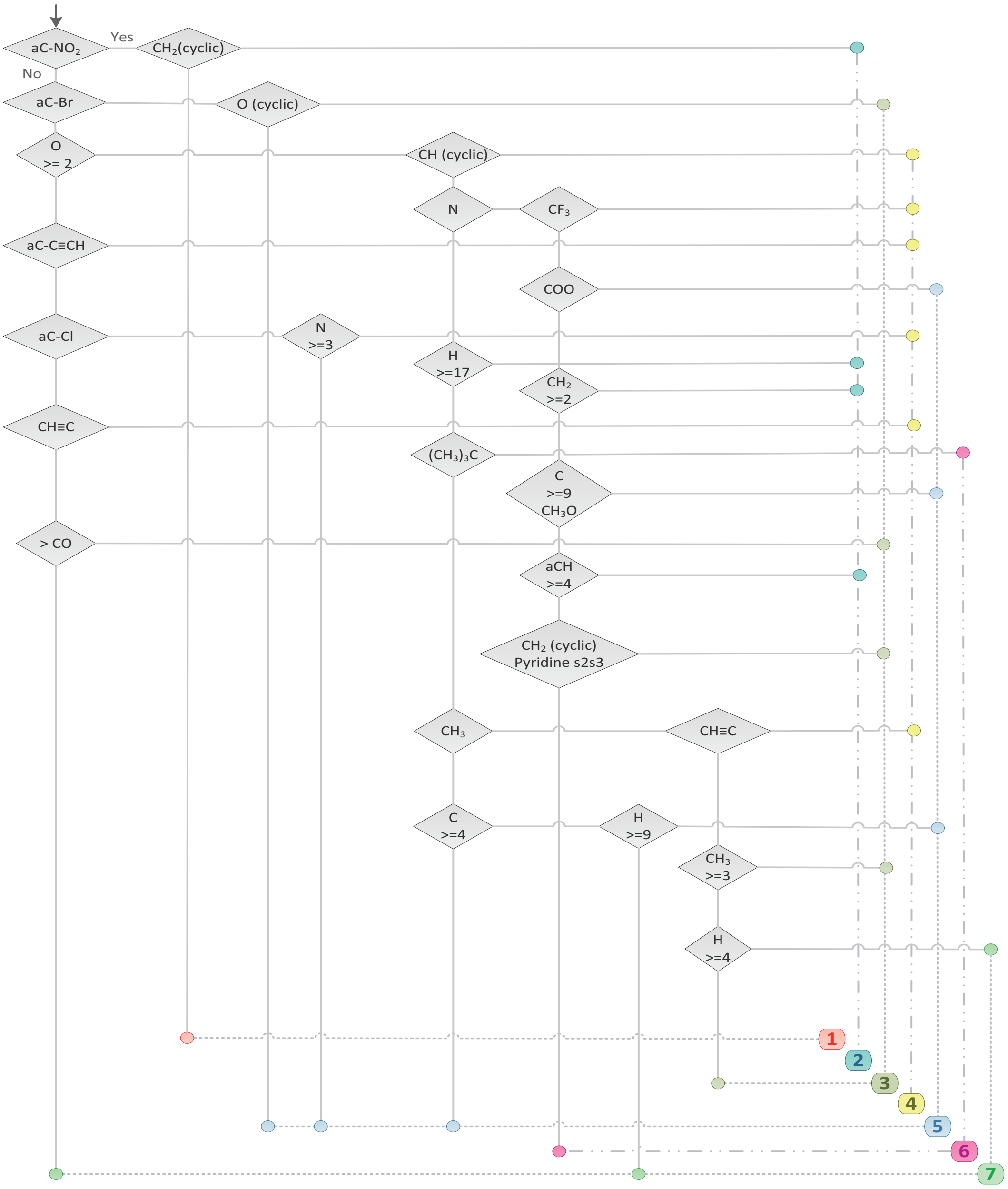


Figure 6.11 – DSC Decision Tree

The performance of the decision tree to assign the DSC curves that did not serve to train the hierarchical clustering was assessed (external validation). For this purpose, the data were assigned to the clusters based on the tree nodes on one hand, and based on the distances to the cluster centroids in the label space, on the other hand. The assignment following the properties is considered the "right" assignment and was compared to the tree outcome. The tree assignment is relatively good in most cases, despite some flaws: it correctly assigns more than 90% of observations destined to cluster 1 (in red in Figure 6.11), but fails to recognize out-of-the-sample data destined to cluster 3 for instance [Mage, 2015].

To prevent from faulty assignment from the tree, the repartition of the dataset is performed with the distance to the cluster centroids in the label space and will now be used to develop local models using GCM modeling. These models are referred to as DSC clusters models as the DSC properties have been used for the clustering.

The modeling is here different than with the QSPR method. The stepwise procedure was not applied as it serves to select among features the most relevant one to build the model, whereas when applying GCM, all groups may eventually contribute to a molecule's properties. Hence the Generalized Linear Model function in Matlab was used in order to assign coefficients to all the groups. Nonetheless, from the 441 groups of the Marrero-Gani framework, 217 are necessary to describe all molecules from the dataset, but not all these 217 groups contribute to each model, nor do they appear in all molecules. In average, one molecule is described by 11 GC^+ groups.

The evaluations of the models built for each cluster are presented in Appendix C.3, Table C.10, and the global evaluation is represented in Table 6.10.

Table 6.10 – DSC Clusters Models Evaluation Summary

Evaluation	Partial Area $\Delta \mathrm{H_r}$	Amplitude Φ_{max}	Max Position $\mathrm{T_{max}}$	Full Width FW	Asymmetry a
R^2_{Tr}	0.756	0.710	0.849	0.736	0.741
R^2_{Val}	0.713	0.526	0.838	0.596	0.490
ARD_{Tr} [%]	166	199	8	39	143
ARD_{Val} [%]	113	103	7	30	160

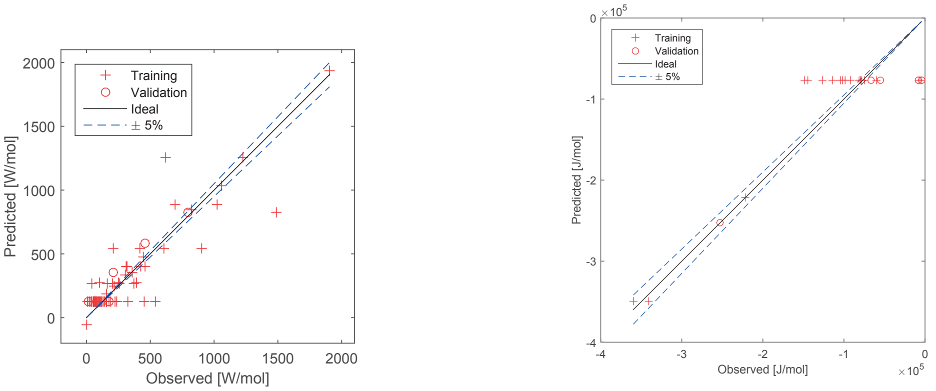
These results reflect a significant improvement relatively to the previous local models proposed as the correlations coefficients are higher and more importantly, the deviations are approximately reduced by half.

Some examples of the obtained responses are shown in Figure 6.12, illustrating the yielded results and their limitations.

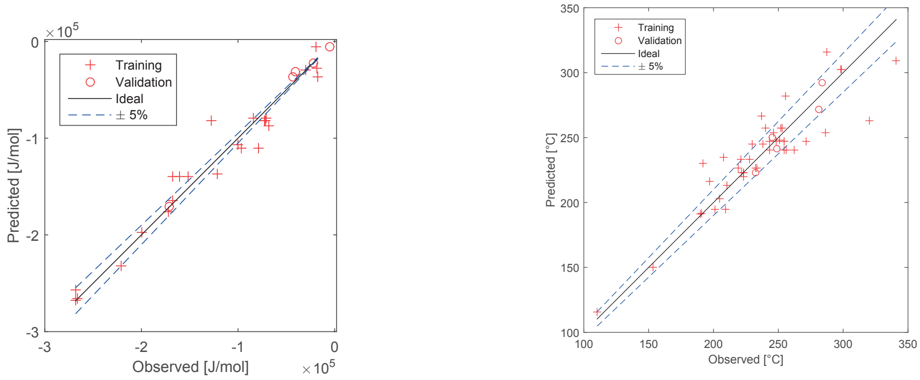
For instance, Figure 6.12 (a) shows the weakest model developed with this procedure. The correlation coefficients are $R^2_{Tr} = 0.671$ and $ARD_{Tr} = 446\%$. However, it exhibits a pattern that is symptomatic of GCM. Indeed, a great number of observations (42 out of 73) with Φ_{max} ranging between 0 and 500 $\mathrm{W\,mol^{-1}}$ are all predicted with the same value $\Phi_{\mathrm{max}} = 128\,\mathrm{W\,mol^{-1}}$, which in fact is the constant term of this model y_o. The underlying reason to these poor predictions is that the model relies on 14 groups, 8 of which only appear in one or two compounds. Hence, 42 molecules in this set are not concerned with any parameter in this model, and therefore their predicted values correspond to the constant term.

The most efficient solution to this issue would be to assign a contribution coefficient to every group in the framework. However, when this was attempted the fitting power of the models were perfect scores $R^2_{Tr} = 1$, while the predictive power and the generalization decreased drastically.

The same type of error appears again in Figure 6.12 (b), where the responses of model $\Delta \mathrm{H_r}$ of DSC Cluster 3 are represented.



(a) Φ_{max} of DSC Cluster 4 **(b)** $\Delta \mathrm{H_r}$ of DSC Cluster 3

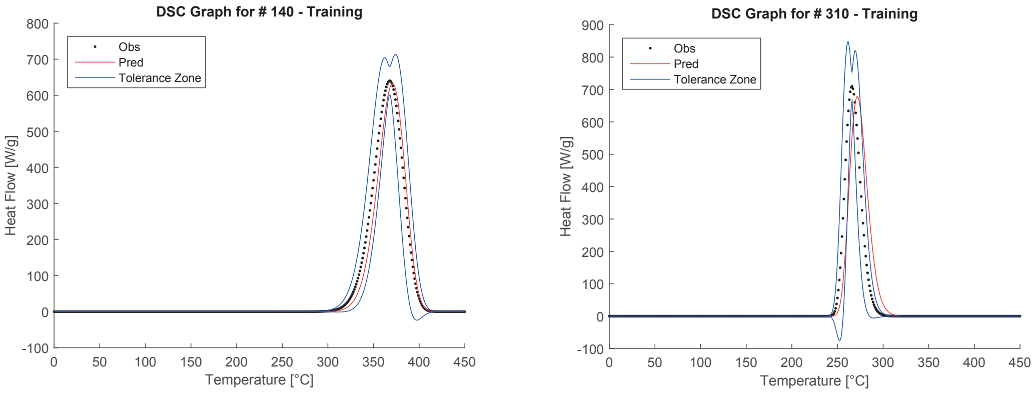


(c) $\Delta \mathrm{H_r}$ of DSC Cluster 6 **(d)** $\mathrm{T_{max}}$ of DSC Cluster 5

Figure 6.12 – Examples of DSC Clusters Models Responses

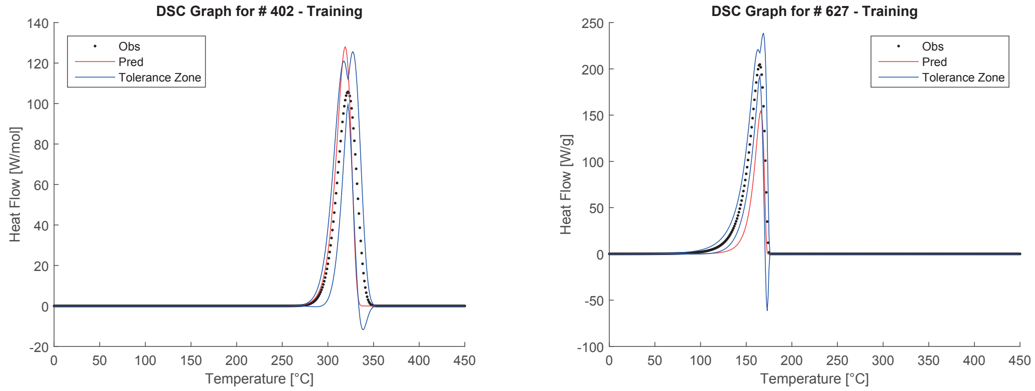
Globally, models are relatively satisfactory and, as the examples shown in Figure 6.12 (c) and (d), the predictions are in good agreement with the observed values. In particular, the models of $\mathrm{T_{max}}$ are highly accurate and the highest average relative deviations are of $ARD_{Tr} = 18\%$ for DSC Cluster 7.

Figure 6.13 shows few examples among the best DSC reconstructions based on simulations performed with the DSC Clusters models. In this figure and in Table **??** (Appendix C.2), the molecules are numbered to show the cluster number in the hundred digits: 2-nitrotoluene belongs to cluster 1 and is the 40^{th} compound in this cluster.



(a) 2-nitrotoluene

(b) 2-bromo-5-(morpholinomethyl)-pyridine



(c) veratrylamine

(d) 2-chloromethyl-pyridine

Figure 6.13 – Examples of DSC Reconstructions from DSC Cluster Models

6.5 Model Application

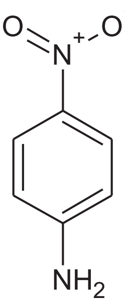
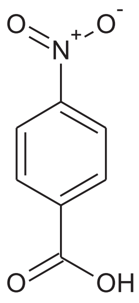
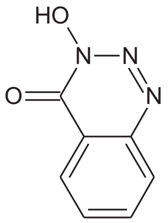
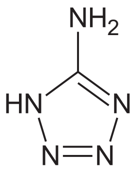
In this section, some examples are detailed to present the overall procedure to output predictive simulations, the obtained results at each step of the entire procedure and to discuss a few elements. Table 6.11 shows the structures of the examples that will be treated here. The chemical families and DSC clusters these molecules belong to are indicated as well.

Assuming only the structures of these 4 compounds are known, the overall procedure to simulate their DSC thermograms is detailed here. The first step is to generate their molecular descriptors for QSPR modeling and to identify and count their GC^+ groups.

A visual inspection of the structure here is sufficient to assign 4-Nitroaniline and 4-Nitrobenzoic acid to the Nitro set, and DHBT and 5-Aminotetrazole to the Azo and Tetrazole set.

Regarding the assignment into DSC clusters, GC^+ groups are required [Hukkerikar et al., 2012].

Table 6.11 – Molecules treated as Examples

	A	B	C	D
Compound	4-Nitroaniline	4-Nitrobenzoic acid	DHBT*	5-Aminotetrazole
Chemical Family	Nitro	Nitro	Azo and Tetrazoles	
# in set	16	17	2	9
DSC Cluster	1	1	2	5
# in set	146	147	202	544
Structure				

*DHBT stands for 3,4-Dihydro-3-hydroxy-4-oxo-1,2,3-benzotriazine

Table 6.12 shows in parallel the structure of DHBT as expressed in function of GC^{+} groups[1] and the answers to all nodes in the decision tree that concern this molecule. Eventually, DHBT is assigned in Cluster 2.

At this point of the procedure, it is already possible to obtain first estimations of the thermal stability of these compounds. Table C.1 in Appendix C.1 reports the average values of the characteristics of the DSC thermograms of all chemical families investigated. Thus, the first estimations for the ΔH_r and T_{max} are:

- $\overline{\Delta H_r} = -1525\,\mathrm{J\,g^{-1}}$ and $\overline{T}_{max} = 293\,°\mathrm{C}$ for Nitro compounds (4-Nitroaniline and 4-Nitrobenzoic acid)
- $\overline{\Delta H_r} = -993\,\mathrm{J\,g^{-1}}$ and $\overline{T}_{max} = 231\,°\mathrm{C}$ for Azo and Tetrazole (DHBT and 5-Aminotetrazole).

In the same manner, the average characteristics for each cluster are also known and gathered in Figure C.6, Appendix C.3.

For instance, from the decision tree, DHBT was assigned to cluster 2, so its DSC thermogram may by approximated by the thermogram shown in Figure 6.14.

---

[1] The GC^{+} groups were numbered in order to show the group order in the hundred digits, and 4^{th} order refers to atom counts and connectivity indices.

Table 6.12 – Example of DSC Cluster Assignment

DHBT Structure			Decision Tree Path	
#	Value	Group	Node	Answer
1015	4	aCH	$aC-NO2$	No
1017	2	aC fused with non aromatic ring	$aC-Br$	No
1029	1	OH	$O \geq 2$	Yes
1176	1	N (cyclic)	CH (cyclic)	No
1180	1	CO (cyclic)	N	Yes
1195	1	$N=N$	CF3	No
3032	1	$aC-$COcyc(fused rings)	COO	No
4002	5	H	$CH2 \geq 2$	Yes
4007	3	N		
4008	2	O		
4011	7	C		
4016	3.54	${}^0\chi$		
4017	1.02	${}^1\chi$		

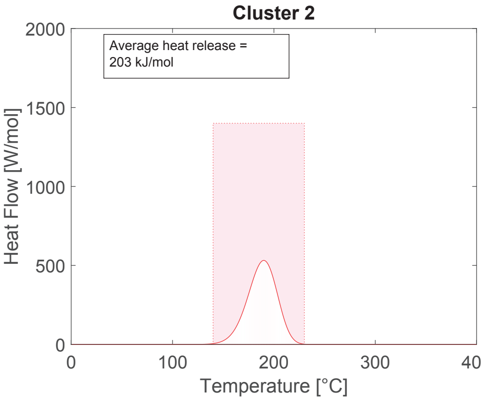


Figure 6.14 – Average Thermogram for Cluster 2

If an average molecular weight of $M = 200\,\mathrm{g\,mol^{-1}}$ is assumed, the $\overline{\Delta H_r}$ obtained with Table C.1 and Figure 6.14 are sensibly the same, i.e. $\overline{\Delta H_r} = -203\,\mathrm{kJ\,mol^{-1}} \approx -1000\,\mathrm{J\,g^{-1}}$.

- $\overline{\Delta H_r} = -1445\,\mathrm{J\,g^{-1}}$ and $\overline{T}_{max} = 343\,°\mathrm{C}$ for Cluster 1 (4-Nitroaniline and 4-Nitrobenzoic acid)
- $\overline{\Delta H_r} = -1015\,\mathrm{J\,g^{-1}}$ and $\overline{T}_{max} = 190\,°\mathrm{C}$ for Cluster 2 (DHBT)
- $\overline{\Delta H_r} = -380\,\mathrm{J\,g^{-1}}$ and $\overline{T}_{max} = 341\,°\mathrm{C}$ for Cluster 5 (5-Aminotetrazole)

From the average characteristics of local subsets, either families or clusters, it is possible to propose a primary risk assessment regarding the studied compounds. This assessment is approximate, but could be a valuable indication to determine the following actions to take.

Thermal risk assessment is performed through a systematic procedure in six steps [Stoessel, 2008], known as the cooling failure scenario. The six steps to develop the cooling failure scenario can be summarized as:

1. Establish the operating conditions necessary for safely conducting a desired reaction. This mainly requires to set the process temperature in function of the reaction heat release rate and the heat removal rate by the cooling system.
2. Evaluate the highest temperature that would be reached in case of cooling failure, by estimating that the desired reaction would follow an adiabatic course (noted MTSR for Maximum Temperature of Synthesis Reaction).
3. Assess if a secondary reaction could be triggered in case MTSR is reached. If yes, evaluate the highest temperature that would be reached by the secondary reaction adiabatic course.
4. Identify the worst moment for the cooling failure to occur, i.e. often when the concentration in potentially reactive chemicals is the highest.
5. Determine the time required to go from the process temperature T_p to MTSR. This requires to know the kinetics of the synthesis reaction, and if unknown, it is approximated to be instantaneous (conservative hypothesis).
6. Evaluate the kinetics of the secondary reaction through the Time to Maximum Rate under adiabatic conditions $\mathrm{TMR}_{\mathrm{AD}}$ as:

$$TMR_{AD} = \frac{C \cdot R \cdot T_o^2}{\Phi_{(T_o)} \cdot E_a} \tag{6.3}$$

where :	C	$\mathrm{J\,g^{-1}\,K^{-1}}$	specific heat capacity
	R	$\mathrm{J\,mol^{-1}\,K^{-1}}$	universal gas constant
	$\mathrm{T_o}$	K	starting temperature from which $\mathrm{TMR}_{\mathrm{AD}}$ is calculated
	$\Phi_{\mathrm{T_o}}$	$\mathrm{W\,g^{-1}}$	heat release rate at $\mathrm{T_o}$
	$\mathrm{E_a}$	$\mathrm{J\,mol^{-1}}$	activation energy

Usually in risk assessment, the risk is evaluated through severity and the probability of an undesired event. For the evaluation of the thermal risk, the probability may be evaluated through the $\mathrm{TMR}_{\mathrm{AD}}$. Strictly speaking, $\mathrm{TMR}_{\mathrm{AD}}$ does not reflect the probability of a cooling failure. However, it indicates how much time the reaction would take to reach its highest rate, if a cooling failure would occur at $\mathrm{T_o}$. For controlled systems, it is considered that if the

$\mathrm{TMR_{AD}}$ is 24h or more, the operating conditions may be considered on the safe side. In case of a cooling failure, there is enough time available to react and take the appropriate response. If $\mathrm{TMR_{AD}}$ is comprised between 8 and 24h the situation is assigned a "Medium" probability, and finally below $\mathrm{TMR_{AD}}$ = 8h, the probability is "High".

The $\mathrm{T_{D24}}$ is the counterpart of the $\mathrm{TMR_{AD}}$: it is the temperature at which $\mathrm{TMR_{AD}}$= 24h.

Regarding the severity, it is assessed through the temperature rise that a loss of temperature control could cause. By assuming that the reaction proceeds under adiabatic conditions, all heat released by the reaction will serve to self-heat the reaction mass, hence the temperature increase $\Delta\mathrm{T_{AD}}$ which may be expressed as:

$$\Delta T_{AD} = \frac{\Delta H_{r,tot}}{C} \tag{6.4}$$

After this short digression, it is now possible to use the results of our primary DSC approximations to average thermograms for chemical families and clusters to compute estimations of $\Delta\mathrm{T_{AD}}$ and $\mathrm{T_{D24}}$.

In the present case, the DSC thermograms represent decompositions reactions of pure compounds, hence from the exposed procedure above, only the information related the "secondary reaction" or decomposition reactions apply. Moreover, one peak only is considered in each thermogram, hence, $\Delta\mathrm{H_{r,tot}}$ is simply $\Delta\mathrm{H_r}$. The other unknown parameters were simulated with standard values as:

R	constant	$\mathrm{R = 8.31\,J\,mol^{-1}\,K^{-1}}$
C	varied	$\mathrm{C = 1.5\ ,\ 2\ and\ 2.5\,J\,g^{-1}\,K^{-1}}$
$\mathrm{E_a}$	varied	$\mathrm{E_a}$ = 50, 100 and 150 $\mathrm{J\,mol^{-1}}$
$\Phi_{\mathrm{T_o}}$	determined from the average thermograms	
$\Delta\mathrm{T_{AD}}$	approximated by $\Delta\mathrm{H_r}$ as C is the same	

The obtained ranges may be represented in a risk matrix to visualize the risk in terms of severity and probability. For the four chemicals serving as examples here, the Nitro and Azo and Tetrazoles families, and Clusters 1, 2 and 5 are placed on the risk matrix shown in Figure 6.15.

All the chemical families and clusters of the examples are in the higher parts of the risk matrix, as the heat release potential is high. This is not surprising as the chosen molecules belong to the chemical families with the highest average $\Delta\mathrm{H_r}$. However, the $\mathrm{T_{D24}}$ evaluation shows some variations. For instance, the DSC Cluster 1, despite showing the highest heat release potential, is not categorized as presenting "High Severity x High Probability" risks, due to the temperature of occurrence of the decompositions reactions that are relatively high and this leads to an estimated $\mathrm{T_{D24}}$ above 200 °C. On the other hand, Cluster 2 for instance, exhibits

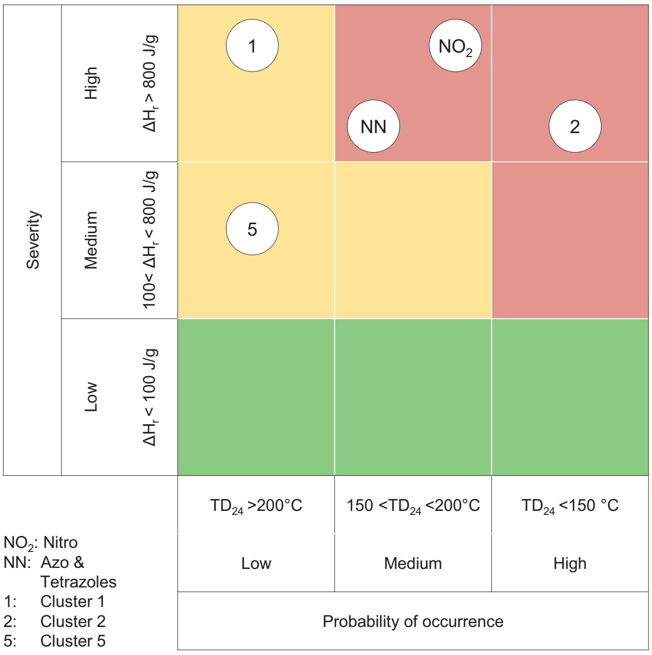


Figure 6.15 – Risk Matrix

lower average T_{max} ($\overline{T}_{max}$ = 190°C see Figure 6.14), hence the T_{D24} estimation reveals a more critical situation and therefore, compounds of Cluster 2 are ranked in "High Severity x High Probability" category.

Considering these results are only indicative approximations, the procedure should continue with the application of the local models to compute more accurate simulations of the DSC thermograms.

Tables 6.13 details the calculations of ΔH_r for Nitro Azo and Tetrazoles families, with the values of descriptors to represent the structures of the four examples treated here. In Appendix B.2, parameters of all models developed are tabulated and the same calculations as presented here can be performed to estimate the five DSC key characteristics of the compounds of interest. The obtained responses are also represented in Tables B.2, B.4, **??** and **??**) in Appendices 6.6 and B.2 .

So far, the results were only discussed from a modeling perspective mainly because they could not be validated properly or because they were relatively unstable and that different run would present different models. Nonetheless, the examples chosen here are among the best results obtained as they present good repeatability and remain stable with iterations. Thus, it is possible here to have a closer look at the model's parameters and their influence.

For instance in Table 6.13, the coefficients in the linear model point out the QSPR descriptors

Table 6.13 – Examples of Enthalpy Calculation with Chemical Family Models

ΔH_r for Nitro Compounds				
i	Coefficient	Parameter	4-Nitroaniline	4-Nitrobenzoic acid
0	1.95×10^5	y_o		
1	1.59×10^4	$AB_{MO,max}$	-2.34	-2.31
2	7.70×10^3	$BO_{C,avg}$	1.17	1.00
3	-1.55×10^4	$E_{R,max(HC)}$	11.1	11.0
4	-3.37×10^3	$FHACA$	0.156	0.188
5	3.99×10^3	S^{γ}_{XZ}	0.756	0.790
Pred		ΔH_r [J g^{-1}]	-2055	-2157
Obs		ΔH_r [J g^{-1}]	-2032	-2143
Rel. Dev.		RD [%]	1.1	0.7
ΔH_r for Azo and Tetrazoles				
i	Coefficient	Parameter	DHBT	5-Aminotetrazole
0	-3.09×10^2	y_o		
1	-4.90×10^3	$FHASA$	0.348	0.559
2	-3.62×10^{-1}	$DPSA-2$	434	171
3	1.31×10^2	$E_{C,tot}/N$	7.37	8.01
Pred		ΔH_r [J g^{-1}]	-1205	-2063
Obs		ΔH_r [J g^{-1}]	-1143	-2033
Rel. Dev.		RD [%]	5.4	1.5

that positively and negatively contribute to ΔH_r: higher bond orders for C atoms $BO_{C,avg}$ and higher antibonding molecular orbital contributions $AB_{MO,max}$ lead to higher ΔH_r, while the highest resonance energy of $C-H$ bonds $E_{R,max(HC)}$ contributes to decrease ΔH_r.

The bond order of C atoms is known to be correlated to the dissociation energies of the $-NO_2$ group in nitro-aromatic compounds [Fayet et al., 2009], and the resonance energies could be related to the molecule's ability to stabilize the $-NO_2$. The influences of $FHACA$ and S^{γ}_{XZ} are relatively limited. They are assigned the smallest coefficients, but also they take values that slightly vary across the dataset of Nitro compounds. These two parameters could serve as adjustements to the model rather than actual descriptors of the molecule's behavior. In this sense, it is understandable why local models would suffer from potential overfitting: if parameters are included in the model to satisfy the fitting to the particular set of observations , it faces higher risks of failure to generalize.

Regarding the Azo and Tetrazole family, the ΔH_r model comprises three parameters related to the molecules polarity and inter-molecular interactions. Electrostatic interactions are accounted for in $E_{C,tot}/N$, which is the only parameter with a positive contribution to ΔH_r, while the other two have negative contributions (thus lead to higher enthalpy release). $DPSA-2$ is the difference between total charge weighted partial positive and negative surface areas. The Charged Partial Surface Areas (CPSA) are a class of descriptors that encode for information relative to solvent-accessible surface area, partial charges and polar interactions [Golmohammadi and Dashtbozorgi, 2010]. Finally, $FHASA$ (Fractional H-acceptors surface area) has the largest influence on this model. It reveals the ability of the molecule to make hydrogen-bonds, which contribute to higher intra- and inter-molecular interactions and increase the potential heat release of decompositions reactions.

Tables 6.14 presents the calculations of $\Delta\mathrm{H_r}$ for Cluster 1, 2 and 5 based on group contributions.

In Table 6.14, the molecular weight of the compounds are also reported, in order to convert $\Delta\mathrm{H_r}$ into comparable units. Indeed, the development of the models for the various families was conducted on $\Delta\mathrm{H_r}$ expressed in $[\mathrm{J\,g^{-1}}]$ and this was changed in the modeling based on DSC clusters to $[\mathrm{J\,mol^{-1}}]$. The results are equivalent, however the coefficients are adjusted to output values in the corresponding units.

The GCM models are more straightforward to interpret than QSPR models. In Cluster 1, which comprises most nitroaromatic compounds of the overall set, it is not surprising that the number of $\mathrm{NO_2}$ groups on an aromatic C (group #1080 $\mathrm{aC-NO2}$) is the most influential parameter of the model. Similarly, in $\Delta\mathrm{H_r}$ model for Cluster 2 on the atom count of C, H and O, but also on the presence of an aromatic ring (accounted for by aCH, and the presence of a ketone group on a cyclic C (group #1180 CO (cyclic)[2]). Finally, $\Delta\mathrm{H_r}$ of 5-Aminotetrazole is computed with only the number of $\mathrm{NH2}$[3], and is rather accurate.

It is important to note that the cluster models include more groups than those represented here, but in Table 6.14, only groups relevant to the examples are represented. For instance, there are 9 parameters in $\Delta\mathrm{H_r}$ model for Cluster 5, but only one is necessary (or relevant) for 5-Aminotetrazole.

Once all DSC key characteristics have been estimated, it is possible to simulate the DSC thermograms. In Figure 6.16, the DSC simulations obtained with the family models and the cluster models (marked as QSPR or GCM in the figure) are compared to the average thermogram construction for the corresponding cluster.

The DSC simulations of 4-nitroaniline, 4-nitrobenzenzoic acid and DHBT show good agreement and match with the projected average of the clusters. In the context of predictive application, it would not be possible to compare to the DSC measurements to verify the adequacy of these predictions, and therefore these mutual agreement confirm the results and reinforce their reliability. Nonetheless, the comparison to the actual measurements are shown in Figure C.14, Appendix C.3, and indeed, these three simulations actually do represent accurately the DSC measurements.

---

[2] 'cyclic' refers to non-aromatic closed structures in GC^+ framework

[3] in GC^+ framework, groups may be defined as a substituent on a another group, and when it is not the case, they are marked with the mention "except as above", for instance the group preceding " $\mathrm{NH2}$ except as above" is $\mathrm{aC-NH2}$.

Table 6.14 – Examples of Enthalpy Calculation with Cluster Models

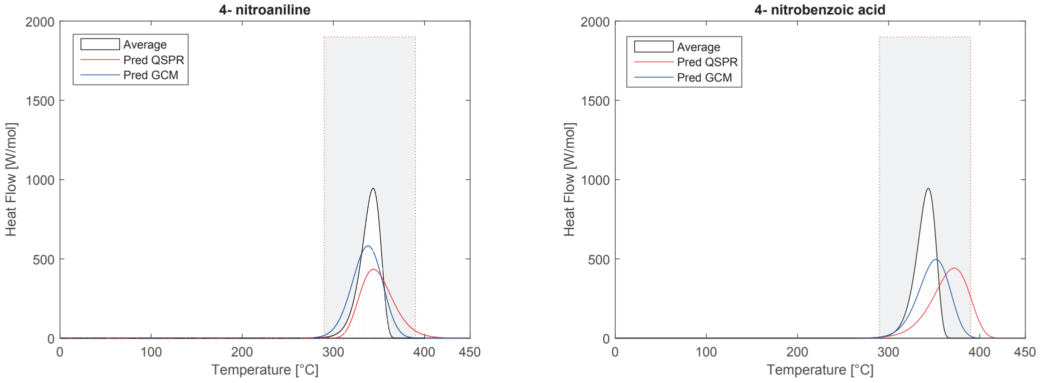
ΔH_r Cluster 1				
i	Coefficient	Parameter	4-Nitroaniline	4-Nitrobenzoic acid
0	-2.28×10^5	y_o		
1080	-2.22×10^5	$aC-NO2$	1	1
4007	-4.27×10^4	N	2	1
4011	2.44×10^4	C	6	7
		M [$g\,mol^{-1}$]	138	167
Pred		ΔH_r [$J\,mol^{-1}$]	-3.89×10^5	-3.22×10^5
Pred		ΔH_r [$J\,g^{-1}$]	-2814	-1924
Obs		ΔH_r [$J\,g^{-1}$]	-2032	-2143
Rel. Dev.		RD [%]	38	10.2

ΔH_r Cluster 2			
i	Coefficient	Parameter	DHBT
0	-2.34×10^5	y_o	
1015	-8.75×10^3	aCH	4
1180	-1.36×10^5	CO (cyclic)	1
4002	1.72×10^4	H	5
4007	7.46×10^4	N	3
4008	-4.31×10^4	O	2
		M [$g\,mol^{-1}$]	163
Pred		ΔH_r [$J\,mol^{-1}$]	-1.82×10^5
Pred		ΔH_r [$J\,g^{-1}$]	-1117
Obs		ΔH_r [$J\,g^{-1}$]	-1143
Rel. Dev.		RD [%]	2.3

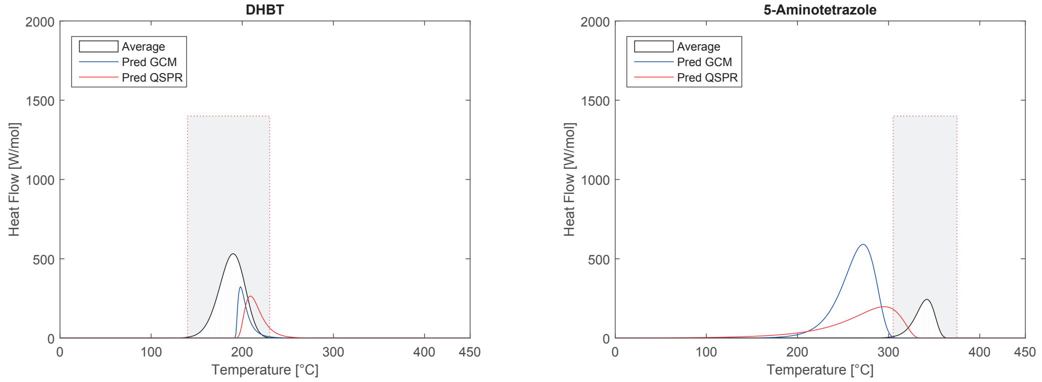
ΔH_r Cluster 5			
i	Coefficient	Parameter	5-Aminotetrazole
0	-6.94×10^4	y_o	
1065	-9.23×10^4	NH2 except as above	1
		M [$g\,mol^{-1}$]	85
Pred		ΔH_r [$J\,mol^{-1}$]	-1.62×10^5
Pred		ΔH_r [$J\,g^{-1}$]	-1900
Obs		ΔH_r [$J\,g^{-1}$]	-2063
Rel. Dev.		RD [%]	6.5

Concerning the 5-aminotetrazole, the two models and the average show divergent results. In such cases, the local models are more reliable than the average projection. Especially in this particular case, as they predict decomposition at a lower temperature and with higher energy release, and hence represent a worse scenario than the projected average. In the case of predictive application of the models, such a result would strongly suggest to take precautions in the use of these simulations and preferably to perform an experiment to obtain the DSC thermogram of this compound.

In any case, when possible the experimental measurement should be favored; however the DSC simulations obtained from the various models developed here could allow for primary estimations, screening of numerous compounds, and a guiding tool in experimental planning.



(a) 4-nitroaniline **(b)** 4-nitrobenzenzoic acid



(c) DHBT **(d)** 5-aminotetrazole

Figure 6.16 – Examples of DSC Reconstructions

6.6 Conclusion

In this chapter, a short literature review showed that regarding thermal stability of chemicals, there were mainly two properties that were studied, the reaction enthalpy ΔH_r and the onset temperature T_o. Besides, most studies focused on structurally similar compounds, most often that belong to a unique chemical family, and often, it would be nitro compounds.

In order to challenge the DSC reconstruction method based on the Fraser-Suzuki equation, it has been decided to perform the initial modeling attempts on nitro compounds, which successfully provided the first set of predictive local models. Then, a second modeling phase integrated nitro compounds into a set of structurally diverse chemicals which showed that a defined substructure to all observations is not a requirement to modeling, and that good performing models could be obtained. Yet, in this phase, the correlations were satisfactory, while the deviations were larger than for the local models built with NO_2 compounds only.

From this initial modeling phase, the following steps were to expand both the local and global models: hence with the dataset of 375 compounds, one large global set and five additional chemical families were studied. The chemical families investigated were : Nitroso and Nitrites, Azo and Tetrazoles, Phenylamines, Ethers and Nitriles. The results were mitigated, as in some cases very well performing models were obtained while others due to overfitting issues were unable to predict out-of-the-set data.

Regarding the global model, it was undoubtedly the weakest set of models obtained. The correlations were poor and the deviations very high. Yet, these results could not be imputed to over-fitting problems as the models failed to predict the validation data but also to describe the training data. The structure of the dataset itself was analyzed and this highlighted the fact that the observations were non-uniformly distributed over the considered ranges. This imbalance in data representation, especially the over representation of compounds of small exoenergetic decompositions, hindered the model building process by depreciating the predictions of properties of the compounds with higher decompositions energies or peak amplitude.

This under-performance on the ’one-size-fits-all’ models raised the interest towards the local models, and posed several interrogations regarding whether alternative subsets could serve for developing local models, and if yes, how to determine which one to apply.

To answer these questions, two approaches to systematically subdivide the dataset were implemented. On the one hand, a k-means clustering was performed on the structural feature space, and on the other hand, a hierarchical clustering was conducted on the DSC curves.

For the latter clustering, a classification system was necessary in order to assign the molecules of unknown DSC into the clusters. The classification tree applies to all molecules of the dataset and outputs the membership of the molecule to one of seven DSC Clusters, but also a "typical DSC curve". From this perspective, the decision tree represents a semi-quantitative global model.

The local models developed within the feature space clusters present limited fitting and predictive powers. The parametrization was closely controlled to avoid over-fitting, yet, this lead to poor models. The second set of local models, created from the clustering of the DSC curves, resulted in better-performing results.

In summary, it was not possible to develop a highly efficient regression global predictive model, however, the decision tree presents a large application range and allows to sort out the compounds into 7 categories. At this stage, the corresponding "average DSC curve" may also be used as a rough estimate for the considered compound. Then, the local models within the clusters allow for a quantitative evaluation of the DSC properties, which are then included into the Fraser-Suzuki model to recover the DSC curves. Moreover, the models corresponding to six chemicals families have also been developed and allow for more accurate estimations.

These models would highly benefit from further testings in order to be more comprehensively evaluated, especially to determine their application domains, for instance. Moreover, they could be further challenged and optimized with the inclusion of additional observations.

Finally, considering the results obtained here, it seems that GCM, despite their simplicity could allow for models as efficient, or even better, than QSPR.

Conclusions

For effective design and implementation of preventive and protective measures, rigorous risk assessment is necessary, and this requires thorough characterization of the process and the involved chemicals. For this purpose, number of safety data are required which may be obtained from knowledge, databases, or experimental measures. Ideally, when this information is available at the early stages of a process design, it allows for an optimal integration of safety measures to the system, or to consider possible substitutions of a hazardous chemical with another, with lower intrinsic hazards, or in lower quantities, to enable a simplified control. Timely availability of these data do not only allow for easing the inherently safer design, it also reduces time and resources necessary for postliminary process modifications or corrections.

Moreover, Product Design is nowadays increasingly employed, especially in biotechnology or life sciences fields where drugs, pesticides or food products are created, and selected for their physico-chemical properties or biological activities *in silico* prior to being physically synthesized. As these compounds are not readily available for experimental analysis, only the simulated properties can be estimated until their preparation in sufficient quantities.

In this context, predictive modeling of safety-related data has gained interest and in the recent years, several applications of structure-based modeling focused on explosive characteristics of chemicals such as the flash point, Auto-Ignition Temperature, flammability limits, or explosion constants, and thermal stability, in particular decomposition enthalpies and onset temperature.

Nevertheless, it was noticed that the Minimal Ignition Energy (MIE) of compounds was not modeled, whereas this characteristic of the compounds' sensitivity to ignite relies on a time and material expensive procedure that could highly profit from the use of predictive simulations. We propose several regression models, a global model and local models depending on the compounds physical state at room temperature. From the obtained results, the model developed from the Dusts subset presented the most accurate fittings and predictions, yet they were not fully satisfactory.

Therefore, a global classification model was proposed in the form of a simple decision tree that includes only 4 structural descriptors and allows for classing the data according to their sensitivity into 4 categories corresponding to the British Standard MIE classification. Then, intra-classes models were developed and the predictions were greatly improved as the relative errors obtained were reduced below 50 %. However, a 50 % error on a safety-related criteria is rather unreliable and could not be reasonably applied in safety studies in itself. It could be however interesting to integrate it as a primary screening tool that could narrow the MIE range to investigate, and hence, reduce the number of experimental measurements required to determine the MIE value.

To improve these models, it is highly probable that the underlying hypothesis should be revised. As it has been noted, the MIE value greatly depends on several factors besides the molecular structure, especially the particle size for dusts, the concentration, and the physical state. Here, these parameters were overlooked. The physical state distinction was only disregarded when developing the global model, in order to gather all the available observations in a unique set; they were separated later into different subsets. On the other hand, due to the absence of information regarding the particle size and mixture concentration, these parameters were not taken into account while their influence on the MIE is established. It would complement the models obtained here if information regarding these aspects could be included for a more comprehensive estimation system.

Regarding the thermal stability of chemicals, several studies were found in the literature review that propose predictive models for the decomposition enthalpy ΔH_r and the onset temperature T_o, temperature at which the decomposition reaction progresses at a significant rate. These two characteristics can be determined from DSC experiments. However, we have showed that the DSC thermograms analysis allows to identify the reaction kinetics as well. For this purpose several DSC thermograms are required and the curve shape should be analyzed.

Without deepening the investigations to the kinetics determination, it was nonetheless decided to propose a modeling method that would allow to preserve the DSC curve shape and limit the data loss. Therefore, for the analyzed thermograms, five key DSC properties were extracted by Gaussian-like fittings with the Fraser-Suzuki equation. With the peak's amplitude, position, width and an asymmetry factor, the entire curve is preserved.

In a similar manner to the procedure for the MIE models, one set of global models, and several local models were developed. Once again, the global regression models yielded weak predictions, while local models were more accurate. Therefore, another strategy relying on a combination of global classification and subsequent local models is suggested. This approach offers a systematic classification of chemicals into different categories, and for each category, a tailored set of models could allow to recover the chemicals DSC key properties, which in turn serve the DSC curve reconstruction.

Besides, the nitro compounds were particularly focused on by previous studies, while limited information was found on other compounds. Indeed the nitro compounds exhibit the highest decomposition reactivity, yet, other families also present hazardous thermal behavior. Therefore, specific models were also created for various chemical families, which ranked on top of the list according to decomposition enthalpy released: nitroso, nitrites, azo compounds, tetrazoles, ethers and nitriles. From these modeling phase, some accurate results were obtained, yet the subsets were rather narrow and would certainly benefit from additional data and further validation.

Actually, this would be beneficial to all models proposed here. Despite the broad dataset investigated (approximately 400 compounds), the structural descriptors were also numerous, and the feature selection methods applied did not sufficiently reduce the structure space dimension to avoid over-fitting issues. By increasing the number of observations, the model parametrization could be improved to retain less descriptors, while increasing the quality and the performance of the models.

It is important to note not all the available data has been exploited. For each thermogram that exhibited several peaks, the DSC key properties were all extracted, yet only the principal peak was held for the present work. The procedure should be extended to incorporate the simulations of several peaks per thermogram, in order to complete the thermal trail simulations and actually deliver comprehensive simulations. This could also be done through the construction of a classification system that would recognize from the molecular structures if compounds would exhibit simple or complex thermal decompositions, and if secondary or even tertiary peaks are expected, they could be modeled following a similar procedure to what we developed here. However, from a safety point of view, the "main peaks" selected here correspond either to the highest energy release or the lowest temperature of decomposition, and represent the critically hazardous thermal event the compounds could undergo. Secondary or tertiary peaks would complement this information, without significantly affecting the outcome of the simulation.

A potential perspective to explore would be the modeling of mixtures. Structure-based models conventionally apply to single molecules, yet, in practice compounds are rarely found in pure composition, unless for storage or transportation of raw materials and final products. For all other operations, especially reactions, reactants are mixed or diluted in solvents and by-products are also present. Therefore, after analyzing the thermal stability, safety studies would consider the compound in its matrix. To extend QSPR methods to mixtures, the "mixing rules" apply [Nieto-Draghi et al., 2015]. Some properties were modeled with such methods as for instance the application of Peng-Robinson equation of state model for binary mixtures [Jaubert and Mutelet, 2004, Peng and Robinson, 1976] or the density of mixtures [Ajmani et al., 2006]. Besides the synergetic effects and interactions that arise in multiple components mixtures, the molar fractions should also be accounted for.

Simulating safety data from the molecular structure offers several advantageous applications in process design. Besides the possibility to predict characteristics which cannot be experimentally measured, there is also a time benefit. Predictions can be made at a very early stage of the process design, so that hazardous behavior could already be anticipated. Moreover, simulations can allow analyzing several alternatives within limited resources, saving them for considering potential substitution of a hazardous compound by a less hazardous one or modification of the process. It is also noteworthy that predictive models help avoiding expendable handling of harmful chemicals.

We propose a method relying on molecular-based approaches to predict an explosive sensitivity evaluation, MIE, and thermal stability through DSC simulations to identify thermal threats without necessarily facing them. In both cases, a combination of global classification and local regressions models are proposed to obtain approximate estimations that are refined to more accurate predictions.

It is important to stress that predictive models should be handled with precaution when applied to sensitive data such as safety related information. They are also limited to pure compounds, whereas matrix should not be disregarded. Thus they are not intended to replace proper experimental investigations, but rather be a helpful tool that allows focusing the experimental work on the most critical compounds. The major benefits of such procedures within process design context are mainly to broaden the number of evaluations within given time and resources, an efficiency gain in testing phase with better resource allocation and valuable timing leading to anticipation.

Appendices

Minimal Ignition Energies Data

Table A.1 – MIE Values for Gaseous Compounds

#	Name	T_b	MIE [mJ]	Reference
1	1,3-butadiene	-4	0.13	c
2	2,2-dimethylpropane	9	1.6	a
3	butane	-1	0.26	i
4	cyclopropane	-33	0.18	i
5	dimethyl amine	7	0.30	g
6	ethane	-89	0.26	i
7	ethyl chloride	12	0.30	g
8	ethyl nitrite	17	0.17	a
9	ethylamine	16	2.4	a
10	ethylene	-103	0.07	i
11	ethylene oxide	11	0.06	i
12	isobutane	-13	0.52	a
13	methylacetylene	-23	0.12	c
14	methylether	-24	0.29	c
15	propane	-42	0.26	i
16	propylene	-47	0.28	a
17	toluene	11	0.24	i
18	vinyl acetylene	6	0.08	a
19	vinyl chloride	-13	0.30	g

Table A.2 – MIE values for liquid compounds

#	Name	T_m	T_b	MIE [mJ]	Reference
20	1,3-cyclopentadiene	-90	40	0.67	a
21	1-heptyne	-80	100	0.56	a
22	2,2,3-trimethylbutane	-26	81	1.0	a
23	2,2-dimethylbutane	-100	50	1.6	a
24	2,3-butadione	-2	88	0.41	a
25	2-pentene	-165	30	0.18	g
26	2-propanol	-89	83	0.65	a
27	acetaldehyde	-123	20	0.38	a
28	acetone	-95	56	1.2	a
29	acetonitrile	-45	81	2.8	a
30	acrolein	-88	53	0.13	a
31	acrylonitrile	-84	77	0.16	c
32	allyl chloride	-135	45	0.78	a
33	alpha-pinene	-64	155	1.4	a
34	aziridine	-78	56	0.48	g
35	benzene	6	80	0.22	c
36	cyclohexane	7	81	0.22	c
37	cyclohexene	-104	83	0.53	a
38	cyclohexene oxide	-40	130	0.74	a
39	cyclopentane	-94	49	0.24	i
40	diethyl ether	-116	35	0.20	i
41	dihydropyran	-70	86	0.36	a
42	diisobutylene	-94	101	0.96	a
43	dimethoxymethane	-105	42	0.42	a
44	dimethyl sulfide	-98	35	0.48	a
45	dioxane	12	101	0.30	g
46	di-tert-butyl peroxide	-40	109	0.41	a
47	epichlorohydrin	-25	118	0.29	a
48	ethyl acetate	-84	77	1.4	c
49	furan	-86	31	0.23	a
50	heptane	-91	98	0.70	a
51	hexane	-96	68	0.29	i
52	iso-octane	-107	99	1.4	a
53	isopentane	-161	28	0.25	i
54	isopropyl alcohol	-89	83	0.65	c
55	isopropyl chloride	-117	35	1.1	i
56	isopropyl ether	-60	68	1.1	a
57	isopropyl mercaptan	-131	57	0.53	a

Table A.2 – MIE Values for Liquid Compounds (continued)

#	Name	T_m	T_b	MIE [mJ]	Reference
58	isopropylamine	-95	31	2.0	a
59	methanol	-98	65	0.14	c
60	methylcyclohexane	-126	101	0.27	c
61	methylethyl ketone	-86	80	0.53	c
62	methylformate	-100	32	0.40	a
63	m-xylene	-48	139	0.20	i
64	n-butyl chloride	-123	78	0.33	i
65	nitroethane	-90	112	0.22	a
66	n-propyl chloride	-128	47	1.1	a
67	o-xylene	-25	144	0.20	i
68	pentane	-130	36	0.51	a
69	propargyl alcohol	-51	114	0.21	a
70	propionaldehyde	-81	46	0.33	a
71	propylene oxide	-112	34	0.14	i
72	p-xylene	13	138	0.20	i
73	pyrrole	-23	129	1.7	a
74	tetrafluoroethylene	-142	131	3.5	i
75	tetrahydrofuran	-108	66	0.54	a
76	tetrahydropyran	-45	88	0.22	c
77	thiophene	-38	84	0.39	a
78	trichloroethylene	-73	87	295	i
79	triethyl amine	-115	90	1.2	a
80	vinyl acetate	-93	73	0.70	a

Table A.3 – MIE Values for Solid Compounds

#	Name	T_b	$C_{Ex,min}$ [g/L]	MIE [mJ]	Reference
81	1,3-bis(4-nitrophenyl)urea	240	0.095	60	h
82	2,4-dichlophenoxy ethyl benzoate	66	0.045	60	d
83	2-acetylamino5-nitrothiazole	263	0.16	40	d
84	2-amino-5-nitrothiazole	195	0.075	30	d
85	4-chloro-2-nitro aniline	116	0.75	140	d
86	a,a’-azo isobutyronitrile	97		25	f
87	aceto acetanilide	83		20	f
88	adipic acid	152		60	c
89	anthranilic acid	146	0.030	35	d
90	ascorbic acid	190	0.070	60	d
91	aspirin	136		16	e
92	azelaic acid	109		25	f
93	benzoic acid	122		12	e
94	benzotriazole	100	0.030	30	d
95	benzoyl peroxide	103		21	e
96	bis(2-hydroxy-5-chlorophenyl)-methane	177	0.040	60	d
97	caprolactam	68		60	e
98	cyclohexanone peroxide	76		21	e
99	dehydroacetic acid	109	0.030	15	d
100	diazo amino benzene	96	0.015	20	d
101	dicyclopentadiene dioxide	185		30	f
102	dimethyl isophtalate	61		15	f
103	dimethyl terephtalate	142		20	f
104	dinitrobenzamide	183	0.040	45	d, h
105	dinitrobenzoic acid	204	0.050	45	d, h
106	dinitrotoluamide	177		15	h
107	diphenyl	69.2	0.065	20	d
108	di-t-butyl p-cresol	69		15	f
109	DL methionine	281		35	d
110	ethylenediaminetetraacetic acid	248	0.075	50	d
111	fumaric acid	287		35	f
112	hexamethylenetetramine	200		10	c
113	isatoic anhydride	235		25	d
114	isophtalic acid	300		25	f
115	lauryl peroxide	53		12	e

Table A.3 – MIE Values for Solid Compounds (continued)

#	Name	T_b	$C_{Ex,min}$ [g/L]	MIE [mJ]	Reference
116	l-sorbose	163	0.065	80	d
117	mannitol	166	0.065	40	d
118	methylamino anthraquinone	170		50	d
119	nitropyridone	285	0.045	35	d
120	o-chloroaceto acetanilide	141	0.035	30	d
121	p-chloroaceto acetanilide	131	0.035	20	d
122	pentaerythritol	260		10	c
123	phosphorus pentasulphide	288	0.050	15	d
124	phtalimide	234	0.030	50	d
125	phthalic anhydride	131	0.015	15	c
126	phytosterol	135	0.025	10	d
127	p-phenylene diamine	145		30	f
128	salicylanilide	136		20	d
129	sorbic acid	135	0.020	15	d, e
130	stearic acid	70		25	d
131	t-butyl benzoic acid	168		25	f
132	terephtalic acid	300		20	f
133	trinitrotoluene	80		75	h

Table A.4 – MIE Values Sources

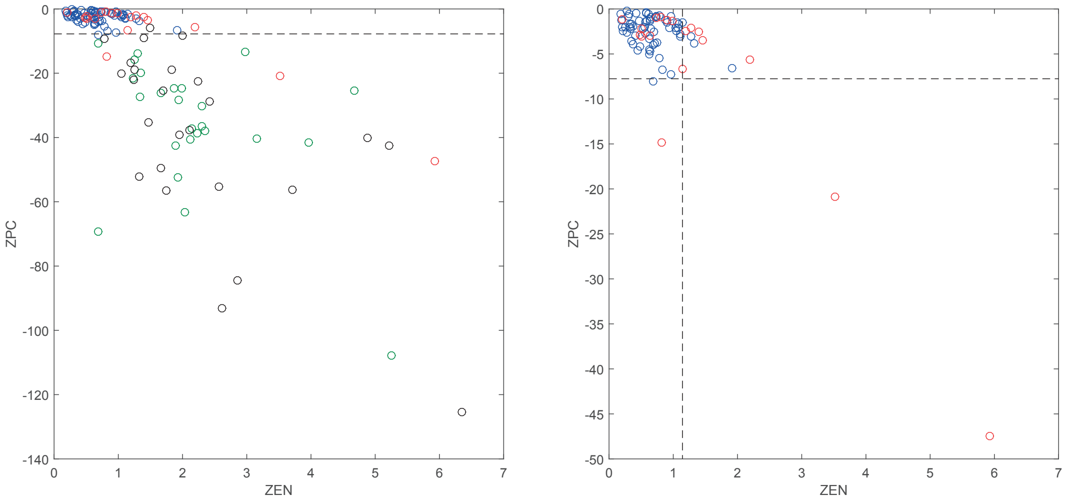
References	
a	Calcote et al. [1952]
b	Calcote et al. [1952]
c	Haase [1977]
d	Cross and Farrer [1982]
e	NFPA [1986]
f	Bartknecht [1989]
g	Berufsgenossenschaften [1992]
h	Hertzberg et al. [1992]
i	Babrauskas [2003]

Table A.5 – Global MIE Parameters

i	Coefficient	Parameter	Name
0	$-3.35 \cdot 10^{2}$	y_o	Intercept
1	$-2.22 \cdot 10^{1}$	N_S	Number of S atoms
2	$1.78 \cdot 10^{1}$	N_{Cl}	Number of Cl atoms
3	$1.54 \cdot 10^{1}$	M_R	Average atom weight
4	-4.15	${}^{1}\chi$	Kier& Hall index (order 1)
5	$-3.53 \cdot 10^{-1}$	${}^{1}BIC$	Bonding Information content (order 1)
6	$-5.39 \cdot 10^{-1}$	ZPC_C	Zefirov's Partial Charges for atom #0000008(C)
7	5.55	$ETS_{P,O}$	Electrotopological state of atom (All pairs, Zefirov's PC) for atom #0000006(O)
8	$2.41 \cdot 10^{-1}$	$ETS_{B,C}$	Electrotopological state of atom (All bonds, Zefirov's PC) for atom #0000011(C)
9	-1.55	$CPSA_{CSA,N}$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000001(N)
10	$4.95 \cdot 10^{1}$	$CPSA_{CSA,C}$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000002(C)
11	-1.69	$SASA_C$	Solvent accessible surface for atom #0000003(C)
12	-3.26	$PNSA1$	PNSA1 Partial negative surface area (Zefirov PC)
13	8.55	$PNSA3$	PNSA3 Atomic charge weighted PNSA (Zefirov PC)
14	$9.54 \cdot 10^{1}$	$FHASA2$	Fractional Area-weighted surface charge of hydrogen bonding acceptor atoms HASA2 (Zefirov PC)
15	$-1.06 \cdot 10^{1}$	$HASA2TS$	Area-weighted surface charge of hydrogen bonding acceptor atoms over square root of Total molecular surface area(Zefirov PC)
16	$3.20 \cdot 10^{1}$	MPC_H	MOPAC Partial Charges for atom #0000014(H)

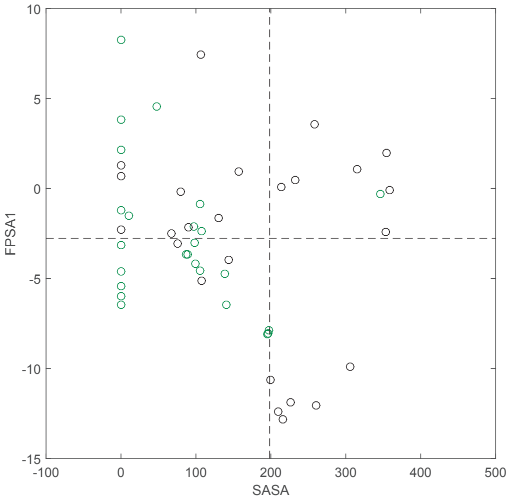
Table A.6 – MIE Dust Model C Parameters

i	Coefficient	Parameter	Name
0	$4.25 \cdot 10^{2}$	y_o	Intercept
1	$-6.14 \cdot 10^{-1}$	^{0}IC	Average Information content (order 0)
2	-1.14	^{0}SIC	Structural Information content (order 0)
3	-1.45	^{1}BIC	Bonding Information content (order 1)
4	$-2.94 \cdot 10^{-1}$	V_M	Molecular volume
5	1.54	S_M	Molecular surface area
6	-3.62	ZEN_C	Sanderson's atomic electronegativities for atom #0000008(C)
7	-1.21	$LOGZEN_C$	Natural logarithm of Sanderson's atomic electronegativities for atom #0000010(C)
8	2.62	$LOGZEN_H$	Natural logarithm of Sanderson's atomic electronegativities for atom #0000012(H)
9	$-4.98 \cdot 10^{-1}$	ZEN_H	Sanderson's atomic electronegativities for atom #0000016(H)
10	3.26	ZPC_C	Zefirov's Partial Charges for atom #0000003(C)
11	2.78	$ZPC_{O,max}$	Max partial charge (Zefirov) for atoms for atom O
12	$3.55 \cdot 10^{-1}$	$ETS_{B,C}$	Electrotopological state of atom (All bonds, Zefirov's PC) for atom #0000003(C)
13	$4.41 \cdot 10^{-1}$	$ETS_{B,C}$	Electrotopological state of atom (All bonds, Zefirov's PC) for atom #0000010(C)
14	1.07	$ETS_{B,C}$	Electrotopological state of atom (All bonds, Zefirov's PC) for atom #0000011(C)
15	1.97	$ETS_{P,H}$	Electrotopological state of atom (All pairs, Zefirov's PC) for atom #0000016(H)
16	7.35	T^{E}_{all}	Topographic electronic index (all bonds)
17	5.63	$SASA_O$	Solvent accessible surface for atom #0000006(O)
18	$-2.69 \cdot 10^{1}$	$CPSA_O$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000006(O)
19	-6.60	$SASA_C$	Solvent accessible surface for atom #0000008(C)
20	$3.05 \cdot 10^{1}$	$CPSA_C$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000008(C)
21	$6.57 \cdot 10^{-1}$	$CPSA_C$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000011(C)
22	-9.85	$SASA_H$	Solvent accessible surface for atom #0000014(H)
23	$-6.01 \cdot 10^{-2}$	$DPSA3$	Difference in CPSAs (PPSA3-PNSA3) (Zefirov PC)
24	$-2.92 \cdot 10^{-2}$	MPC_C	MOPAC Partial Charges for atom #0000003(C)
25	$-6.51 \cdot 10^{-1}$	$MPPC_O$	MOPAC Partial Charges for atom #0000007(O)
26	$4.86 \cdot 10^{-1}$	$FHASA$	Fractional HASA H-acceptor surface area HASA-1/TMSA (HASA/TMSA) (MOPAC PC)
27	$5.56 \cdot 10^{-1}$	BO_{N-O}	MOPAC Bond Orders for bond #0000001(N) - #0000007(O)



(a) First Node

(b) Second Node



(c) Third and Fourth Nodes

Figure A.1 – Graphical Representation of the MIE Classification Tree. Blue: Class 1, Red: Class 2, Green: Class3, Black: Class 4

Table A.7 – MIE Classes Models Parameters

Class	i	Coefficient	Parameter	Name
1	0	-2.51	y_o	Intercept
	1	$2.81 \cdot 10^{-1}$	$N_{SINGLE,R}$	Relative number of single bonds
	2	-1.12	$^{3}\chi$	Randic index (order 3)
	3	$6.93 \cdot 10^{-1}$	$^{3}\chi$	Kier&Hall index (order 3)
	4	$1.71 \cdot 10^{-2}$	^{0}IC	Average Information content (order 0)
	5	$1.11 \cdot 10^{-1}$	S_{XY}	XY Shadow / XY Rectangle
	6	9.67	$LOGZEN$	Natural logarithm of Sanderson's atomic electronegativities for atom #0000006(O)
	7	$-3.68 \cdot 10^{-2}$	$CPSA_{CSA}$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000003(C)
	8	$1.28 \cdot 10^{-1}$	$CPSA_{CSA}$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000012(H)
	9	$-2.31 \cdot 10^{-1}$	$WPSA1$	Weighted Partial positive surface area (Zefirov PC)
	10	-1.27	$FHACAM$	Fractional H-acceptors charged surface area (MOPAC PC)
2	0	$5.83 \cdot 10^{-1}$	y_o	Intercept
	1	8.56	I_C	Moments of inertia C
	2	$4.51 \cdot 10^{-1}$	ZEN	Sanderson's atomic electronegativities for atom #0000007(O)
	3	$-2.09 \cdot 10^{-1}$	ZPC	Zefirov's Partial Charges for atom #0000003(C)
	4	$4.05 \cdot 10^{-1}$	$CPSA_{CSA}$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000001(N)
	5	$7.74 \cdot 10^{-2}$	$SASA$	Solvent accessible surface for atom #0000006(O)
3	0	$4.25 \cdot 10^{1}$	y_o	Intercept
	1	$-8.83 \cdot 10^{2}$	I_C	Moments of inertia C
	2	$-9.87 \cdot 10^{-1}$	^{2}SIC	Average Structural Information content (order 2)
	3	$2.17 \cdot 10^{-2}$	$SASA$	Solvent accessible surface for atom #0000001(N)
	4	$4.72 \cdot 10^{-1}$	$CPSA_{CSA}$	Charge density on solvent accessible surface (Zefirov's PC) for atom #0000012(H)
	5	$2.71 \cdot 10^{1}$	$HDCA2$	HA dependent HDCA-2 (Zefirov PC)
4	0	$6.95 \cdot 10^{2}$	y_o	Intercept
	1	$-1.30 \cdot 10^{1}$	$N_{SINGLE,R}$	Relative number of single bonds
	2	$-2.40 \cdot 10^{2}$	$N_{TRIPLE,R}$	Relative number of triple bonds
	3	$2.75 \cdot 10^{-1}$	$LOGZEN$	Natural logarithm of Sanderson's atomic electronegativities for atom #0000011(C)
	4	$1.62 \cdot 10^{-1}$	ETS_B	Electrotopological state of atom (All bonds, Zefirov's PC) for atom #0000011(C)
	5	$-9.57 \cdot 10^{-1}$	$PPSA1$	Partial positive surface area (Zefirov PC)
	6	$-2.40 \cdot 10^{1}$	$HDCA1$	HA dependent HDCA-1 (Zefirov PC)
	7	$-1.11 \cdot 10^{3}$	$HDCA2T$	HA dependent HDCA-2/TMSA (Zefirov PC)
	8	-1.41	MPC	MOPAC Partial Charges for atom #0000004(C)

Table A.8 – Models Responses

Data				Global Model			Local Models			Classification			Class Models	
#	**Name**	**MIE [mJ]**		**Pred**	**ARD %**		**Pred**	**ARD %**		**Actual**	**Tree**		**Pred**	**ARD %**
1	1,3-butadiene	0.13		-8.30	6484		0.15	13		1	1	v	0.21	63
2	2,2-dimethylpropane	1.57		-8.75	657		1.51	4	v	2	1		1.87	19
3	butane	0.26		0.02	93	v	0.25	5		1	1		0.28	8
4	cyclopropane	0.18		-2.36	1414		0.16	13		1	1		0.19	7
5	dimethyl amine	0.30		6.39	2030		0.34	12		1	1		0.34	14
6	ethane	0.26		-1.49	672	v	0.29	12		1	1		0.39	48
7	ethyl chloride	0.30		15.2	4955		0.42	41		1	1		0.26	13
8	ethyl nitrite	0.17		-5.81	3520		0.18	4		1	1		0.20	18
9	ethylamine	2.40		6.19	158		1.22	49		2	1	v	2.71	13
10	ethylene	0.07		6.97	9856		0.07	1		1	1		0.09	31
11	ethylene oxide	0.06		-8.88	14428		0.10	66		1	1		0.12	95
12	isobutane	0.52	v	-3.43	760		0.45	14		1	1		0.53	1
13	methylacetylene	0.12		11.1	9558		0.14	19		1	1		0.10	13
14	methylether	0.29		-2.13	834		0.17	43		1	1		0.15	50
15	propane	0.26		-5.14	2076	v	0.17	33		1	1	v	0.19	25
16	propylene	0.28		0.51	84		0.26	8		1	1		0.23	17
17	toluene	0.24	v	3.72	1452		0.22	9	v	1	1		0.23	4
18	vinyl acetylene	0.08	v	-6.32	7790		0.08	0		1	1		0.12	41
19	vinyl chloride	0.30		-3.43	1244		0.32	5		1	1		0.28	7
20	1,3-cyclopentadiene	0.67		6.67	896		0.54	19		1	1		0.38	43
21	1-heptyne	0.56		-7.08	1365		0.56	0		1	1		0.44	22
22	2,2,3-trimethylbutane	1.00		-8.23	923		0.78	22		1	1		0.71	29
23	2,2-dimethylbutane	1.64		-2.96	280		0.73	56		2	1		1.42	13
24	2,3-butadione	0.41		8.94	2081		0.42	2		1	1		0.45	11
25	2-pentene	0.18		11.3	6157		0.33	81		1	1		0.34	90
26	2-propanol	0.65		-0.40	162		0.72	11		1	1		0.88	35
27	acetaldehyde	0.38		-4.12	1183		0.62	63		1	1		0.36	6
28	acetone	1.15		-5.36	566		0.65	44		2	1		1.04	10
29	acetonitrile	2.80		1.01	64		1.83	35		2	1		2.78	1
30	acrolein	0.13	v	-3.82	3041		0.10	26		1	1		0.15	16
31	acrylonitrile	0.16		-16.0	10099		0.22	37		1	1		0.23	43
32	allyl chloride	0.78		3.44	342		0.41	47		1	1		0.40	49

Table A.8 – Models Responses (continued)

Data				**Global Model**		**Local Models**			**Classification**			**Class Models**	
#	**Name**	**MIE [mJ]**		**Pred**	**ARD %**	**Pred**	**ARD %**		**Actual**	**Tree**		**Pred**	**ARD %**
33	alpha-pinene	1.40		21.6	1440	1.72	23		2	2		1.24	12
34	aziridine	0.48		-8.16	1801	0.62	30		1	1		0.27	44
35	benzene	0.22		-5.67	2679	0.24	7		1	1		0.18	19
36	cyclohexane	0.22		-4.09	1961	0.38	71		1	1		0.38	71
37	cyclohexene	0.53		7.95	1414	0.31	41		1	1		0.38	27
38	cyclohexene oxide	0.74		-0.94	228	0.48	36		1	1		0.42	43
39	cyclopentane	0.24		-3.46	1541	0.27	14		1	1		0.28	17
40	diethyl ether	0.20		0.82	309	0.26	31		1	1		0.28	39
41	dihydropyran	0.36	v	9.58	2561	0.26	27		1	1		0.57	59
42	diisobutylene	0.96		3.66	281	0.90	6		1	1		0.86	11
43	dimethoxymethane	0.42		11.9	2743	0.29	31		1	2		0.30	29
44	dimethyl sulfide	0.48		17.6	3562	0.48	1		1	1		0.30	38
45	dioxane	0.30		10.5	3397	0.34	12		1	2		0.26	14
46	di-tert-butyl peroxide	0.41		-0.95	331	0.41	1		1	2		0.50	23
47	epichlorohydrin	0.29		21.1	7171	0.27	6	v	1	4	v	0.28	5
48	ethyl acetate	1.42		-5.24	469	1.46	3		2	2	v	0.45	68
49	furan	0.23		-0.41	281	0.22	0		1	1	v	0.33	45
50	heptane	0.70	v	2.36	238	0.53	24		1	1		0.45	36
51	hexane	0.29		7.81	2592	0.29	2		1	1		0.27	5
52	iso-octane	1.35		-5.40	500	1.56	16		2	2	v	1.22	9
53	isopentane	0.25	v	0.48	91	0.32	26		1	1		0.42	66
54	isopropyl alcohol	0.65		3.42	425	0.52	20		1	1		0.53	18
55	isopropyl chloride	1.08		6.70	521	1.25	16		2	1		1.19	10
56	isopropyl ether	1.14		4.54	298	0.87	23	v	2	2		1.19	4
57	isopropyl mercaptan	0.53		-6.77	1378	1.13	113		1	1		0.56	6
58	isopropylamine	2.00		7.56	278	3.64	82		2	1		1.85	7
59	methanol	0.14		-9.78	7089	0.14	2		1	1		0.15	11
60	methylcyclohexane	0.27		7.16	2554	0.39	43		1	1		0.42	55
61	methylethyl ketone	0.53		2.09	293	0.84	58		1	1		0.49	7
62	methylformate	0.40	v	-7.75	2038	0.60	50		1	1		0.51	27
63	m-xylene	0.20		9.19	4495	0.35	76		1	1	v	0.34	72
64	n-butyl chloride	0.33		1.53	365	0.56	70		1	1		0.51	55
65	nitroethane	0.22		1.17	430	0.20	10		1	1		0.22	1

Table A.8 – Models Responses (continued)

Data				Global Model			Local Models			Classification			Class Models	
#	Name	MIE [mJ]		Pred	ARD %		Pred	ARD %		Actual	Tree		Pred	ARD %
66	n-propyl chloride	1.08		12.1	1023		0.74	32		2	1		1.17	8
67	o-xylene	0.20		0.92	358		0.16	22		1	1		0.26	28
68	pentane	0.51		3.00	489		0.35	32	v	1	1		0.38	26
69	propargyl alcohol	0.21		-3.28	1662		0.24	15		1	1		0.21	2
70	propionaldehyde	0.33		-0.49	250		0.50	55		1	1		0.25	25
71	propylene oxide	0.14		-10.5	7566		0.13	7		1	1		0.16	18
72	p-xylene	0.20		4.81	2303		0.18	8		1	1		0.20	1
73	pyrrole	1.70		12.8	651		1.43	16		2	1	v	1.92	13
74	tetrafluoroethylene	3.5		6.71	92		3.95	13		2	3		3.46	1
75	tetrahydrofuran	0.54		10.6	1871		0.19	65		1	1		0.21	61
76	tetrahydropyran	0.22		4.00	1720		0.25	15		1	1		0.42	92
77	thiophene	0.39		10.9	2694		0.41	6		1	1		0.42	9
78	trichloroethylene	295		285	3		-	-		5	4		-	-
79	triethyl amine	1.15		6.45	461		1.05	8		2	2		1.29	12
80	vinyl acetate	0.70		-1.44	305		0.73	5		1	1		0.62	11
81	1,3-bis(4-nitrophenyl)urea	60		43.8	27	v	62.6	4	v	4	4	v	47.7	20
82	2,4-dichlophenoxy ethyl benzoate	60	v	73.7	23	v	45.7	24		4	4		61.0	2
83	2-acetylamino5-nitrothiazole	40		48.5	21		38.1	5		4	4		40.7	2
84	2-amino-5-nitrothiazole	30		48.9	63		26.7	11	v	4	4		32.8	9
85	4-chloro-2-nitro aniline	140		89.0	36		140	0		4	4		136	3
86	a,a'-azo isobutyronitrile	25		24.1	3		25.3	1	v	3	4		24.7	1
87	aceto acetanilide	20		30.2	51		20.4	2		3	4		19.2	4
88	adipic acid	60		28.9	52		50.5	16		4	4	v	77.6	29
89	anthranilic acid	35		43.0	23		35.4	1	v	4	4		33.9	3
90	ascorbic acid	60		62.2	4		25.1	58		4	4		56.5	6
91	aspirin	16		24.5	53		31.6	98		3	3		17.0	6
92	azelaic acid	25		19.0	24		11.7	53		3	3		24.3	3
93	benzoic acid	12		11.4	5		29.8	149		3	3		11.9	1
94	benzotriazole	30		33.1	10	v	39.3	31		4	4		38.1	27
95	benzoyl peroxide	21		7.98	62		60.4	188	v	3	4		19.5	7
96	bis(2-hydroxy-5-chlorophenyl)-methane	60		57.9	3		57.5	4		4	4		59.7	1
97	caprolactam	60		12.4	79		19.5	67		4	2		62.5	4
98	cyclohexanone peroxide	21		23.5	12	v	30.7	46		3	3	v	22.6	8

Table A.8 – Models Responses (continued)

Data				**Global Model**			**Local Models**			**Classification**			**Class Models**	
#	**Name**	**MIE [mJ]**		**Pred**	**ARD %**		**Pred**	**ARD %**		**Actual**	**Tree**		**Pred**	**ARD %**
99	dehydroacetic acid	15		30.4	103	v	-12.8	185		3	3		16.9	13
100	diazo amino benzene	20		17.0	15		31.7	59		3	4		23.8	19
101	dicyclopentadiene dioxide	30		14.0	53	v	15.8	47		4	4		30.9	3
102	dimethyl isophtalate	15		29.4	96		16.5	10		3	3		16.8	12
103	dimethyl terephtalate	20		19.9	1		45.7	129		3	3		15.3	23
104	dinitrobenzamide	45		58.1	29		46.9	4		4	4	v	51.6	15
105	dinitrobenzoic acid	45	v	42.2	6	v	43.9	2	v	4	4		47.4	5
106	dinitrotoluamide	15		44.5	197		19.5	30		3	4		16.8	12
107	diphenyl	20		-5.71	129		18.5	7		3	4		20.3	2
108	di-t-butyl p-cresol	15		16.4	9		48.1	220		3	4		16.3	8
109	DL methionine	35		48.9	40		49.9	43		4	4		34.1	2
110	ethylenediaminetetraacetic acid	50		31.2	38		34.3	31		4	3		37.6	25
111	fumaric acid	35	v	27.8	21		8.90	75	v	4	3		34.5	1
112	hexamethylenetetramine	10		19.1	91		26.2	162		2	2	v	1.7	83
113	isatoic anhydride	25		34.8	39		23.9	5		3	3		21.1	16
114	isophtalic acid	25	v	28.0	12		12.4	50		3	3		22.1	12
115	lauryl peroxide	12		24.2	102		78.5	554	v	3	3	v	11.5	4
116	l-sorbose	80		50.3	37		42.9	46		4	4		83.9	5
117	mannitol	40		45.0	13		52.8	32		4	4		42.0	5
118	methylamino anthraquinone	50		24.4	51	v	30.9	38		4	4		41.2	18
119	nitropyridone	35		21.8	38		20.3	42		4	4		31.6	10
120	o-chloroaceto acetanilide	30		38.9	30		27.8	7		4	4		47.8	59
121	p-chloroaceto acetanilide	20		44.0	120		14.3	28		3	4	v	20.4	2
122	pentaerythritol	10		31.6	216		15.3	53		2	3		10.0	0
123	phosphorus pentasulphide	15		15.7	5		52.6	251		3	4	v	30.0	100
124	phtalimide	50	v	28.1	44		16.8	66	v	4	4		42.5	15
125	phthalic anhydride	15		16.0	7		9.70	35		3	3		18.2	21
126	phytosterol	10		12.4	24		30.4	204		2	4		10.0	0
127	p-phenylene diamine	30		28.5	5		16.6	45		4	4	v	-56.0	287
128	salicylanilide	20		29.8	49		20.9	4	v	3	4		20.5	2
129	sorbic acid	15		24.9	66		24.3	62		3	3		13.7	9
130	stearic acid	25		16.1	35		17.0	32		3	3		24.1	3
131	t-butyl benzoic acid	25	v	12.7	49		16.5	34		3	3	v	19.6	22

Table A.8 – Models Responses (continued)

Data				**Global Model**			**Local Models**		**Classification**			**Class Models**	
#	**Name**	**MIE [mJ]**		**Pred**	**ARD %**		**Pred**	**ARD %**	**Actual**	**Tree**		**Pred**	**ARD %**
132	terephtalic acid	20		33.4	67	v	71.4	257	3	3		21.5	7
133	trinitrotoluene	75	v	51.6	31	v	56.8	24	4	4	v	36.3	52

Differential Scanning Calorimetry

B.1 Tolerance Zone Construction

In chapter 6, the experimental error of DSC measurements is estimated through a repeatability study. As the DSC curves are fitted by Fraser-Suzuki models, as shown in equation 1, the inference of the error on each individual parameter included in the Fraser-Suzuki model has been assessed in order to estimate the overall impact on the DSC curve.

$$\Phi = \Phi_{max} \cdot \exp\left[\frac{-\ln(2)}{a^2} \cdot \ln^2\left(1 + 2a\frac{(T - T_{max})}{FW}\right)\right] \tag{1}$$

The results have been computed following the procedure detailed in chapter 2, section 2.5.3. The intermediate steps are fully expressed here. The tolerance zone, which is constructed around the DSC curves, and is based on the estimation of the overall deviation of the curve σ_{Φ} depending on the deviations for each of the parameter $\sigma_{\bar{x}_i}$:

$$\sigma_{\Phi} = \sqrt{\sum_i \left[\left(\frac{\partial \Phi}{\partial x_i}\right)_B \cdot \sigma_{\bar{x}_i}\right]^2} \tag{2}$$

where $\frac{\partial \Phi}{\partial x_i}$ are:

$$\frac{\partial \Phi}{\partial \Phi_{max}} = \exp\left[\frac{-\ln(2)}{a^2} \cdot \ln^2\left(1 + 2a\frac{(T - T_{max})}{FW}\right)\right] \tag{3}$$

$$\frac{\partial \Phi}{\partial T_{max}} = \Phi \cdot \frac{\ln(2)}{a^2} \cdot \frac{4a\ln\left(1 + 2a\frac{(T - T_{max})}{FW}\right)}{FW + 2a(T - T_{max})} \tag{4}$$

$$\frac{\partial \Phi}{\partial FW} = \Phi \cdot \frac{2\ln(2)}{a^2} \cdot \ln\left(1 + 2a\frac{(T - T_{max})}{FW}\right) \cdot \frac{(2a(T - T_{max}) + FW)}{FW^2} \tag{5}$$

$$\begin{aligned}\frac{\partial \Phi}{\partial a} = \Phi \cdot \Bigg[& \left(\frac{2\ln(2)}{a^3} \cdot \ln^2\left(1 + 2a\frac{(T - T_{max})}{FW}\right)\right) + \\ & \left(\frac{-4\ln(2)(T - T_{max})}{a^2(FW + 2a(T - T_{max}))} \cdot \ln\left(1 + 2a\frac{(T - T_{max})}{FW}\right)\right)\Bigg]\end{aligned} \tag{6}$$

B.2 Nitro Compounds Study

Table B.1 – Nitro Compounds Models Parameters

Property	i	Coefficient	Parameter	Name
ΔH_r	0	$1.95 \cdot 10^5$	y_o	Intercept
	1	$1.59 \cdot 10^4$	$AB_{MO,max}$	Max anti-bonding contribution of one MO
	2	$7.70 \cdot 10^3$	$BO_{C,avg}$	Average bond order for atom C
	3	$-1.55 \cdot 10^4$	$E_{R,max(HC)}$	Max resonance energy for bond H-C
	4	$-3.37 \cdot 10^3$	$FHACA$	Fractional H-acceptor ability of the molecule (HACA/TMSA) (MOPAC PC)
	5	$3.99 \cdot 10^3$	S^{γ}_{XZ}	Relative shadow area: ZX Shadow / ZX Rectangle
$\Phi_{\max}$	0	$1.20 \cdot 10^3$	y_o	Intercept
	1	11.2	^0SIC	Structural Information content (order 0)
	2	−2.27	$HDSA2$	HA dependent H-donors surface area (Zefirov PC)
	3	$-1.29 \cdot 10^3$	$BO_{\sigma-\sigma,max}$	Max $\sigma - \sigma$ bond order
	4	−15.9	$E_{C,tot/N}$	Total molecular electrostatic interaction / # of atoms
	5	16.3	OK	Image of the Onsager-Kirkwood solvation energy
	6	11.6	S_{tot}/N	Total entropy (300K) /# atoms
$T_{\max}$	0	$-9.89 \cdot 10^2$	y_o	Intercept
	1	$-4.67 \cdot 10^{-1}$	$PPSA2$	Total charge weighted PPSA (MOPAC PC)
	2	$6.80 \cdot 10^2$	$B_{MO,max}$	Max bonding contribution of one MO
	3	$6.88 \cdot 10^3$	$BO_{\sigma-\pi,max}$	Max $\sigma - \pi$ bond order
	4	$5.07 \cdot 10^3$	$NRI_{C,min}$	Min nucleophilic reaction index for atom C
	5	−79.3	$BO_{C,avg}$	Average bond order for atom C
FW	0	$6.57 \cdot 10^3$	y_o	Intercept
	1	21.3	$E_{ne(CC),max}$	Max nuclear-electron attraction for bond C-C
	2	$-2.05 \cdot 10^{-2}$	ν_H	Highest normal mode vibration frequency
	3	$-1.16 \cdot 10^2$	$E_{aC,max}$	Max atomic state energy for atom C
	4	1.66	$WPSA2$	Weighted PPSA (PPSA2*TMSA/1000) (Zefirov PC)
	5	$-5.97 \cdot 10^1$	OK	Image of the Onsager-Kirkwood solvation energy
a	0	−43.3	y_o	Intercept
	1	$7.84 \cdot 10^{-3}$	$HDSAM$	H-donors surface area (MOPAC PC)
	2	78.8	$R_{C,max}$	Max 1-electron reaction index for atom C
	3	1.54	$E_{R,max(NO)}$	Max resonance energy for bond N-O
	4	−72.4	$NRI_{C,min}$	Min nucleophilic reaction index for atom C
	5	$-1.82 \cdot 10^{-2}$	S_{vib}	Vibrational entropy (300K)

Table B.2 – Nitro Models Responses

Data			Partial area		(J/g)	Amplitude		(W/g)	Position		(°C)	Half Width		(°C)	Asymmetry		(-)
#	Name	Set	Obs	Pred	RD[%]	Obs	Pred	RD[%]	Obs	Pred	RD[%]	Obs	Pred	RD[%]	Obs	Pred	RD[%]
1	nitrobenzene	Tr	-2648	-2513	5.1	5.70	5.34	6.3	410	404	1.6	28.13	29.36	4.4	-0.32	-0.21	34.4
2	2-nitroaniline	Tr	-3000	-2976	0.8	2.52	3.40	35.1	353	352	0.2	71.20	68.96	3.1	-0.37	-0.42	16.0
3	1,2-dinitrobenzene	Tr	-4370	-4432	1.4	4.94	4.82	2.5	388	390	0.5	54.12	48.54	10.3	-0.26	-0.38	43.6
4	1,3-dinitrobenzene	Val	-3832	-4026	5.1	11.91	7.19	39.6	418	427	2.2	18.31	8.58	53.1	-0.53	-0.75	42.7
5	1,4-dinitrobenzene	Tr	-3989	-3950	1.0	7.75	7.87	1.6	383	387	1.2	32.26	32.43	0.5	0.06	0.09	44.9
6	2,4-dinitrotoluene	Tr	-1877	-1848	1.5	9.77	9.84	0.6	315	317	0.5	12.03	7.97	33.8	-0.06	-0.15	145.1
7	2,6-dinitrotoluene	Tr	-3168	-3121	1.5	19.57	19.43	0.7	343	342	0.4	7.00	9.11	30.3	-0.99	-0.87	11.7
8	2-nitrobenzoic acid	Tr	-3148	-3074	2.3	2.79	2.60	6.8	319	322	0.9	70.43	73.67	4.6	0.12	0.19	67.7
9	2-nitrophenol	Tr	-2436	-2441	0.2	3.52	3.73	5.9	312	310	0.8	43.40	40.81	6.0	0.05	0.03	34.5
10	2-nitrotoluene	Tr	-2759	-2827	2.4	4.67	4.42	5.4	368	373	1.4	36.73	36.33	1.1	-0.15	-0.26	77.5
11	3,4-dinitrotoluene	Tr	-3472	-3538	1.9	11.82	12.12	2.6	328	324	1.2	18.07	18.34	1.5	0.23	0.26	13.6
12	3-nitroaniline	Val	-2432	-2029	16.6	2.75	-1.75	163.6	343	363	5.7	51.53	68.93	33.8	0.45	-0.01	102.7
13	3-nitrobenzoic acid	Tr	-1776	-1881	5.9	4.95	4.42	10.7	382	383	0.3	21.35	25.81	20.9	-0.39	-0.38	1.1
14	3-nitrophenol	Tr	-3456	-3449	0.2	2.95	2.51	14.9	361	360	0.4	72.69	74.29	2.2	-0.18	-0.10	41.7
15	3-nitrotoluene	Val	-2397	-2449	2.2	4.38	7.20	64.4	376	370	1.6	34.14	30.81	9.8	0.13	-0.98	876.5
16	4-nitroaniline	Tr	-2032	-2055	1.1	3.10	3.14	1.4	344	344	0.0	39.98	41.00	2.5	0.28	0.23	15.8
17	4-nitrobenzoic acid	Tr	-2143	-2157	0.7	2.64	2.64	0.0	374	372	0.6	50.12	45.71	8.8	-0.21	-0.27	29.5
18	4-nitrophenol	Tr	-2795	-2787	0.3	2.69	3.12	15.9	330	332	0.8	63.54	65.34	2.8	0.27	0.25	6.1
19	4-nitrotoluene	Tr	-2296	-2309	0.6	6.91	6.84	1.1	364	362	0.5	20.66	24.01	16.2	-0.15	-0.07	51.6

Table B.3 – Miscellaneous Set Models Parameters

Property	i	Coefficient	Parameter	Name
ΔH_r	0	$3.24 \cdot 10^3$	y_o	Intercept
	1	$-1.09 \cdot 10^3$	$E_{C,tot}/N$	Total molecular electrostatic interaction / # of atoms
	2	$2.87 \cdot 10^5$	$PCSA_C$	Partial Charged Surface Area for atom C
	3	$7.14 \cdot 10^3$	$FHASA$	Fractional H-acceptors surface area HASA-1/TMSA (Zefirov PC)
	4	4.19	$NCSA$	Negatively Charged Surface Area (MOPAC PC)
	5	$-4.68 \cdot 10^3$	OK	Image of the Onsager-Kirkwood solvation energy
Φ_{max}	0	2.60	y_o	Intercept
	1	$-3.20 \cdot 10^2$	N_P/N	Relative number of P atoms
	2	$9.99 \cdot 10^{-2}$	$DPSA3$	Difference in CPSAs (PPSA3-PNSA3) (MOPAC PC)
	3	$-2.68 \cdot 10^2$	$E_{C,avg}$	Average electrophilic reaction index for atom C
	4	-6.18	T^E_{all}	Topographic electronic index (all bonds)
	5	$1.28 \cdot 10^{-1}$	N_{HD}	count of H-donors sites (Zefirov PC) (all)
T_{max}	0	$4.67 \cdot 10^2$	y_o	Intercept
	1	$6.26 \cdot 10^2$	$V_{M,XYZ}$	Molecular Volume / XYZ Box
	2	-3.64	$PNSA3$	Atomic charge weighted PNSA (MOPAC PC)
	3	$2.49 \cdot 10^2$	$AB_{MO,max}$	Max anti-bonding contribution of one Molecular Orbital
	4	$-1.38 \cdot 10^3$	$E_{C,max}$	Max electrophilic reaction index for atom C
FW	0	$-1.11 \cdot 10^3$	y_o	Intercept
	1	$1.28 \cdot 10^3$	N_{Br}/N	Relative number of Br atoms
	2	$7.01 \cdot 10^1$	$\rho_{HC,min}$	Min coulombic interaction for bond H-C
	3	456	$FPSA_{HD}$	H-donors Fractional partial positively charged surface area (version 2)
	4	574	$BO_{\sigma\pi,max}$	Max SIGMA-PI bond order
	5	$2.19 \cdot 10^1$	$E_{nn(HC),min}$	Min nuclear repulsion for bond H-C
a	0	7.82	y_o	Intercept
	1	$1.01 \cdot 10^1$	$N_{C,max}$	Max nucleophilic reaction index for atom C
	2	$-6.92 \cdot 10^{-1}$	V_C, max	Max valency for atom C
	3	$-3.31 \cdot 10^{-2}$	$\Delta H_f/N$	Final heat of formation / # atoms
	4	-1.39	$\rho_{HC,max}$	Max coulombic interaction for bond H-C
	5	$1.11 \cdot 10^1$	$R_{C,max}$	Max 1-electron reaction index for atom C

Table B.4 – Miscellaneous Set Models Responses

Data			Partial area		(J/g)	Amplitude		(W/g)	Position		(°C)	Half Width		(°C)	Asymmetry		(-)
#	Name	Set	Obs	Pred	RD[%]	Obs	Pred	RD[%]	Obs	Pred	RD[%]	Obs	Pred	RD[%]	Obs	Pred	RD[%]
1	2,2-di(tert-butylperoxy) butane	Tr	-963	-1200	24.6	2.64	2.12	19.8	155	178	15.2	22.2	23.3	4.9	-0.28	-0.26	8.9
2	acetone	Tr	-26	-46	76.3	0.02	0.06	182.4	160	168	5.3	63.8	62.1	2.7	-0.03	-0.03	3.0
3	bis(2,2,2-trichloroethyl) azodicarboxylate	Tr	-431	-496	15.1	0.81	0.11	86.9	226	207	8.6	32.9	33.9	2.8	0.22	0.09	61.0
4	bromobenzene	Tr	-516	-651	26.0	0.24	0.24	2.6	186	193	3.5	139.2	139.2	0.0	-0.15	-0.06	61.6
5	butyl nitrile	Tr	-2427	-2197	9.5	5.62	4.68	16.7	209	210	0.6	26.7	21.3	20.1	-0.19	-0.33	75.8
6	di-(4-chlorobenzyl) azodicarboxylate	Val	-531	-1345	153.5	1.25	2.19	74.7	217	236	8.7	22.8	31.7	39.1	-0.61	-0.20	67.3
7	di-2-methoxyethyl azodicarboxylate	Tr	-769	-1122	45.8	1.35	1.72	26.6	190	189	0.6	35.4	32.8	7.5	-0.09	-0.06	33.7
8	dicumyl peroxide	Tr	-874	-609	30.3	2.00	3.22	61.2	168	192	14.2	26.2	26.9	2.6	-0.35	-0.33	5.8
9	diisopropyl azodicarboxylate	Tr	-879	-648	26.2	2.12	1.88	11.4	238	226	4.8	24.3	22.1	9.1	-0.42	-0.36	15.8
10	di-t-butyl peroxide	Tr	-1171	-941	19.6	2.40	2.69	11.9	182	191	5.2	28.8	26.6	7.5	-0.40	-0.29	29.4
11	ethyl diazoacetate	Tr	-354	-391	10.5	3.21	3.91	21.8	142	155	9.0	6.5	8.9	36.0	-0.37	-0.34	8.4
12	heptane	Tr	-42	-423	898.7	0.46	0.25	45.2	222	214	3.5	5.6	5.3	5.4	-0.16	-0.15	6.5
13	heptene	Tr	-52	-44	15.2	0.12	-0.25	306.2	176	157	10.7	24.2	12.4	48.8	0.41	0.32	21.7
14	isobutyl nitrate	Tr	-2251	-2449	8.8	6.65	6.55	1.5	207	216	4.2	20.7	21.1	2.0	-0.26	-0.21	22.1
15	isopentyl nitrite	Tr	-1987	-1829	7.9	5.00	5.09	1.8	207	207	0.2	24.3	30.9	27.0	-0.22	-0.24	8.8
16	methyl phenyl sulfoxide	Tr	-474	-233	50.8	8.92	8.79	1.5	278	276	0.8	2.3	1.6	27.8	-1.04	-1.09	5.1
17	NN-dimethyl formamide	Tr	-24	39	265.8	0.05	-0.34	727.5	182	198	8.6	23.7	20.3	14.5	-0.63	-0.59	5.8
18	t-butyl peroxyacetate	Val	-969	-1480	52.7	2.25	1.63	27.6	152	171	12.3	25.0	60.9	143.3	-0.46	-0.23	49.9
19	toluene	Tr	-21	-27	25.0	0.25	0.92	275.3	232	211	9.4	5.4	12.3	127.0	-0.07	0.05	172.5
20	triethyl phosphate	Tr	315	319	1.5	-7.48	-7.48	0.0	304	311	2.3	2.6	11.3	330.3	-0.05	-0.05	6.1
21	2-nitroaniline	Tr	-3000	-3281	9.4	2.52	3.59	42.5	353	305	13.5	71.2	75.2	5.6	-0.37	-0.45	24.3
22	1,4-dinitrobenzene	Tr	-3989	-3822	4.2	7.75	7.05	9.0	383	427	11.4	32.3	35.8	11.0	0.06	-0.11	279.0
23	2-nitrobenzoic acid	Tr	-3148	-3213	2.1	2.79	2.77	0.9	319	311	2.7	70.4	73.6	4.4	0.12	0.10	10.9
24	3,4-dinitrotoluene	Val	-3472	-4780	37.7	11.82	5.55	53.1	328	307	6.4	18.1	14.0	22.5	0.23	-0.12	150.6
25	3-nitrophenol	Tr	-3456	-3277	5.2	2.95	2.83	3.9	361	339	6.1	72.7	64.8	10.8	-0.18	-0.09	51.2

Confidential Data

The following tables, figures and results have been developed based on a proprietary database, and therefore some elements are subject to confidentiality request from the industrial partner of this project.

C.1 Chemical Families Study

Table C.1 – Chemical Families

Family	Defining Group	N	Partial Area ΔH_r [J g^{-1}]	Amplitude Φ_{max} [Wg^{-1}	Max Position T_{max} [°C]	Full Width FW [°C]	Asymmetry a [-]
Nitro Compounds	NO_2	45	-1525	5.58	295	23.7	-0.31
Nitroso Compounds, Nitrites	NO	13	-1184	4.54	271	28.2	-0.24
Azo Compounds and Tetrazoles	NN	14	-993	2.29	231	32.0	0.00
Phenylanilines	$PhNH_2$	28	-849	2.60	330	26.8	-0.24
Ethers	ROR	79	-613	1.91	282	29.5	-0.18
Amines	NH_2	145	-584	1.79	268	27.5	-0.04
Alcohols	OH	66	-582	1.65	252	32.4	0.00
Nitriles	CN	27	-564	2.06	295	26.1	-0.32
Chlorine Compounds	Cl	88	-556	2.06	297	28.5	-0.04
Sulfur Compounds	S	28	-506	2.84	262	19.5	-0.37
Bromide Compounds	Br	44	-471	1.82	298	25.0	-0.06
Fluorine Compounds	Fl	32	-363	1.27	251	22.3	-0.13
Esters	RO(O)R	50	-314	0.72	271	33.0	-0.10
Acids	COOH	30	-212	0.62	213	19.5	-0.19

Table C.2 – Chemical Families Models Evaluation Summary

Evaluation	Partial Area $\Delta \mathrm{H_r}$	Amplitude Φ_{max}	Max Position $\mathrm{T_{max}}$	Full Width FW	Asymmetry a
Nitroso and Nitrites					
R^2_{Tr}	0.958	0.986	0.953	0.944	0.966
R^2_{Val}	0.929	0.946	0.685	0.916	0.990
ARD_{Tr} [%]	7	16	7	14	45
ARD_{Val} [%]	653	145	7	23	28
Parameters	3	3	4	4	3
Dataset Size	13	Training	10	Validation	3
Azo and Tetrazoles					
R^2_{Tr}	0.936	0.886	0.894	0.924	0.638
R^2_{Val}	0.942	0.705	0.949	0.845	0.979
ARD_{Tr} [%]	11	40	10	15	1171
$ARD_{Tr,c}$ [%]					92
ARD_{Val} [%]	18	31	7	15	80
Parameters	3	3	3	4	4
Dataset Size	14	Training	10	Validation	4
Phenylamines					
R^2_{Tr}	0.948	0.781	0.788	0.770	0.738
R^2_{Val}	0.955	0.774	0.035	0.557	0.613
ARD_{Tr} [%]	37	72	7	41	221
ARD_{Val} [%]	16	1013	35	117	76
$ARD_{Val,c}$ [%]		10			
Parameters	5	5	5	7	9
Dataset Size	28	Training	23	Validation	5
Nitriles					
R^2_{Tr}	0.900	0.909	0.896	0.429	0.798
R^2_{Val}	0.823	0.260	0.010	0.150	0.146
ARD_{Tr} [%]	211	190	7	64	388
ARD_{Val} [%]	48	1030	39	103	1912
$ARD_{Val,c}$ [%]		100	26		255
Parameters	5	5	6	2	4
Dataset Size	27	Training	21	Validation	6
Ethers					
R^2_{Tr}	0.744	0.641	0.520	0.356	0.365

Table C.2 – Chemical Families Models Evaluation Summary (continued)

Evaluation	Partial Area $\Delta \mathrm{H_r}$	Amplitude Φ_{max}	Max Position $\mathrm{T_{max}}$	Full Width FW	Asymmetry a
R^2_{Val}	0.212	0.510	0.042	0.292	0.578
ARD_{Tr} [%]	248	263	21	82	246
ARD_{Val} [%]	247	343	28	165	554
$ARD_{Val,c}$[%]	176	139		106	128
Parameters	8	4	7	4	7
Dataset Size	78	Training	72	Validation	6

Tables C.3 to C.7 are subject to the confidentiality clause and are for this reason withheld.

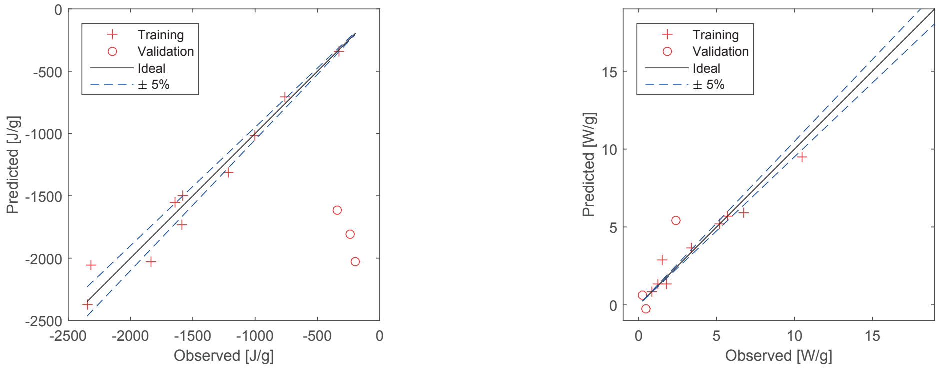
Table C.3 – Nitroso and Nitrites Models Parameters

Table C.4 – Azo and Tetrazoles Models Parameters

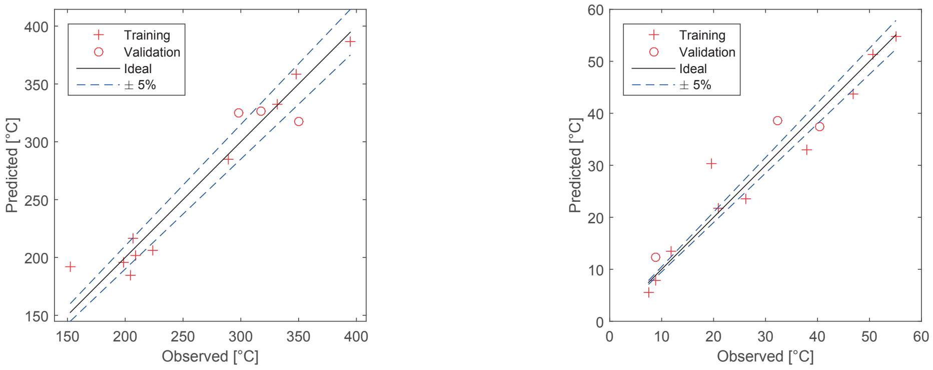
Table C.5 – Phenylamines Models Parameters

Table C.6 – Nitrile Models Parameters

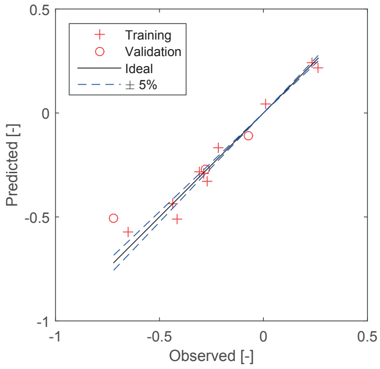
Table C.7 – Ethers Models Parameters



(a) Partial Area ΔH_r **(b)** Amplitude Φ_{max}

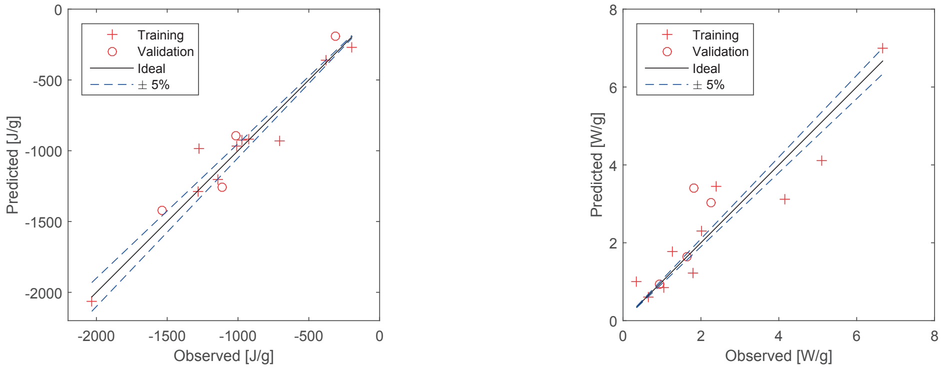


(c) Max Position T_{max} **(d)** Full Width FW



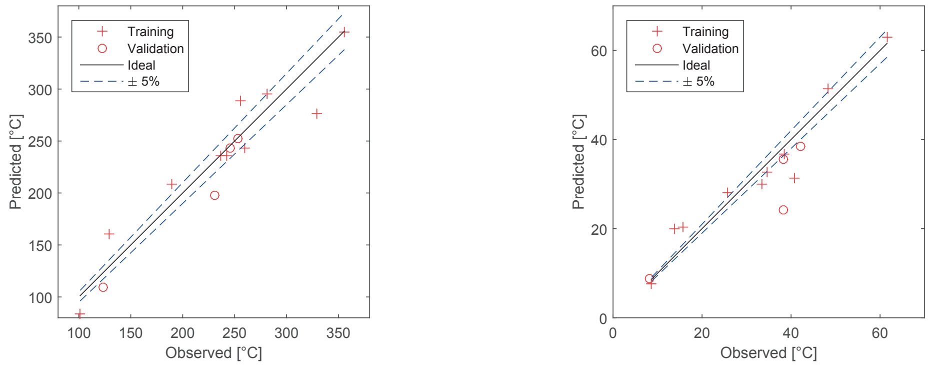
(e) Asymmetry a

Figure C.1 – Graphical Representations of the Models for the Nitroso and Nitrites Set



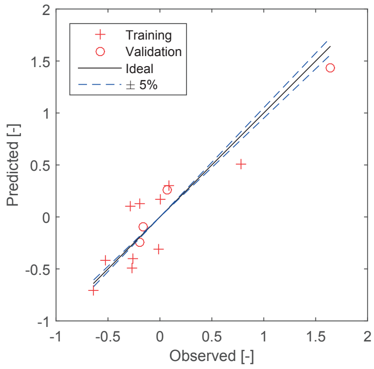
(a) Partial Area ΔH_r

(b) Amplitude Φ_{max}



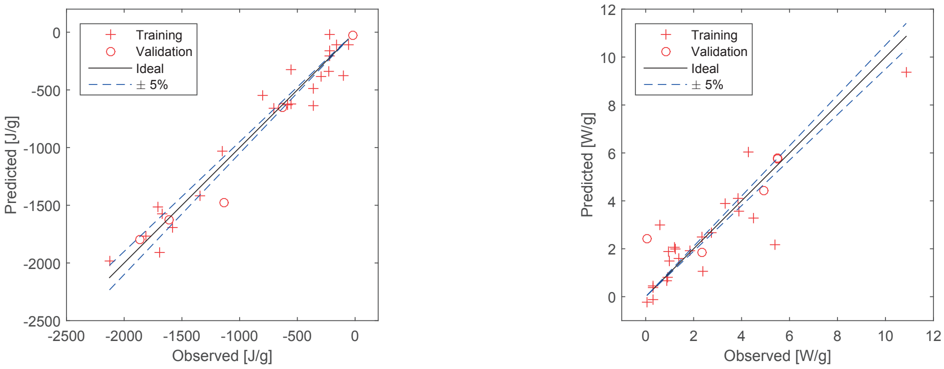
(c) Max Position T_{max}

(d) Full Width FW

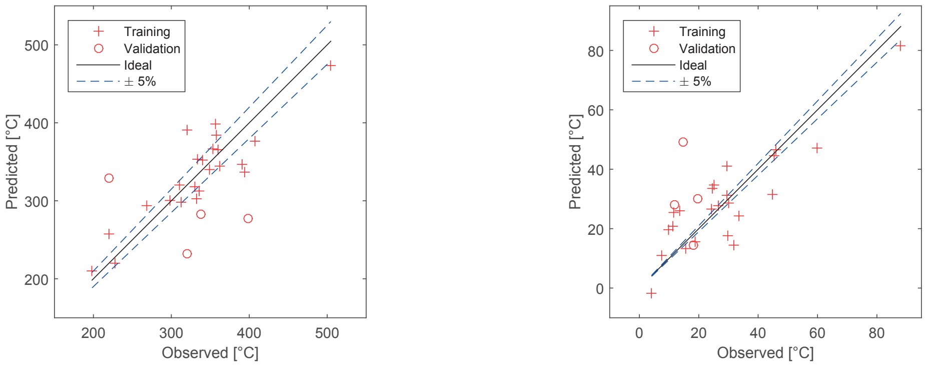


(e) Asymmetry a

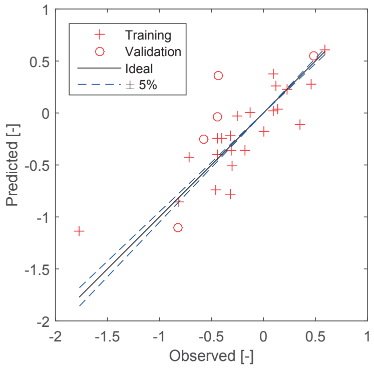
Figure C.2 – Graphical Representations of the Models for Azo and Tetrazoles



(a) Partial Area $\varDelta H_r$ **(b)** Amplitude $\varPhi_{max}$

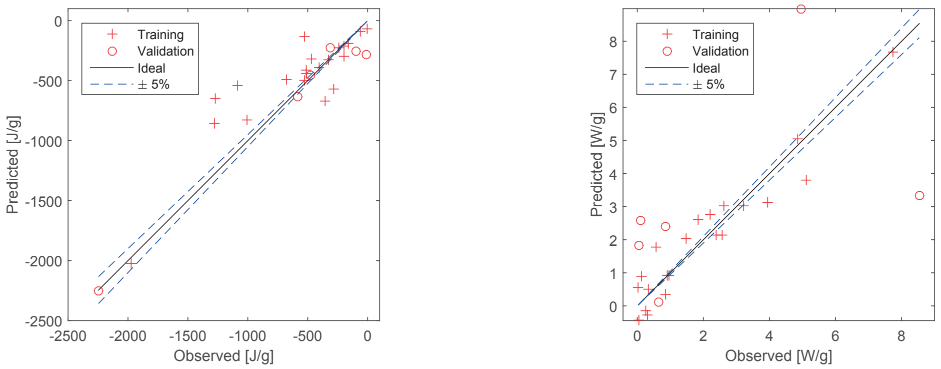


(c) Max Position T_{max} **(d)** Full Width FW

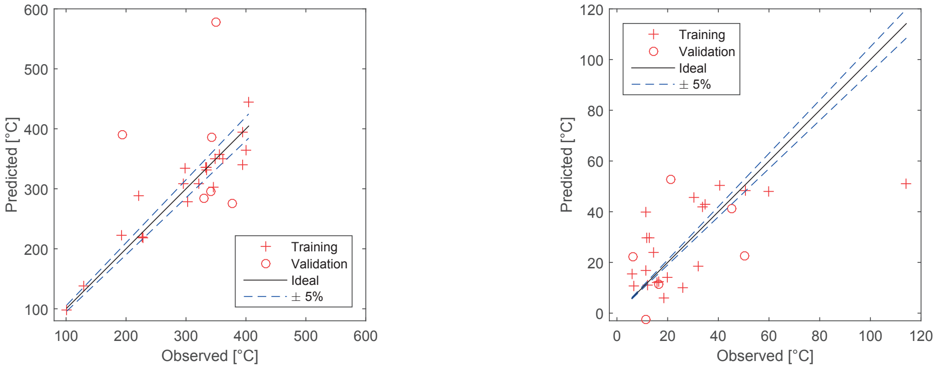


(e) Asymmetry a

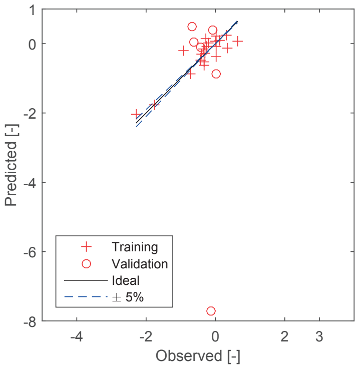
Figure C.3 – Graphical Representations of the Models for Phenylamines



(a) Partial Area ΔH_r **(b)** Amplitude Φ_{max}

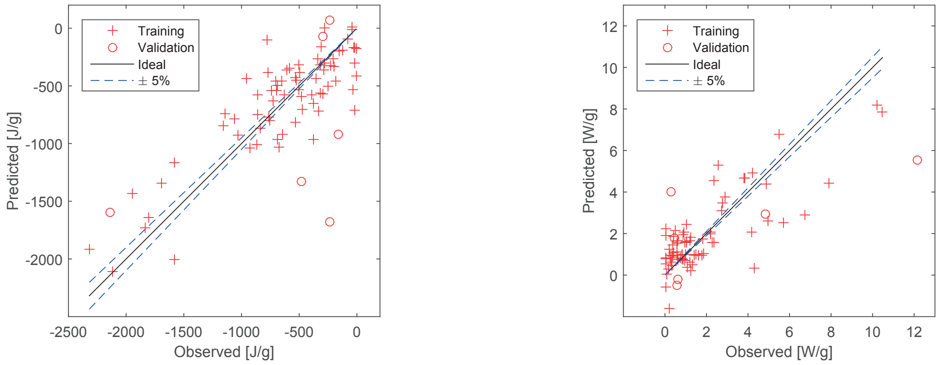


(c) Max Position T_{max} **(d)** Full Width FW

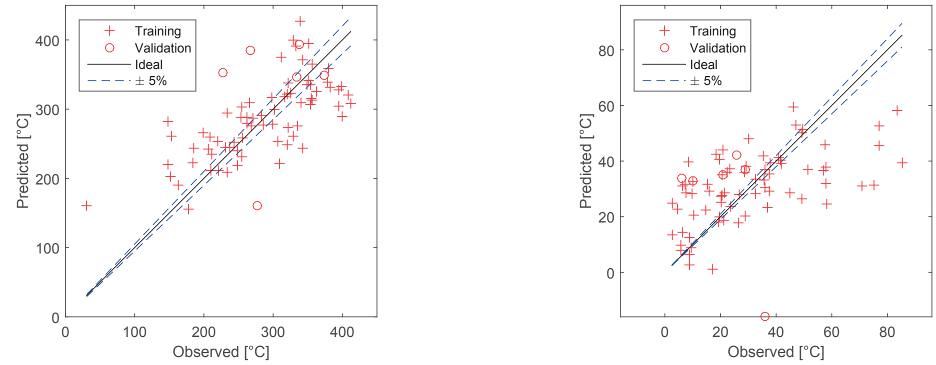


(e) Asymmetry a

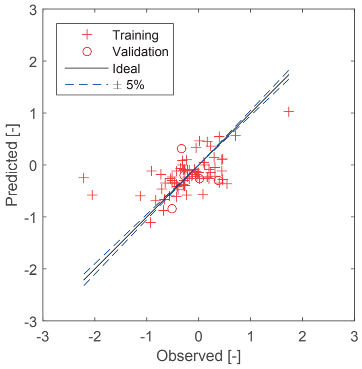
Figure C.4 – Graphical Representations of the Models for Nitriles



(a) Partial Area $\Delta\mathrm{H}_\mathrm{r}$ **(b)** Amplitude Φ_{max}



(c) Max Position $\mathrm{T}_{\mathrm{max}}$ **(d)** Full Width FW



(e) Asymmetry a

Figure C.5 – Graphical Representations of the Models for Ethers

Table C.8 is subject to the confidentiality clause and is for this reason withheld.

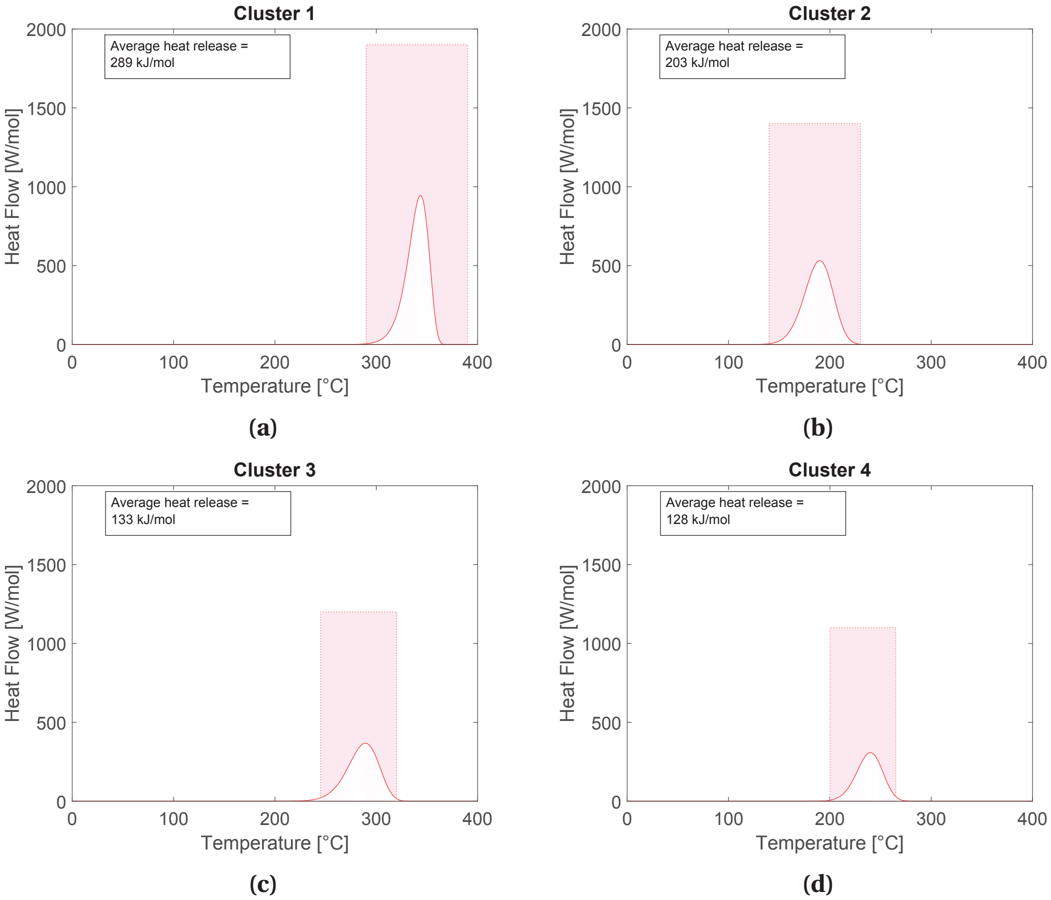
Table C.8 – Chemical Families Models Responses

C.2 Models on Clusters

Table C.9 – Clusters Models Evaluation Summary

Evaluation	Partial Area $\Delta \mathrm{H}_\mathrm{r}$	Amplitude Φ_max	Max Position T_max	Full Width FW	Asymmetry a
Cluster 1					
R^2_{Tr}	0.710	0.663	0.638	0.637	0.656
R^2_{CV}	0.935	0.441	0.478	0.908	0.194
ARD_{Tr} [%]	649	526	24	62	185
ARD_{CV} [%]	15	50	15	80	117
Parameters	10	12	19	29	35
Dataset Size	131	Training	126	Validation	5
Cluster 2					
R^2_{Tr}	0.700	0.606	0.646	0.638	0.635
R^2_{CV}	0.737	0.189	0.077	0.029	0.841
ARD_{Tr} [%]	571	988	18	67	209
ARD_{CV} [%]	562	192	53	125	220
Parameters	13	13	31	32	41
Dataset Size	131	Training	126	Validation	5
Cluster 3					
R^2_{Tr}	0.932	0.976	0.963	0.685	0.913
R^2_{CV}	0.035	0.124	0.451	0.474	0.681
ARD_{Tr} [%]	71	40	7	25	26
ARD_{CV} [%]	170	119	18	43	58
Parameters	3	3	3	2	3
Dataset Size	15	Training	10	Validation	5
Cluster 4					
R^2_{Tr}	0.621	0.619	0.625	0.789	0.773
R^2_{CV}	0.843	0.198	0.181	0.223	0.932
ARD_{Tr} [%]	276	131	25	41	224
ARD_{CV} [%]	30	53	35	48	143
Parameters	5	5	5	9	12
Dataset Size	55	Training	50	Validation	5

C.3 DSC Properties Clustering



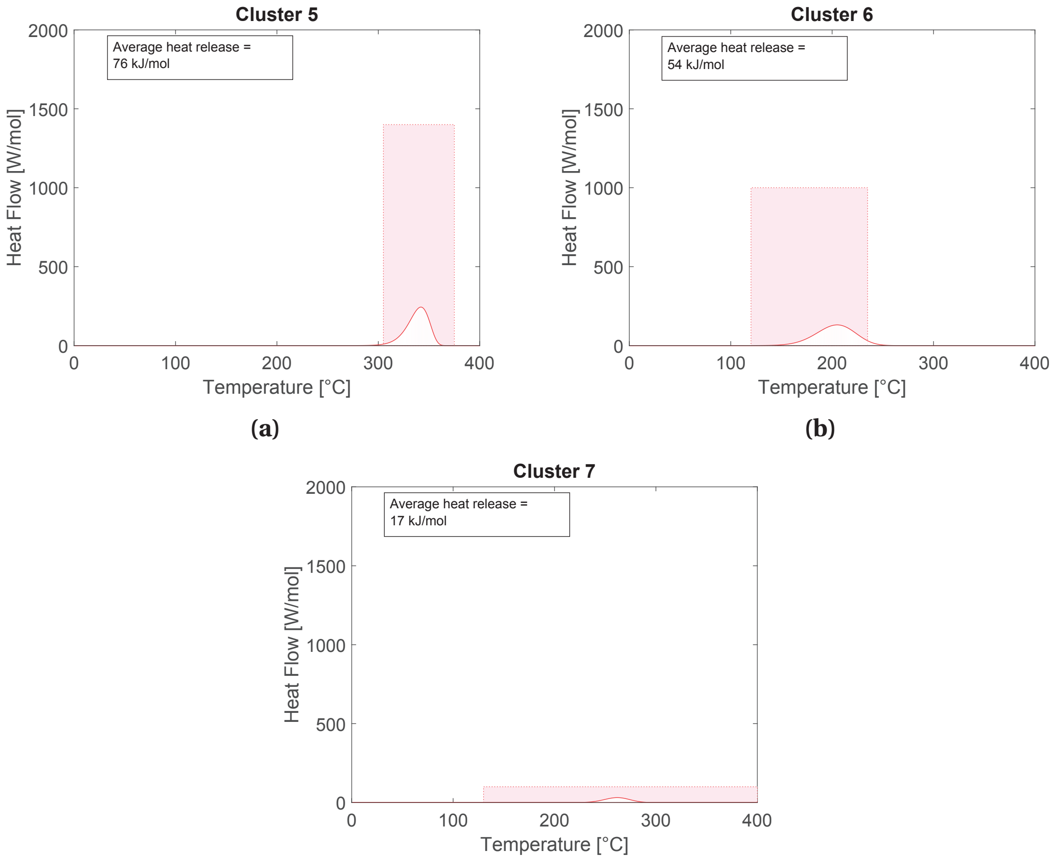
(a)

(b)

(c)

(d)

Figure C.6 – DSC Clusters Representation



(a)

(b)

(c)

Figure C.6 – DSC Clusters Representation (continued)

Table C.10 – DSC Clusters Models Evaluation Summary

Evaluation	Partial Area $\Delta \mathrm{H_r}$	Amplitude Φ_{max}	Max Position $\mathrm{T_{max}}$	Full Width FW	Asymmetry a
DSC Cluster 1					
R^2_{Tr}	0.650	0.730	0.822	0.662	0.632
R^2_{CV}	0.084	0.179	0.487	0.105	0.306
ARD_{Tr} [%]	39	57	7.3	34	103
ARD_{CV} [%]	19	28	11	19	39
Parameters	17	12	7	15	17
Dataset Size	48	Training	43	Cross-Validation	5
DSC Cluster 2					
R^2_{Tr}	0.900	0.862	0.928	0.940	0.989
R^2_{CV}	0.441	0.132	0.923	0.989	0.976
ARD_{Tr} [%]	159	288	10.4	21	21
ARD_{CV} [%]	14	26	15	6.8	19
Parameters	8	5	4	5	7
Dataset Size	18	Training	13	Cross-Validation	5
DSC Cluster 3					
R^2_{Tr}	0.821	0.918	0.909	0.903	0.656
R^2_{CV}	0.926	0.477	0.992	0.962	0.370
ARD_{Tr} [%]	233	146	3.9	43	45
ARD_{CV} [%]	513	163	1.0	22	49
Parameters	3	7	10	7	8
Dataset Size	22	Training	17	Cross-Validation	5
DSC Cluster 4					
R^2_{Tr}	0.763	0.671	0.737	0.304	0.697
R^2_{CV}	0.867	0.929	0.960	0.104	0.528
ARD_{Tr} [%]	409	446	8.3	51	101
ARD_{CV} [%]	169	189	4.1	66	558
Parameters	14	14	17	24	22
Dataset Size	73	Training	68	Cross-Validation	5
DSC Cluster 5					
R^2_{Tr}	0.808	0.791	0.822	0.730	0.593
R^2_{CV}	0.798	0.499	0.907	0.462	0.075
ARD_{Tr} [%]	91	141	4.8	35	225
ARD_{CV} [%]	9.5	42	3.1	28	61

Table C.10 – DSC Clusters Models Evaluation Summary (continued)

Evaluation	Partial Area $\Delta\mathrm{H}_\mathrm{r}$	Amplitude Φ_max	Max Position T_max	Full Width FW	Asymmetry a
Parameters	9	6	10	13	7
Dataset Size	48	Training	43	Cross-Validation	5
DSC Cluster 6					
R^2_{Tr}	0.963	0.982	0.998	0.934	0.877
R^2_{CV}	0.996	0.993	1.000	0.904	0.800
ARD_{Tr} [%]	18	54	0.9	29	347
ARD_{CV} [%]	8.4	177	0.7	14	309
Parameters	13	9	24	9	7
Dataset Size	29	Training	24	Cross-Validation	5
DSC Cluster 7					
R^2_{Tr}	0.389	0.018	0.729	0.676	0.744
R^2_{CV}	0.883	0.471	0.594	0.645	0.377
ARD_{Tr} [%]	215	261	18	60	157
ARD_{CV} [%]	56	96	14	54	86
Parameters	22	15	56	46	44
Dataset Size	130	Training	125	Cross-Validation	5

Tables C.11 to C.17 are subject to the confidentiality clause and are for this reason withheld.

Table C.11 – DSC Cluster 1 Models Parameters

Table C.12 – DSC Cluster 2 Models Parameters

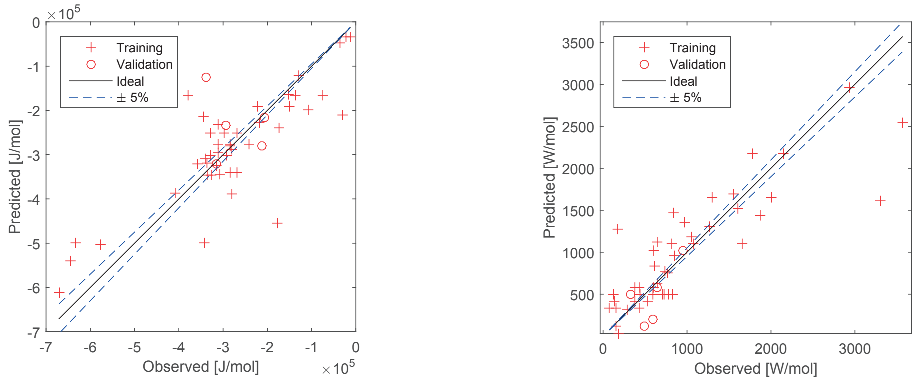
Table C.13 – DSC Cluster 3 Models Parameters

Table C.14 – DSC Cluster 4 Models Parameters

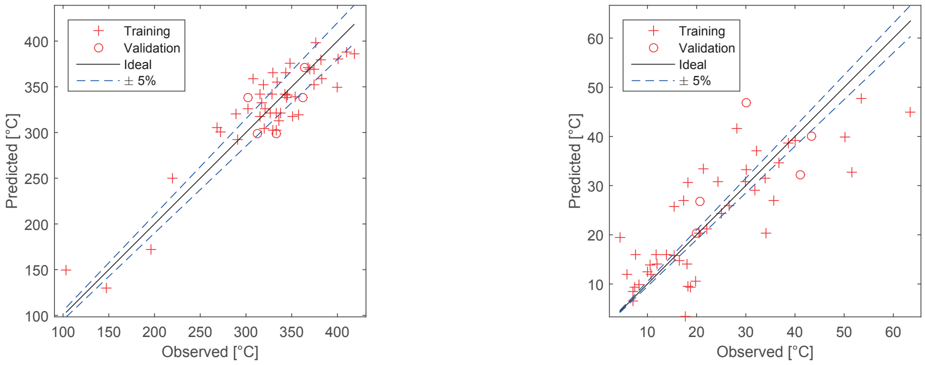
Table C.15 – DSC Cluster 5 Models Parameters

Table C.16 – DSC Cluster 6 Models Parameters

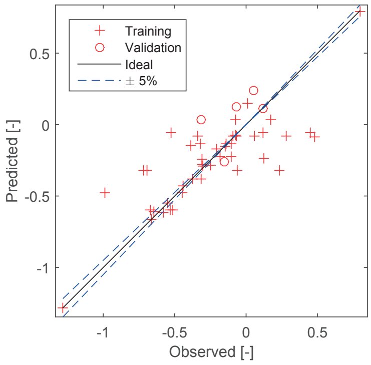
Table C.17 – DSC Cluster 7 Models Parameters



(a) Partial Area ΔH_r **(b)** Amplitude Φ_{max}

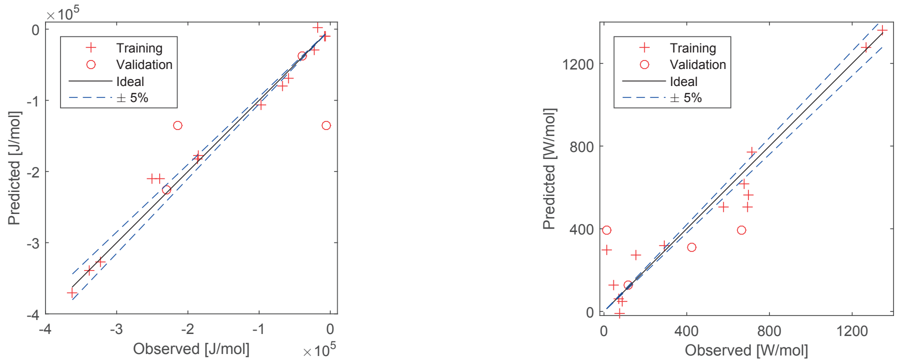


(c) Max Position T_{max} **(d)** Full Width FW

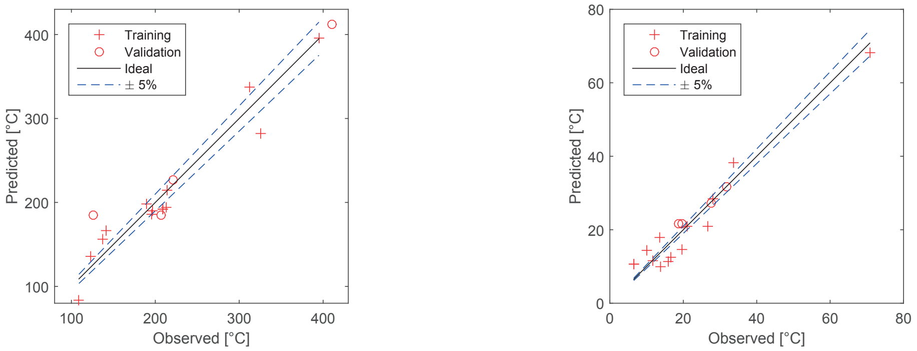


(e) Asymmetry a

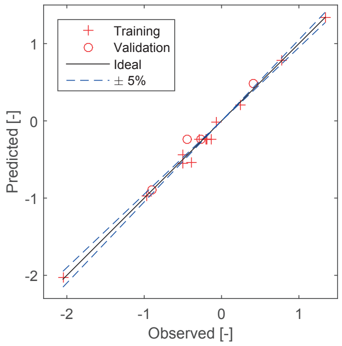
Figure C.7 – Graphical Representations of the Models for DSC Cluster 1



(a) Partial Area $\varDelta H_r$ **(b)** Amplitude $\varPhi_{max}$

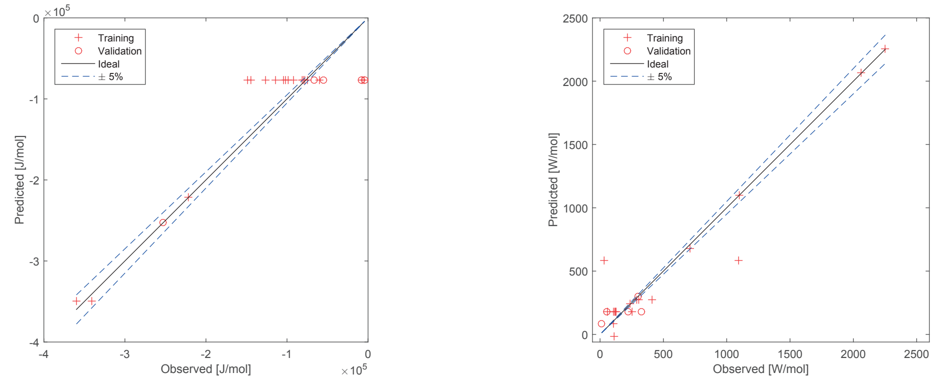


(c) Max Position T_{max} **(d)** Full Width FW



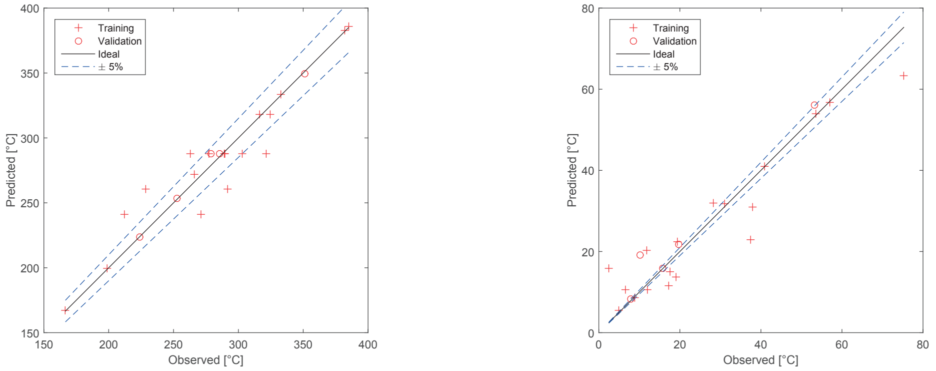
(e) Asymmetry a

Figure C.8 – Graphical Representations of the Models for DSC Cluster 2



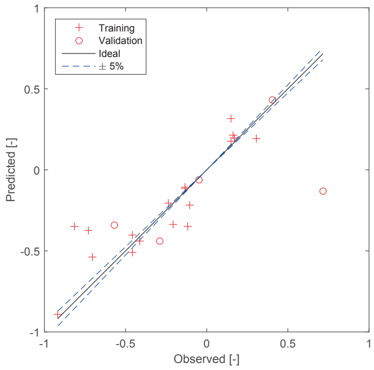
(a) Partial Area ΔH_r

(b) Amplitude Φ_{max}



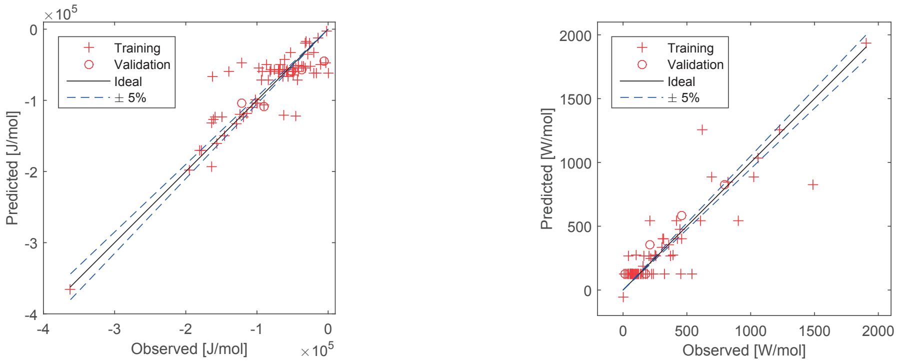
(c) Max Position T_{max}

(d) Full Width FW

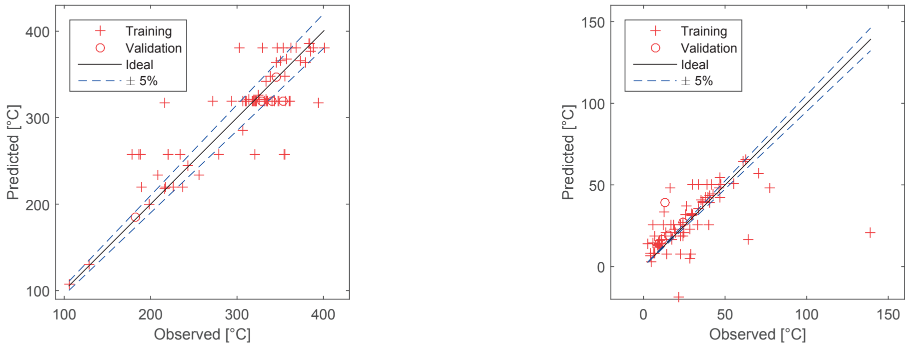


(e) Asymmetry a

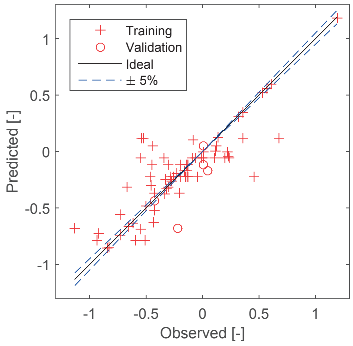
Figure C.9 – Graphical Representations of the Models for DSC Cluster 3



(a) Partial Area $\varDelta\mathrm{H}_{\mathrm{r}}$ **(b)** Amplitude $\varPhi_{\mathrm{max}}$

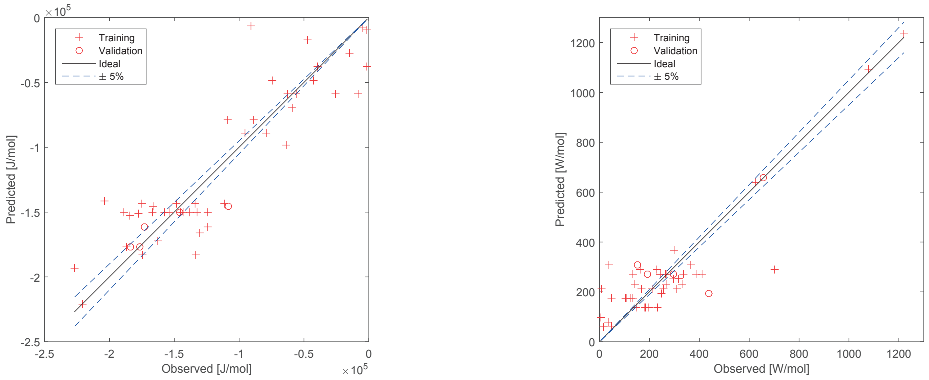


(c) Max Position $\mathrm{T}_{\mathrm{max}}$ **(d)** Full Width FW

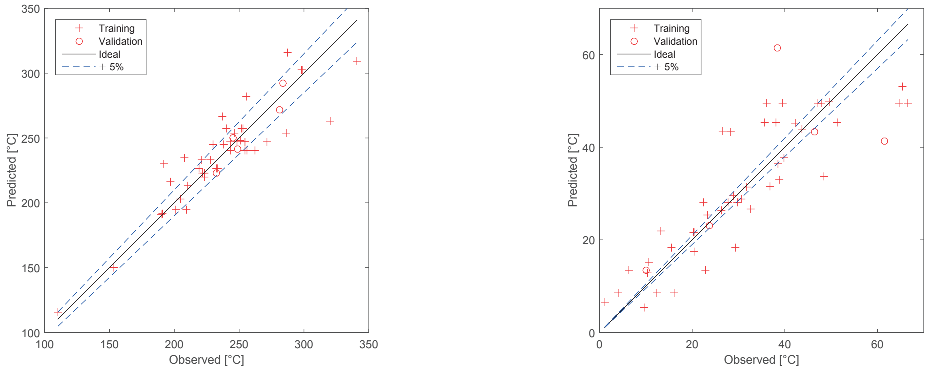


(e) Asymmetry a

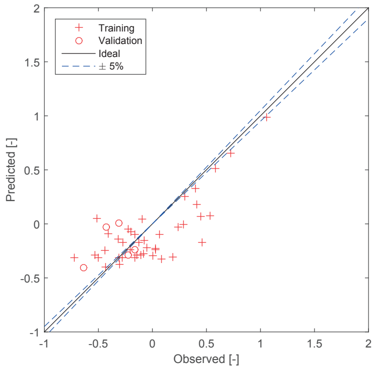
Figure C.10 – Graphical Representations of the Models for DSC Cluster 4



(a) Partial Area ΔH_r **(b)** Amplitude Φ_{max}

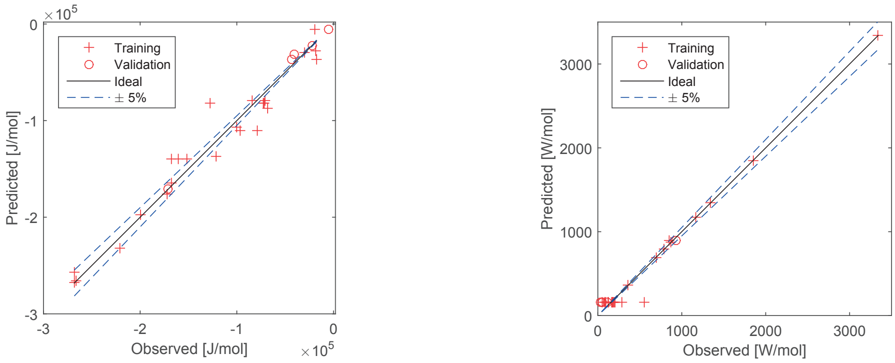


(c) Max Position T_{max} **(d)** Full Width FW

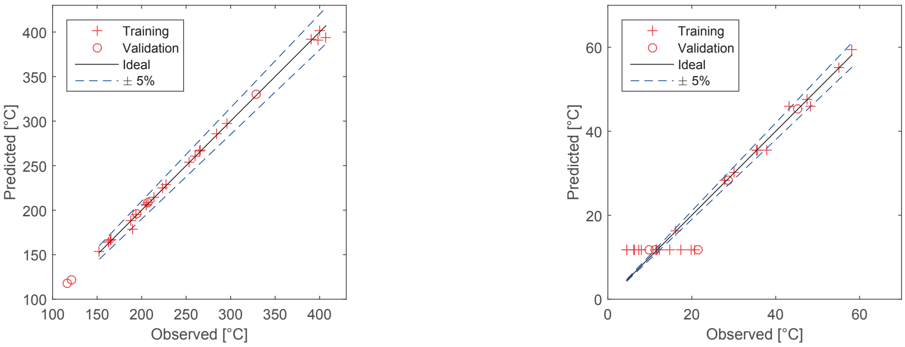


(e) Asymmetry a

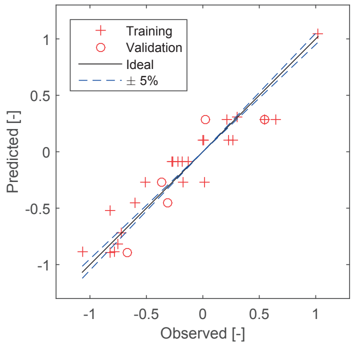
Figure C.11 – Graphical Representations of the Models for DSC Cluster 5



(a) Partial Area $\varDelta H_r$ **(b)** Amplitude $\varPhi_{max}$

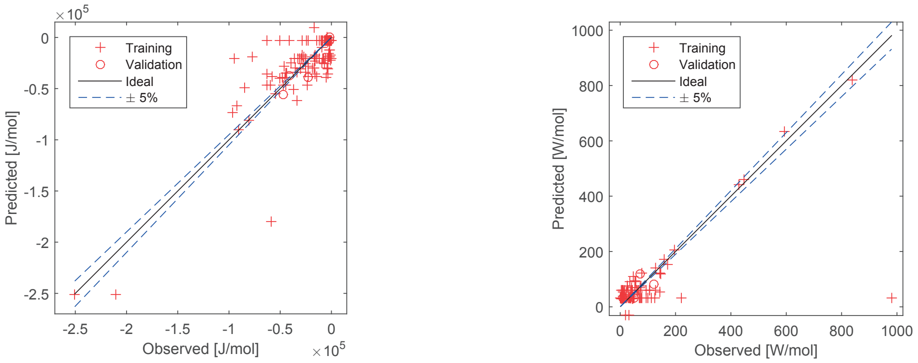


(c) Max Position T_{max} **(d)** Full Width FW

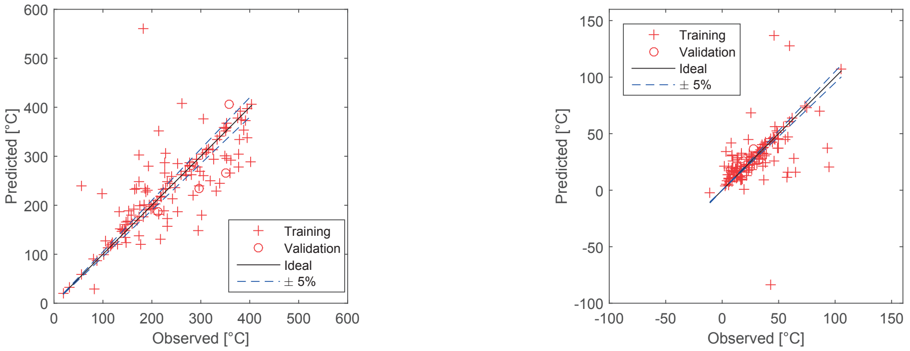


(e) Asymmetry a

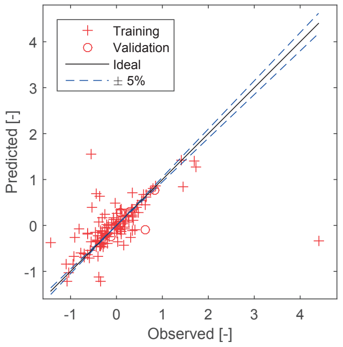
Figure C.12 – Graphical Representations of the Models for DSC Cluster 6



(a) Partial Area $\varDelta H_r$ **(b)** Amplitude $\varPhi_{max}$



(c) Max Position T_{max} **(d)** Full Width FW

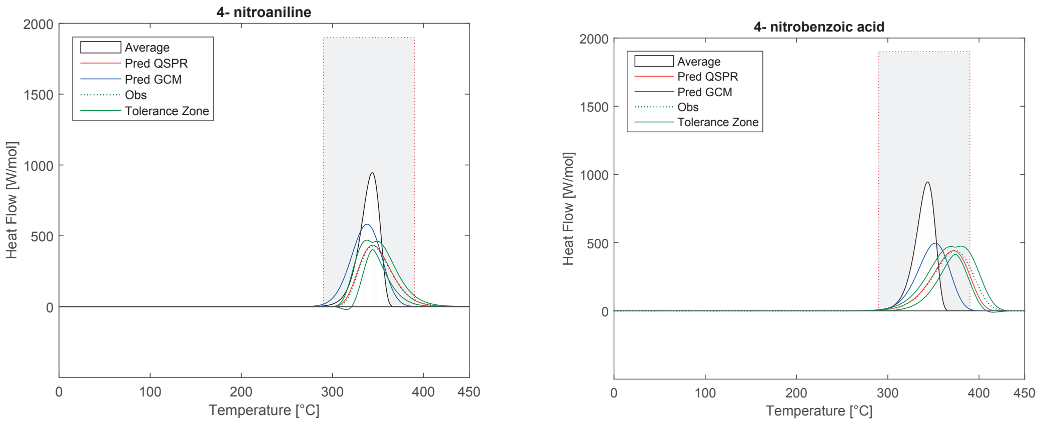


(e) Asymmetry a

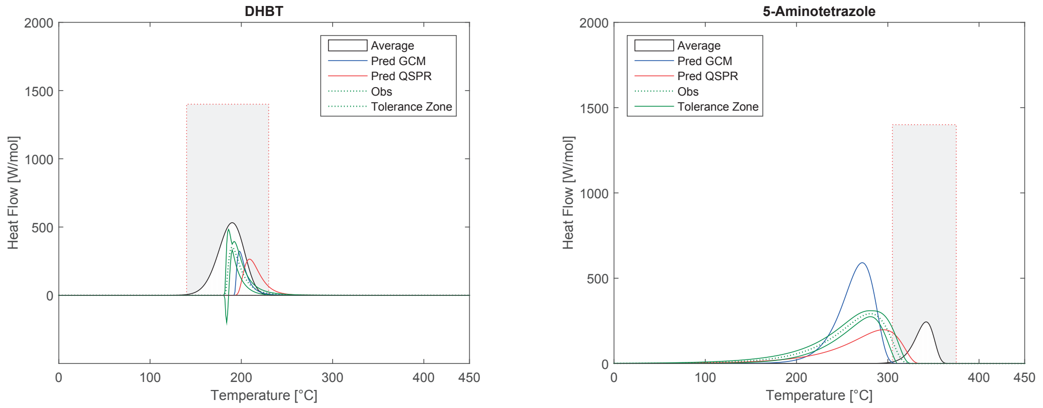
Figure C.13 – Graphical Representations of the Models for DSC Cluster 7

Table C.18 is subject to the confidentiality clause and is for this reason withheld.

Table C.18 – DSC Clusters Models Responses



(a) 4-nitroaniline **(b)** 4-nitrobenzenzoic acid



(c) DHBT **(d)** 5-aminotetrazole

Figure C.14 – Examples of DSC Reconstructions Compared to Actual Measurements

Bibliography

Ajmani, S., Rogers, S. C., Barley, M. H., and Livingstone, D. J. Application of QSPR to mixtures. *Journal of Chemical Information and Modeling*, 46:2043–2055, 2006. ISSN 15499596. doi: 10.1021/ci050559o.

Akaike, H. A New Look at the Statistical Model Identification. *IEE Transactions on Automatic Control*, 19:716–723, 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

AKTS SA. AKTS-Thermokinetics Software Version 3.91 DB v1.34. www.akts.ch, 2000.

Albahri, T. A. Flammability Characteristics of Pure Hydrocarbons. *Chemical Engineering Science*, 58(16):3629–3641, August 2003. ISSN 00092509. doi: 10.1016/S0009-2509(03)00251-3.

Albahri, T. A. MNLR and ANN Structural Group Contribution Methods for Predicting the Flash Point Temperature of Pure Compounds in the Transportation Fuels Range. *Process Safety and Environmental Protection*, 93(January):182–191, 2014. ISSN 09575820. doi: 10.1016/j.psep.2014.03.005.

Albert, A. *Selective Toxicity: The Physico-Chemical Basis of Therapy*, volume 27. Springer, 2013. ISBN 1489971300.

Ando, T., Fujimoto, Y., and Morisaki, S. Analysis of Differential Scanning Calorimetric Data for Reactive Chemicals. *Journal of Hazardous Materials*, 28:pp51–280, 1991. doi: 10.1016/0304-3894(91)87079-H.

ASTM International. Standard Test Method for Minimum Ignition Energy of a Dust Cloud in Air. Technical report, ASTM, 2007.

Baati, N. Feasibility Study on the Prediction of Thermal Stability of Chemicals. Master's thesis, École polytechnique fédérale de Lausanne, 2011.

Babrauskas, V. *Ignition Handbook, Principles and Applications to Fire Safety Engineering, Fire Investigation, Risk Management and Forensic Science.* Fire Science Publishers, 2003. ISBN 0972811133. doi: 10.1177/1042391504042549.

Bagheri, M., Rajabi, M., Mirbagheri, M., and Amin, M. BPSO-MLR and Anfis Based Modeling of Lower Flammability Limit. *Journal of Loss Prevention in the Process Industries*, 25(2): 373–382, March 2012. ISSN 09504230. doi: 10.1016/j.jlp.2011.10.005.

Bartknecht, W. *Dust Explosions: Course, Prevention, Protection.* Springer Verlag Berlin Heidelberg, 1989.

Beal, D. J. SAS Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria. In *SESUG 2005: The Proceedings of the SouthEast SAS Users Group, Portsmouth, VA, 2005*, page SA01_05, 2005.

Ben Talouba, I., Balland, L., Mouhab, N., and Abdelghani-Idrissi, M. A. Kinetic Parameter Estimation for Decomposition of Organic Peroxides by Means of DSC Measurements. *Journal of Loss Prevention in the Process Industries*, 24:391–396, 2011. ISSN 09504230. doi: 10.1016/j.jlp.2011.02.001.

Benson, S. W. and Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *The Journal of Chemical Physics*, 29(3):546, 1958. ISSN 00219606. doi: 10.1063/1.1744539.

Berman, J. J. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information.* Kaufmann, M., 2013. ISBN 0124047246.

Bernard, S., Lebecki, K., Gillard, P., Youinou, L., and Baudry, G. Statistical Method for the Determination of the Ignition Energy of Dust Cloud-experimental Validation. *Journal of Loss Prevention in the Process Industries*, 23:404–411, 2010. ISSN 09504230. doi: 10.1016/j.jlp.2010.01.006.

Berufsgenossenschaften. Richtlinien für die Vermeidung von Zündgefahren infolge elektrostatischer Aufladungen – Richtlinien „Statische Elektrizität“ –. Bundesverband der Unfallkassen, Hauptverband der gewerblichen Berufsgenossenschaften, January 1992.

Bishop, C. M. *Pattern Recognition and Machine Learning*, volume 4. Springer-Verlag New York, 2006. ISBN 9780387310732. doi: 10.1117/1.2819119.

Blake, E. S., Hammann, W. C., Edwards, J. W. F., Reichard, T. E., and Ort, M. R. Thermal Stability as a Function of Chemical Structure. *Journal of Chemical & Engineering Data*, 6(1):87–98, January 1961. ISSN 0021-9568. doi: 10.1021/je60009a020.

Boersma, S. L. A Theory of Differential Thermal Analysis and New Methods of Measurement and Interpretation. *Journal of the American Ceramic Society*, 38(8):281–284, 1955. ISSN 1551-2916. doi: 10.1111/j.1151-2916.1955.tb14945.x.

Borchardt, H. J. and Farrington, D. The Application of Differential Thermal Analysis to the Study of Reaction Kinetics. *Journal of the American Chemical Society*, 79(1):41–46, January 1957. ISSN 0002-7863. doi: 10.1021/ja01558a009.

Brown, J. S., Zilio, C., and Cavallini, A. Thermodynamic Properties of Eight Fluorinated Olefins. *International Journal of Refrigeration*, 33(2):235–241, 2010. doi: 10.1016/j.ijrefrig.2009.04.005.

Calcote, M. F., Gregory, C. A., Barnett, C. M., and Giemer, R. B. Spark Ignition. Effect of Molecular Structure. *Industrial & Engineering Chemistry*, 44(11):2656–2662, 1952.

Cances, E., Defranceschi, M., Kutzelnigg, W., Le Bris, C., and Maday, Y. Computational Quantum Chemistry : A Primer. *Handbook of Numerical Analysis*, 10:3–270, 2003. ISSN 1570-8659. doi: 10.1016/S1570-8659(03)10003-8.

Carson, P. and Mumford, C. Toxic Chemicals. In *Hazardous Chemicals Handbook*, chapter 5, pages 67–177. Elsevier Ltd., 2002.

CCPS, C. f. C. P. S. *Inherently Safer Chemical Processes - A Life Cycle Approach*. John Wiley and sons, Hoboken, NJ, 2nd editio edition, 2009. ISBN 978-0471-77892-9.

Cesana, C. and Siwek, R. *MIKE3 Manual*. Birsfelden, 2010.

Chickos, J. S., Hesse, D. G., Hosseini, S., Liebman, G., J. F.and David Mendenhall, Verevkin, S. P., Rakus, K., Beckhaus, H. D., and Rüchardt, C. F. Enthalpies of Vaporization of Some Highly Branched Hydrocarbons. *The Journal of Chemical Thermodynamics*, 27(6):693–705, 1995. doi: 10.1006/jcht.1995.0071.

Chouhan, T. R. The unfolding of Bhopal disaster. *Journal of Loss Prevention in the Process Industries*, 18(4-6):205–208, July 2005. ISSN 09504230. doi: 10.1016/j.jlp.2005.07.025.

Constantinou, L., Prickett, S. E., and Mavrovouniotis, M. L. Estimation of Properties of Acyclic Organic Compounds Using Conjugation Operators. *Industrial & Engineering Chemistry Research*, 33(2):395–402, 1994. ISSN 0888-5885. doi: 10.1021/ie00026a034.

Council of the European Union. Council Directive 1999/92/EC Of 16 December 1999 On Minimum Requirements For Improving The Safety And Health Protection Of Workers Potentially At Risk From Explosive Atmospheres, 1999. (15th individual Directive within the meaning of Article 16(1) of Directive 89/391/EEC).

Council of the European Union. Council Directive 2012/18/EU of 4 July 2012 on the Control of Major-accident Hazards Involving Dangerous Substances, 2012. amending and subsequently repealing Council Directive 96/82/EC.

Cross, J. and Farrer, D. *Dust Explosions*. Springer Science & Business Media, 1982. ISBN 1461568692.

DDBSST, D. ARTIST - Thermophysical Properties from Molecular Structure. www.ddbst.com/artist-property-estimation.html, 2009. Accessed: 2016-01-08.

Dehmer, M., Varmuza, K., and Bonchev, D. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, volume 2. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, February 2012. ISBN 9783527645121. doi: 10.1002/9783527645121.

Dewar, M. J. S. and Thiel, W. Ground States of Molecules. 38. the MNDO Method. Approximations and Parameters. *Journal of the American Chemical Society*, 99(15):4899–4907, 1977.

Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *Journal of the American Chemical Society*, 107(13): 3902–3909, 1985.

Duh, Y.-S., Lee, C., Hsu, C.-C., Hwang, D.-R., and Kao, C.-S. Chemical Incompatibility of Nitrocompounds. *Journal of Hazardous Materials*, 53:183–194, 1997. ISSN 03043894. doi: 10.1016/S0304-3894(96)01829-8.

Eckerman, I. *The Bhopal Saga: Causes and Consequences of the World's Largest Industrial Disaster*. Universities Press, 2005. ISBN 8173715157.

Eckhoff, R. K. *Dust Explosions in the Process Industries*, volume 1. Gulf Professional Publishing, 2003. ISBN 9780750676021.

Egolf, L. M. and Jurs, P. C. Estimation of Autoignition Temperatures of Hydrocarbons, Alcohols, and Esters from Molecular Structure. *Industrial & Engineering Chemistry Research*, 31: 1798–1807, 1992.

Emeis, S. The Discovery of Latent Heat 250 Years Ago. *Meteorologische Zeitschrift*, 13(4): 329–333, July 2004. ISSN 09412948. doi: 10.1127/0941-2948/2004/0013-0329.

Fayet, G., Del Rio, A., Rotureau, P., Joubert, L., and Adamo, C. QSPR Modeling of Thermal Stability of Nitroaromatic Compounds: DFT vs AM1 Calculated Descriptors. *Journal of Molecular Modelling*, 16:805–812, 2009.

Fayet, G., Del Rio, A., Rotureau, P., Joubert, L., and Adamo, C. Predicting Explosibility Properties of Chemicals from Quantitative Structure-Property Relationships. *Process Safety Progress*, 29(4):359–371, 2010. ISSN 10668527. doi: 10.1002/prs.10379.

Fayet, G., Del Rio, A., Rotureau, P., Joubert, L., and Adamo, C. Predicting the Thermal Stability of Nitroaromatic Compounds Using Chemoinformatic Tools. *Molecular Informatics*, 30 (6-7):623–634, 2011. ISSN 1868-1751. doi: 10.1002/minf.201000077.

Fayet, G., Rotureau, P., Prana, V., and Adamo, C. Global and Local Quantitative Structure-property Relationship Models to Predict the Impact Sensitivity of Nitro Compounds. *Process Safety Progress*, 31(3):291–303, 2012. ISSN 10668527. doi: 10.1002/prs.11499.

Felinger, A. *Data Analysis and Signal Processing in Chromatography*. Elsevier, 1998. ISBN 0080525563.

Flynn, J. H. The Isoconversional Method for Determination of Energy of Activation at Constant Heating Rates. *Journal of Thermal Analysis*, 27(1):95–102, May 1983. ISSN 0368-4466. doi: 10.1007/BF01907325.

Fredenslund, A., Jones, R. L., and Prausnitz, J. M. Group-contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE Journal*, 21(6):1086–1099, November 1975. ISSN 0001-1541. doi: 10.1002/aic.690210607.

Gani, R. ICAS Version 17.0. www.capec.kt.dtu.dk/Software/ICAS-and-its-Tools, 1999.

Gani, R. and Constantinou, L. Molecular Structure Based Estimation of Properties for Process Design. *Fluid Phase Equilibria*, 116:75–86, 1996.

Gebauer, S., Knütter, I., Hartrodt, B., Brandsch, M., Neubert, K., and Thondorf, I. Three-Dimensional Quantitative Structure - Activity Relationship Analyses of Peptide Substrates of the Mammalian H+/Peptide Cotransporter PEPT1. *Journal of Medicinal Chemistry*, 46 (26):5725–5734, 2003. ISSN 00222623. doi: 10.1021/jm030976x.

Gharagheizi, F. Prediction of Upper Flammability Limit Percent of Pure Compounds from Their Molecular Structures. *Journal of Hazardous Materials*, 167(1-3):507–10, August 2009. ISSN 1873-3336. doi: 10.1016/j.jhazmat.2009.01.002.

Glor, M. Ignition Hazard Due to Static Electricity in Particulate Processes. *Powder Technology*, 135-136:223–233, 2003. ISSN 00325910. doi: 10.1016/j.powtec.2003.08.017.

Gmelin, E. and Sarge, S. M. Calibration of Differential Scanning Calorimeters. *Pure and Applied Chemistry*, 67(11):1789–1800, 1995. ISSN 0033-4545. doi: 10.1351/pac199567111789.

Golmohammadi, H. and Dashtbozorgi, Z. Quantitative structure–property relationship studies of gas-to-wet butyl acetate partition coefficient of some organic compounds using genetic algorithm and artificial neural network. *Structural Chemistry*, 21(6):1241–1252, November 2010. ISSN 1040-0400. doi: 10.1007/s11224-010-9669-8.

Grüber, C. and Buß, V. Quantum-mechanically Calculated Properties for the Development of Quantitative Structure-activity Relationships (QSAR's). Pka-values of Phenols and Aromatic and Aliphatic Carboxylic Acids. *Chemosphere*, 19(10–11):1595 – 1609, 1989. ISSN 0045-6535. doi: http://dx.doi.org/10.1016/0045-6535(89)90503-1.

Grewer, T. The Influence of Chemical Structure on Exothermic Decomposition. *Thermochimica Acta*, 187:133–149, 1991. ISSN 00406031. doi: 10.1016/0040-6031(91)87188-3.

Grewer, T., Frurip, D. J., and Keith Harrison, B. Prediction of Thermal Hazards of Chemical Reactions. *Journal of Loss Prevention in the Process Industries*, 12(5):391–398, 1999. ISSN 09504230. doi: 10.1016/S0950-4230(99)00011-X.

Grossel, S. S. Safety Considerations in Conveying of Bulk Solids and Powders. *Journal of Loss Prevention in the Process Industries*, 1(April):62–74, 1988.

Guerard, J. B. Regression Analysis and Forecasting Models. In *Introduction to Financial Forecasting in Investment Analysis*, pages 19–45. Springer New York, 2013. ISBN 978-1-4614-5238-6. doi: 10.1007/978-1-4614-5239-3.

Guha, R. On the Interpretation and Interpretability of Quantitative Structure-activity Relationship Models. *Journal of Computer-Aided Molecular Design*, 22(12):857–871, 2008. ISSN 0920654X. doi: 10.1007/s10822-008-9240-5.

Haase, H. *Electrostatic Hazards : Their Evaluation and Control.* Weinheim ; New York : Verlag Chemie,, 1977.

Hada, S. and Harrison, B. K. Prediction of Energy Release Hazards Using a Simplified Adiabatic Temperature Rise. *Journal of Loss Prevention in the Process Industries*, 20:151–157, February 2007. ISSN 09504230. doi: 10.1016/j.jlp.2007.02.001.

Hall, L. H. and Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Modeling*, 35(6):1039–1045, 1995. ISSN 1549-9596. doi: 10.1021/ci00028a014.

Hansch, C. and Dunn, W. J. Linear Relationships between Lipophilic Character and Biological Activity of Drugs. *Journal of Pharmaceutical Sciences*, 61(1):1–19, 1972. ISSN 00223549.

Hansch, C., Lien, E. J., and Helmer, F. Structure-activity Correlations in the Metabolism of Drugs. *Archives of Biochemistry and Biophysics*, 128(2):319–330, 1968. ISSN 00039861.

Harder, A., Escher, B. I., and Schwarzenbach, R. P. Applicability and Limitation of QSARs for the Toxicity of Electrophilic Chemicals. *Environmental Science and Technology*, 37(4955-4961): 4955–4961, 2003. ISSN 0013936X. doi: 10.1021/es0341992.

Hertzberg, M., Cashdollar, K. L., Green, G. M., and Zlochowera, I. A. Explosive Dust Cloud Combustion. In *Twenty-Fourth Symposium (International) on Combustion / The Combustion Institute*, pages 1837–1843, 1992.

Höhne, G. W. H., Hemminger, W., and Flammersheim, H.-J. *Differential Scanning Calorimetry*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996. ISBN 978-3-662-03304-3. doi: 10.1007/978-3-662-03302-9.

Holmes, F. L. *Lavoisier and the Chemistry of Life: An Exploration of Scientific Creativity.* Univ of Wisconsin Press, 1987. ISBN 0299099849.

Homberger, E., Reggiani, G. M., Sambeth, J., and Wipf, H. K. The Seveso Accident: Its Nature, Extent and Consequences. *Annals of Occupational Hygiene*, 22(4):327–370, January 1979. ISSN 0003-4878. doi: 10.1093/annhyg/22.4.327.

Hooper, H. H., Michel, S., and Prausnitz, J. M. M. Correlation of Liquid-liquid Equilibria for Some Water-organic Liquid Systems in the Region 20-250 °C. *Industrial and Engineering Chemistry Research*, 27(11):2182–2187, 1988.

Hukkerikar, A. S., Sarup, B., Ten Kate, A., Abildskov, J., Sin, G., and Gani, R. Group − contribution (GC) Based Estimation of Properties of Pure Components: Improved Property Estimation and Uncertainty Analysis. *Fluid Phase Equilibria*, 321:25–43, 2012. ISSN 03783812. doi: 10.1016/j.fluid.2012.02.010.

Ice calorimeter. Ice Calorimeter, late 18th century, January 2007.

International Electrotechnical Commission. Electrical Apparatus for Use in the Presence of Combustible Dusts, Part 2: Test Method, Section 3: Method for Determining Minimum Ignition Energy of Dust-Air Mixtures. Technical report, IEC, 1994.

Janes, A., Chaineaux, J., Carson, D., and Le Lore, P. A. {MIKE} 3 Versus {HARTMANN} Apparatus: Comparison of Measured Minimum Ignition Energy (MIE). *Journal of Hazardous Materials*, 152(1):32 – 39, 2008. ISSN 0304-3894. doi: http://dx.doi.org/10.1016/j.jhazmat.2007.06.066.

Janes, A., Chaineaux, J., Lesné, P., Mauguen, G., Petit, J. M., Sallé, B., and Marc, F. Mise en œuvre de la réglementation relative aux atmosphères explosives guide méthodologique. Technical report, Institut National de Recherche et de Sécurité, 2011.

Jaubert, J.-N. and Mutelet, F. VLE predictions with the Peng–Robinson equation of state and temperature dependent kij calculated through a group contribution method. *Fluid Phase Equilibria*, 224(2):285–304, 2004. ISSN 03783812. doi: 10.1016/j.fluid.2004.06.059.

Joback, K. G. Knowledge Bases for Computerized Physical Property Estimation. *Fluid Phase Equilibria*, 185(1-2):45–52, 2001. ISSN 03783812. doi: 10.1016/S0378-3812(01)00455-1.

Joback, K. G. and Reid, R. C. Estimation of Pure-component Properties from Group-contributions. *Chemical Engineering Communications*, 57(1-6):233–243, July 1987. ISSN 0098-6445. doi: 10.1080/00986448708960487.

Joint Committee for Guides in Metrology (JCGM). Evaluation of Measurement Data: Guide to the Expression of Uncertainty in Measurement. Technical Report September, Joint Committee for Guides in Metrology (JCGM), 2008.

Jun, Z., Xin-lu, C., Bi, H., and Xiang-dong, Y. Neural Networks Study on the Correlation between Impact Sensitivity and Molecular Structures for Nitramine Explosives. *Structural Chemistry*, 17(5):501–507, September 2006. ISSN 1040-0400. doi: 10.1007/s11224-006-9101-6.

Karcher, W. and Devillers, J. *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology.* Kluwe Academic Publishers, Dordrecht, the Netherlands, 1990.

Karelson, M., Lobanov, V. S., and Katritzky, A. R. Quantum-chemical Descriptors in QSAR / QSPR Studies. *Chemical Reviews*, 96:1027–1043, 1996. doi: 10.1021/cr950202r.

Katritzky, A. R., Stoyanova-Slavova, I. B., Dobchev, D. A., and Karelson, M. QSPR Modeling of Flash Points: an Update. *Journal of Molecular Graphics & Modelling*, 26(2):529–36, September 2007. ISSN 1093-3263. doi: 10.1016/j.jmgm.2007.03.006.

Keshavarz, M. H. Simple Method for Prediction of Activation Energies of the Thermal Decomposition of Nitramines. *Journal of Hazardous Materials*, 162(2-3):1557–62, March 2009a. ISSN 0304-3894. doi: 10.1016/j.jhazmat.2008.06.049.

Keshavarz, M. H. A New Method to Predict Activation Energies of Nitroparaffins. *Indian Journal of Engineering & Materials Sciences*, 16(December):429–432, 2009b.

Keshavarz, M. H. and Ghanbarzadeh, M. Simple Method for Reliable Predicting Flash Points of Unsaturated Hydrocarbons. *Journal of Hazardous Materials*, 193:335–41, October 2011. ISSN 1873-3336. doi: 10.1016/j.jhazmat.2011.07.044.

Keshavarz, M. H., Pouretedal, H. R., and Semnani, A. Reliable Prediction of Electric Spark Sensitivity of Nitramines: A General Correlation with Detonation Pressure. *Journal of Hazardous Materials*, 167(1-3):461–6, August 2009a. ISSN 1873-3336. doi: 10.1016/j.jhazmat.2009.01.009.

Keshavarz, M. H., Pouretedal, H. R., and Semnani, A. Relationship between Thermal Stability and Molecular Structure of Polynitro Arenes. *Sciences-New York*, 16(February):61–64, 2009b. ISSN 09714588.

Kletz, T. *What Went Wrong?* Gulf Professional Publishing, Houston, fourth edition edition, 1999. ISBN 978-0-88415-920-9. doi: http://dx.doi.org/10.1016/B978-088415920-9/50004-9.

Kletz, T. Inherently Safer Design-Its Scope and Future. *Process Safety and Environmental Protection*, 81(6):401–405, 2003. ISSN 09575820. doi: 10.1205/095758203770866566.

Klincewicz, K. M. and Reid, R. C. Estimation of Critical Properties with Group Contribution Methods. *AIChE Journal*, 30(1):137–142, January 1984. ISSN 0001-1541. doi: 10.1002/aic.690300119.

Klos, J., Nowicki, P., and Cizmarik, J. Thermal Parameters of Phenylcarbamic Acid Derivatives Using Calculated Molecular Descriptors with MLR and ANN: Quantitative Structure-Property Relationship Studies. *Journal of Thermal Analysis and Calorimetry*, 91(1):203–212, 2008. ISSN 13886150. doi: 10.1007/s10973-007-8354-7.

Kohn, W. and Sham, L. J. Self-consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133.

Kolská, Z., Zábranský, M., and Randová, A. *Group Contribution Methods for Estimation of Selected Physio-Chemical Properties of Organic Compounds.* Ricardo Morales-Rodriguez, 2012.

Larsen, B., Rasmussen, P., and Fredenslund, A. Modified UNIFAC Group-contribution Model for Prediction of Phase Equilibria and Heats of Mixing. *Industrial and Engineering Chemistry Research*, 26(11):2274–2286, 1987.

Lazzús, J. A. Neural Network/particle Swarm Method to Predict Flammability Limits in Air of Organic Compounds. *Thermochimica Acta*, 512(1-2):150–156, January 2011. ISSN 00406031. doi: 10.1016/j.tca.2010.09.018.

Lazzús, J. A. A Group Contribution Method to Predict the Thermal Decomposition Temperature of Ionic Liquids. *Journal of Molecular Liquids*, 168:87–93, 2012. ISSN 01677322. doi: 10.1016/j.molliq.2012.01.011.

Li, X.-R. and Koseki, H. Thermal Decomposition Kinetic of Liquid Organic Peroxides. *Journal of Loss Prevention in the Process Industries*, 18(4-6):460–464, July 2005. ISSN 09504230. doi: 10.1016/j.jlp.2005.07.003.

Liang, Y., Xu, Q.-S., Li, H.-D., and Cao, D.-S. Support Vector Machines and QSAR/QSPR. In *Support Vector Machines and Their Application in Chemistry and Biotechnology*, chapter 6, pages 115–147. CRC Press, May 2011. ISBN 978-1-4398-2127-5. doi: doi:10.1201/b10911-7.

Lothrop, W. C. and Handrick, G. R. The Relationship Between Performance and Compounds. *Chemical Reviews*, 44:419–445, 1948.

Lu, Y., Ng, D., and Mannan, M. S. Prediction of the Reactivity Hazards for Organic Peroxides Using the QSPR Approach. *Industrial & Engineering Chemistry Research*, 50(3):1515–1522, February 2011. ISSN 0888-5885. doi: 10.1021/ie100833m.

Lydersen, A. L. Estimation of Critical Properties of Organic Compounds by the Method of Group Contributions. Technical report, University of Wisconsin. Engineering Experiment Station., Madison, WI, 1955.

Mage, L. Machine Learning for Thermal Decomposition Prediction from Group Contribution. Master's thesis, École polytechnide fédérale de Lausanne, Lausanne, 2015.

Mallakpour, S., Hatami, M., Khooshechin, S., and Golmohammadi, H. Evaluations of Thermal Decomposition Properties for Optically Active Polymers Based on Support Vector Machine. *Journal of Thermal Analysis and Calorimetry*, 116(2):989–1000, 2014. ISSN 1388-6150. doi: 10.1007/s10973-013-3587-0.

Man, C. K. and Harris, M. L. Participation of Large Particles in Coal Dust Explosions. *Journal of Loss Prevention in the Process Industries*, 27:49 – 54, 2014. ISSN 0950-4230. doi: http://dx.doi.org/10.1016/j.jlp.2013.11.004.

Marrero, J. and Gani, R. Group-contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilibria*, 183-184:183–208, 2001. ISSN 03783812. doi: 10.1016/S0378-3812(01)00431-9.

Marrero-Morejón, J. and Pardillo-Fontdevila, E. Estimation of Pure-compound Properties Using Group-interaction Contributions. *AIChE Journal*, 45(3):615–621, 1999. ISSN 00011541. doi: 10.1002/aic.690450318.

MathWorks. MATLAB Version R2014b (8.4.0.150421). www.mathworks.com, 2014.

McCarty, L. S., Hudson, P. V., Craig, G. R., and Kaiser, K. L. E. The Use of Quantitative Structure-Activity Relationships to Predict the Acute and Chronic Toxicities of Organic Chemicals to Fish. *Environmental Toxicology and Chemistry*, 4(5):595–606, 1985. ISSN 07307268.

Melhem, G. A. and Shanley, E. S. On the Estimation of Hazard Potential for Chemical Substances. *Process Safety Progress*, 15(3):168–172, 1996. ISSN 1066-8527. doi: 10.1002/prs.680150311.

Milano Chemometrics & QSAR Research Group. Molecular Descriptors. www.moleculardescriptors.eu/softwares/softwares.htm, 2007. Accessed: 2016-01-08.

Millán, J. D. R. and Chavarriaga, R. Data Analysis and Model Classification. Lecture 10: Self-organizing Maps. Course, 2013.

Mitchell, B. E. and Jurs, P. C. Prediction of Autoignition Temperatures of Organic Compounds from Molecular Structure. *Journal of Chemical Information and Computer Sciences*, 37: 538–547, 1997.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, 2012. ISBN 026201825X, 9780262018258.

Murphy, M. R., Singh, S. K., and Shanley, E. S. Computationally Evaluate Self-Reactivity Hazards. *CEP*, pages 54–61, 2003.

Nefati, H., Cense, J.-M., and Legendre, J.-J. Prediction of the Impact Sensitivity by Neural Networks. *Journal of Chemical Information and Computer Sciences*, 36(4):804–810, 1996. ISSN 1549-9596. doi: 10.1021/ci950223m.

Nendza, M. and Russom, C. L. QSAR Modelling of the ERL-D Fathead Minnow Acute Toxicity Database. *Xenobiotica*, 21:147–170, August 1991. doi: 10.3109/00498259109039458.

NFPA. Nfpa 68: Standard on explosion protection by deflagration venting. National Fire Protection Association, Quincy, Massachusetts, 1986.

Nieto-Draghi, C., Fayet, G., Creton, B., Rozanska, X., Rotureau, P., de Hemptinne, J.-C., Ungerer, P., Rousseau, B., and Adamo, C. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chemical Reviews*, 2015. ISSN 1520-6890. doi: 10.1021/acs.chemrev.5b00215.

Ozawa, T. A New Method of Analyzing Thermogravimetric Data. *Bulletin of the Chemical Society of Japan*, 38(11):1881–1886, 1965. ISSN 0009-2673. doi: 10.1246/bcsj.38.1881.

Palm, K., Luthman, K., Ungell, A. L., Strandlund, G., Beigi, F., Lundahl, P., and Artursson, P. Evaluation of Dynamic Polar Molecular Surface Area As Predictor of Drug Absorption: Comparison with Other Computational and Experimental Predictors. *Journal of Medicinal Chemistry*, 41(27):5382–92, December 1998. ISSN 0022-2623. doi: 10.1021/jm980313t.

Pan, Y., Jiang, J., Wang, R., Cao, H., and Cui, Y. Predicting the Auto-ignition Temperatures of Organic Compounds from Molecular Structure Using Support Vector Machine. *Journal of Hazardous Materials*, 164(2-3):1242–9, May 2009. ISSN 1873-3336. doi: 10.1016/j.jhazmat.2008.09.031.

Pan, Y., Jiang, J., Ding, X., Wang, R., and Jiang, J. Prediction of Flammability Characteristics of Pure Hydrocarbons from Molecular Structures. *AIChE Journal*, 56(3):690–701, 2010. ISSN 00011541. doi: 10.1002/aic.12007.

Pan, Y., Jiang, J. C., Wang, R., and Jiang, J. J. Predicting the Net Heat of Combustion of Organic Compounds from Molecular Structures Based on Ant Colony Optimization. *Journal of Loss Prevention in the Process Industries*, 24(1):85–89, January 2011. ISSN 09504230. doi: 10.1016/j.jlp.2010.11.001.

Panthananickal, A., Hansch, C., Leo, A., and Quinn, F. R. Structure-activity Relationships in Antitumor Aniline Mustards. *Journal of Medicinal Chemistry*, 21(1):16–26, 1978. ISSN 00222623.

Pareto, V. *Manual of Political Economy*. Kelley, A. M., New York, 1971. ISBN 9780678008812.

Peng, D.-Y. and Robinson, D. B. A New Two-Constant Equation. In *4th International Heat Transfer Conference*, volume 15, pages 59–64, 1976. ISBN 9780735410060. doi: 10.1063/1.3686399.

Pesatori, A. C., Consonni, D., Rubagotti, M., Grillo, P., and Bertazzi, P. A. Cancer incidence in the population exposed to dioxin after the "Seveso accident": twenty years of follow-up. *Environmental health : a global access science source*, 8:39, January 2009. ISSN 1476-069X. doi: 10.1186/1476-069X-8-39.

Petrukhin, R., Karelson, M., Katritzky, A. R., Lomaka, A., Petrukhina, I., and Tatham, D. CODESSA Pro Software Version 1.0 RC2. www.codessa-pro.com, 2001.

Polishchuk, P. G., Kuźmin, V. E., Artemenko, A. G., and Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/ QSPR Models. *Molecular Informatics*, 32(9-10):843–853, 2013. ISSN 18681751. doi: 10.1002/minf.201300029.

Pople, J. A. and Segal, G. A. Approximate Self-consistent Molecular Orbital Theory II. Calculations With Complete Neglect Of Differential Overlap. *The Journal of Chemical Physics*, 43 (10):S136–S151, 1965.

Pople, J. A. and Segal, G. A. Approximate Self-consistent Molecular Orbital Theory. III. CNDO Results for AB2 and AB3 Systems. *The Journal of Chemical Physics*, 44(9):3289–3296, 1966.

Pople, J. A., Beveridge, D. L., and Dobosh, P. A. Approximate Self-consistent Molecular-orbital Theory. V. Intermediate Neglect Of Differential Overlap. *The Journal of Chemical Physics*, 47 (6):2026–2033, 1967.

Randić, M. On Molecular Branching. *Acta Chimica Slovenica*, 44(1):57–77, 1997. ISSN 13180207. doi: 10.1021/ja00856a001.

Rawlinson, C. Differential Scanning Calorimetry. Course School of Pharmacy, 2006.

Reyes, O. J., Patel, S. J., Mannan, M. S., and Kay O 'connor, M. Quantitative Structure Property Relationship Studies for Predicting Dust Explosibility Characteristics (K st , P max) of Organic Chemical Dusts. *Industrial & Engineering Chemistry*, 50:2373–2379, 2011. doi: 10.1021/ie1013663.

Rogers, R. L., Broeckmann, B., and Maddison, N. Explosion Safety Document for the Atex 137 Directive – New Name for a Fire and Explosion Hazard Assessment ? *IChemE Hazards XVII Symposium Series*, 149:431–443, 2003.

Rohrbaugh, R. H. and Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/activity and Structure/property Relationships. *Analytica Chimica Acta*, 199:99–109, 1987. ISSN 0003-2670. doi: doi:DOI:10.1016/S0003-2670(00)82801-9.

Rojas, R. *Neural Networks: A Systematic Introduction.* Springer-Verlag New York, Inc., New York, NY, USA, 1996. ISBN 3-540-60505-3. doi: 10.1016/0893-6080(94)90051-5.

Rowley, R. L. and Wilding, W. V. Estimation of the Flash Point of Pure Organic Chemicals from Structural Contributions. *Process Safety Progress*, 29(4):353–358, 2010. doi: 10.1002/prs10401.

Saito, Y., Saito, K., and Atake, T. Base Line Drawing for the Determination of the Enthalpy of Transition in Classical DTA, Power-compensated DSC and Heat-flux DSC. *Thermochimica Acta*, 104:275–283, 1986. ISSN 00406031. doi: 10.1016/0040-6031(86)85202-9.

Saraf, S. R., Rogers, W. J., and Mannan, M. S. Prediction of Reactive Hazards Based on Molecular Structure. *Journal of Hazardous Materials*, 98(1-3):15–29, March 2003. ISSN 0304-3894. doi: 10.1016/S0304-3894(02)00314-X.

Sarge, S. M. Determination of Characteristic Temperatures with the Scanning Calorimeter. *Thermochimica Acta*, 187:323–334, 1991. ISSN 00406031. doi: 10.1016/0040-6031(91)87208-E.

Sarge, S. M., Höhne, G. W. H., and Hemminger, W. Methods of Calorimetry. In *Calorimetry: Fundamentals, Instrumentation and Applications*, number 2003 in 1, chapter 1, pages 9–18. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2014a. doi: 10.1002/9783527649365.part1.

Sarge, S. M., Höhne, G. W. H., and Hemminger, W. Calorimeters. In *Calorimetry: Fundamentals, Instrumentation and Applications*, chapter 7, pages 125–211. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2014b. doi: 10.1002/9783527649365.part1.

Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 0090-5364. doi: 10.1214/aos/1176344136.

Seaton, W. H., Freedman, E., and Treweek, D. N. CHETAH: The ASTM Chemical Thermodynamic and Energy Release Evaluation Program. *D.S 51*, 1974.

Selassie, C., Verma, R. P., and Abraham, D. J. History of Quantitative Structure-Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, volume 1, pages 1–48. John Wiley & Sons, Inc., 2003. ISBN 0471266949. doi: 10.1002/0471266949.bmc001.pub2.

Shanley, E. S. and Melhem, G. A. A Review of ASTM CHETAH 7.0 Hazard Evaluation Criteria. *Journal of Loss Prevention in the Process Industries*, 8(5):261–264, 1995. ISSN 09504230. doi: 10.1016/0950-4230(95)00014-R.

Sourour, S. and Kamal, M. R. Differential Scanning Calorimetry of Epoxy Cure: Isothermal Cure Kinetics. *Thermochimica Acta*, 14(1-2):41–59, 1976. ISSN 00406031. doi: 10.1016/0040-6031(76)80056-1.

Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods I. Method. *Journal of Computational Chemistry*, 10(2):209–220, 1989. ISSN 1096-987X. doi: 10.1002/jcc.540100208.

Stoessel, F. *Thermal Safety of Chemical Processes: Risck Assessment and Process Design.* Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, February 2008. ISBN 9783527621606. doi: 10.1002/9783527621606.

Stoessel, F. Static Electricity as Ignition Source. Course Chemical process safety EPFL Master in Chemical Engineering, 2014.

Stramaglia, S., Angelini, L., Marangi, C., Nitti, L., and Pellicoro, M. Statistical Physics and the Clustering Problem. In *New Directions in Statistical Physics*, pages 253–272. Wille, Luc T. Springer Berlin Heidelberg, 2004. doi: 10.1007/978-3-662-08968-2.

Suter, G. Protection against Explosions. Training Course Swissi Process Safety, 2008.

Suzuki, T. Quantitative Structure-Property Relationships for Auto-Ignition Temperatures of organic compounds. *Fire and Materials*, 18:81–88, 1994.

Swiss Confederation. Ordinance on Protection against Major Accidents, 1991.

Tiegs, D., Gmehling, J., Rasmussen, P., and Fredenslund, A. Vapor-liquid Equilibria by UNIFAC Group Contribution. 4. Revision and Extension. *Industrial and Engineering Chemistry Research*, 26(1):159–161, 1987.

Todeschini, R. and Consonni, V. *Handbook of Molecular Descriptors.* Wiley-VCH Verlag GmbH, 2012. ISBN 3527304657. doi: 10.1002/9783527654307.ch1.

Valenzuela, E. M., Vázquez-Román, R., Patel, S., and Mannan, M. S. Prediction Models for the Flash Point of Pure Components. *Journal of Loss Prevention in the Process Industries*, 24(6): 753–757, November 2011. ISSN 09504230. doi: 10.1016/j.jlp.2011.04.010.

Wang, G. C. S. and Jain, C. L. *Regression Analysis: Modeling & Forecasting*. Graceway Pub., New York, 2003. ISBN 9780932126504.

Wang, R., Jiang, J. C., and Pan, Y. QSPR Study on Electric Spark Sensitivity of Nitro Arenes. *Advanced Materials Research*, 284-286:197–200, July 2011. ISSN 1662-8985. doi: 10.4028/www.scientific.net/AMR.284-286.197.

Watson, E. S., O'Neill, M. J., Justin, J., and Brenner, N. A Differential Scanning Calorimeter for Quantitative Differential Thermal Analysis. *Analytical Chemistry*, 36(7):1233–1238, 1964. ISSN 0003-2700. doi: 10.1021/ac60213a019.

Witten, I. H., Frank, E., and Hall, M. A. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, second edi edition, 2011. ISBN 0120884070.

Zeman, S., Pelikán, V., and Majzlík, J. Electric Spark Sensitivity of Nitramines . Part I . Aspects of Molecular Structure *). *Central Europe Journal of Enrgetic Materials*, 3(April):27–44, 2006.

Zhang, Y., Pan, Y., Jiang, J., and Ding, L. Prediction of Thermal Stability of Some Reactive Chemicals Using the QSPR Approach. *Journal of Environmental Chemical Engineering*, 2(2): 868–874, 2014. ISSN 22133437 (ISSN). doi: 10.1016/j.jece.2014.02.020.

Zhi, C., Cheng, X., and Zhao, F. The Correlation between Electric Spark Sensitivity of Polynitroaromatic Compounds and Their Molecular Electronic Properties. *Propellants, Explosives, Pyrotechnics*, 35(6):555–560, 2010. ISSN 07213115. doi: 10.1002/prep.200900092.

Abbreviations

Acronyms

AIC	Akaike Information Criterion
AIT	Auto-Ignition Temperature
AM1	Austin Model 1
ANN	Artificial Neural Network
ARC	Accelerating Rate Calorimeter
ARD	Average Relative Deviation
ASTM	American Society of the International Association for Testing and Materials
BIC	Bayesian Information Criterion
CART	Calculated Adiabatic Rise of Temperature
CHETAH	Chemical Thermodynamic and Energy Evaluation Program
CPU	Central Processing Unit
DFT	Density Functional Theory
DHBT	3,4-Dihydro-3-hydroxy-4-oxo-1,2,3-benzotriazine
DSC	Differential Scanning Calorimetry
DSC	Differential Scanning Calorimetry
DTA	Differential Thermal Analysis
EGA	Evolved Gas Analysis
EHT	Extended Hückel Theory
FWHM	Full Width at Half Maximum
GA	Genetic Algorithm
GC^+	Marrero-Gani Group Contribution
GCM	Group Contributions Method
GMM	Gaussian Mixture Models
IEC	International Electrotechnical Committee
LEL	Lower Explosive Limit
MART	Maximum Adiabatic Rise of Temperature
MIE	Minimum Ignition Energy
MLR	Multiple Linear Regression

MNDO	Modified Neglect of Diatomic Overlap
MTSR	Maximum Temperature of Synthesis Reaction
OB	Oxygen Balance
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PLS	Partial Least Squares
PM3	Parametrized Model 3
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure-Property Relationship
RC	Repeatability Coefficients
RMSE	Root Mean Square Error
RTP	Room Temperature and Pressure
SCF	Self-Consistent Field
SOM	Self-Organizing Maps
SSE	Sum of Squared Errors
SSR	Sum of Squares of the Regression
SST	Total Sum of Squares
SVM	Support Vector Machine
$\mathbf{T_{D24}}$	Temperature at which $\mathrm{TMR_{AD}}$=24h
TCDD	2,3,7,8-tetrachlorodibenzo-p-dioxin
$\mathbf{TMR_{AD}}$	Time to Maximum Rate in Adiabatic conditions
TZC	Tolerance Zone Coefficients
UEL	Upper Explosive Limit

Symbols

Symbol	Name	Units
a	Asymmetric Factor	-
A	Activation Function	-
C	Heat Capacity	$\mathrm{J\,K^{-1}}$
C	Concentration	$\mathrm{mol\,m^{-3}}$
d	Particle Size	m
D	Diameter	m
e	Residual Error	-
E	Error Function	-
E_S	Statistic Energy	mJ
E_a	Activation Energy	$\mathrm{J\,mol^{-1}}$
FW	Full Width Maximum Height	°C
G	Group Contribution	-
ΔH	Enthalpy	$\mathrm{J\,g^{-1}}$ or $\mathrm{J\,mol^{-1}}$
i	Iteration	-
k	Proportionality Factor	-
k	Kinetic Rate Constant	$\mathrm{s^{-1}}$
k	Number of Cluster	-
k_o	Preexponential Factor (Arrhenius Law)	$\mathrm{s^{-1}}$
M	Molecular Weight	$\mathrm{g\,mol^{-1}}$
n	Reaction Order	-
n	Molar Quantity	mol
n	Number of Observations	-
O	Output Function	-
P	Pressure	Pa
p	Number of Parameters	-
R	Universal Gas Constant	$\mathrm{J\,mol^{-1}\,K^{-1}}$
r	Reaction Rate	$\mathrm{mol\,m^{-3}\,s}$
T	Temperature	°C or K
V	Volume	$\mathrm{m^3}$
x	Structural Descriptor Value	-

Greek Symbols

Symbol	Name	Units
α	Parameter Coefficient	-
β	Scan Rate	$\mathrm{K\,s^{-1}}$
χ	Randiç Index	-
Φ	Heat Flow Rate	$\mathrm{W\,g^{-1}}$ or $\mathrm{W\,mol^{-1}}$
σ	Standard Deviation	-
τ	Time Constant	s

Subscripts

Subscripts	Meaning
adj	Adjusted
b	Boiling
c	Corrected
crit	Critical
F	Furnace
f	Final
i	Initial
M	Measurement Point
max	Maximum, refers to Peak's Maximum
o	Onset
p	Process
P	Programmed
R	Reference
r	Reaction
S	Sample
th	Thermal

Copyright Credits

- Figure 4.1 (a) reproduced from:

  Science Museum / Science & Society Picture Library
  Ice calorimeter, late 18th century.
  URL http://www.sciencemuseum.org.uk/images/I059/10325932

  Copyright ©2004 Science & Society Picture Library.

- Figure 4.1 (b) reproduced from:

  S. M. Sarge, G. W. H. Hohne, and W. Hemminger.
  In *Calorimetry: Fundamentals, Instrumentation and Applications*
  Figure 1.1 Calorimeter of Lavoisier and Laplace (according to Kleiber, 1975).

  Copyright ©2014, John Wiley and Sons.

Nadia BAATI

Route du Bois, 10
1024 Ecublens, Switzerland

nadia.baati@gmail.com

+41 79 786 70 53

13.07.1988

Single

Tunisian

Education

> Doctoral Studies in Chemistry and Chemical Engineering Program 02/2012 – 03/2016
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

> Master of Sciences in Chemical Engineering and Biochemistry 09/2009 – 8/2011
Minor degree in Management and Technology Entrepreneurship
EPFL, Section of Chemistry and Chemical Engineering

> Bachelor of Sciences in Chemistry 10/2006 – 08/2010
EPFL, Section of Chemistry and Chemical Engineering

> Baccalauréat S 09/2005 – 06/2006
French High School Diploma in Advanced Sciences, with Honors
Lycée Français Saint-Louis of Stockholm, Sweden

Professional Experiences

EPFL – Section of Chemistry and Chemical Engineering

> Doctoral Assistant – supervised by Dr. T. Meyer & Prof. F. Stoessel 02/2012 – 03/2016
"Predictive Models for Thermal Stability of and Explosive Properties of Chemicals from Molecular Structure"

Swissi Process Safety GmbH

> Master Thesis Student – supervised by Prof. F. Stoessel 03/2011 – 08/2011
Feasibility study on thermal safety prediction from chemical structure

> Research Intern at the Process Safety Group 10/2011 – 12/2011
Support and participate in ongoing projects
Course material for continuing education

SIPHAT – Society of Pharmaceutical Industries of Tunisia

> Technical Intern at the Quality Control Service 08/2008
Support and participate in quality analysis and evaluation of various medicine and sensitive packaging

Languages Skills

> French	Bilingual
> Arabic	Bilingual
> English	Fluent
> German	Beginner
> Swedish	Beginner

IT Skills

General

> C/C++ programming
> VBA programming
> Microsoft Office Suite
> LaTeX editing

Technical

> MATLAB
> Codessa Pro
> AKTS
> Aspen Plus