





UNIVERSITY OF MISKOLC ANTAL KERPELY DOCTORAL SCHOOL OF MATERIALS AND SCIENCE TECHNOLOGY AT THE FACULTY OF MATERIALS AND CHEMICAL ENGINEERING PROF.Dr. MERTINGER VALÉRIA

Thesis of Ph.D. Dissertation

titled as

A Data-Driven Modelling Approach for Heterogeneous Catalysis Research

presented by

Alexandra Jakab-Nácsa

SUPERVISOR

PROF.DR. BÉLA VISKOLCZ

BORSODCHEM EXPERT

László Farkas

INSTITUTE OF CHEMISTRY UNIVERSITY OF MISKOLC MISKOLC, HUNGARY 2024







1. INTRODUCTION

Looking at the centuries-old history of the science of catalysis, the hope of fully recognising and understanding it seems to be slipping away. Despite today's technological tools, which allow a more detailed study of the phenomenon, the infinite pile of accumulated data creates confusion in clear vision. However, artificial intelligence has opened new horizons on data transformation into information and into knowledge¹. Catalyst design is an area in which computational chemistry and machine learning techniques lead to outstanding results².

My doctoral dissertation aims to emphasize the data-driven catalyst design. The objective is to elevate the catalyst basic research results to the next level on the path to industrial application by using them in a practice-oriented way.

My research work is based on two pillars, whose database was built from more than 15.000 data points published in scientific literature. The first pillar is the establishment of MIRA21 (**MI**skolc **RA**nking 20**21**) model, a functional and practical mathematical model of catalyst characterization and exact comparison of each other. The second pillar of our research is the application of EDA (Exploratory Data Analysis), which refers to preliminary work on predicting catalyst composition through machine learning.

2. OBJECTIVE OF THE DOCTORAL STUDY



Figure 1 From data to knowledge - schematic of PhD study

My doctoral thesis summarizes my research in the field of data-based catalyst design and contributes to new methods of catalyst research. The goal is to build a hierarchy of data-information-knowledge from the data point through various methods and achieve catalyst design strategies in advance (Figure 1). The main objective of catalyst design is to optimize catalyst composition through analysing literature and experimental data, using various mathematical models and computer software to predict and apply catalyst composition.

During literature research, the structure of specialized literature publications, how research results are discussed, and the data content of publications became known. It was determined which parameters are useful for extracting information from publications and which are appropriate for the characterization of catalysts.

The literature contains inconsistent data that are difficult to compare each other directly. Since data sets are difficult to handle in this way, in collaboration to the University of Miskolc Catalyst Group, a database was created and then standardized with the MIRA21 model into a single quantifiable data per catalyst. The catalysts have thus become rankable and classable. In the next step of information collection, data were cleaned up and parameters selected, and the correlation between parameters and factors influencing the composition of optimal catalysts were investigated. As the process of exploring data analysis continued, data appeared to be very well used by artificial intelligence, especially in machine learning. Thus, catalyst design guidelines and machine learning data sets were created as a result of data analysis processes before machine learning.

Throughout my doctoral studies, new knowledge was gained in the field of catalysis informatics, which helps to promote the use of artificial intelligence in the design of catalysts. Given our considerable accumulation of knowledge on the semantic aspects of publications based on human learning, this research also forms a solid basis for utilizing the potential of semantic searches.

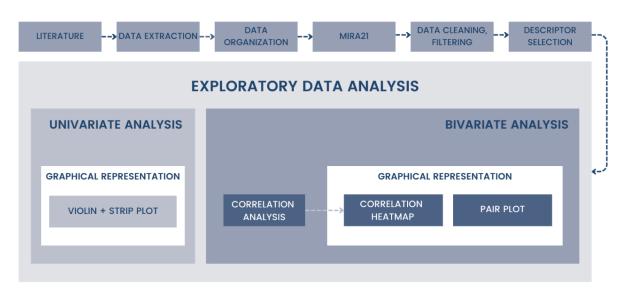
3. Methods

3.1. MIRA model

The Miskolc Ranking 2021 (MIRA21) methodology is a multi-step process for comparing catalysts employed in a specific reaction and identifying new patterns among parameters characterizing catalytic processes. This is a systematic approach to create a simple, general parameter from parameter vectors for the comparison and analysis of catalysts data.

The main purpose of the MIRA21 model is to provide a standard for assessing the "goodness" of a catalyst with objective numerical data and to compare and classify it accordingly. The classification promotes the efficient selection of appropriate links relevant to support the design of a new catalyst or improve existing ones. The comparison of special catalysts for a reaction allows the monitoring of trends in research and development. Standardization of access data in MIRA21 will also promote accurate and consistent data in future publications.

The application field of the model has been reduced to catalytic reactions, mainly heterogeneous catalytic reactions. This methodology was developed using the hydrogenation reaction of aromatic nitrogen compounds.



3.2. Process of EDA in this study

Figure 2 Process of our exploratory data analysis

The aim of the study is also to investigate how to obtain conscious data analysis from the collection of data guided by chemical intuition and to obtain results suitable for the construction of machine learning algorithms.

Exploratory data analysis (EDA) is the application of several statistical techniques aimed at investigating, describing, and summarizing the nature of data³. This allows us to identify possible errors, reveal the existence of an outlier, check the relationship between

variables (correlations) and their possible redundancy, and conduct a descriptive analysis of data using graphical representations and summaries of the most important aspects⁴. EDA of previous catalytic data reveals the exploration of correlations between the physicochemical properties and performance of catalysts^{5–11}.

Data analysis can be described in a multi-step process, as a result of which the necessary database is prepared for machine learning and a catalyst design strategy can be complied (Figure 2). The EDA was carried out in python programming environment by using NumPy, Pandas, Seaborn, and Matplotlib libraries¹²⁻¹⁵.

4. New scientific results

1. A MATHEMATICAL FRAMEWORK WAS DEVELOPED FOR CHARACTERIZING CATALYSTS BY A SINGLE QUANTITATIVE DATA.



Figure T 1 Visualization of MIRA21 number

As a result of the research MIRA (MIskolc RAnking) model was established, which can be used to characterize catalysts described in scientific research using a single quantitative data. 13 parameters that can be used to characterize catalyst were determined. 13-parameter descriptor system was created, each characterized by quantifiable data. The parameters are divided into four groups, namely the variables that characterize the performance of the catalyst, the parameters that describe the reaction conditions, the properties of catalyst and the sustainability parameters. A mathematical procedure was developed that allows easy identification and comparison of catalysts. In mathematical procedures, parameters are normalized, then weighted by group, and the number of MIRAs is generated using mathematical formulas. Catalysts ranked and classified become comparative by the number of MIRAs.

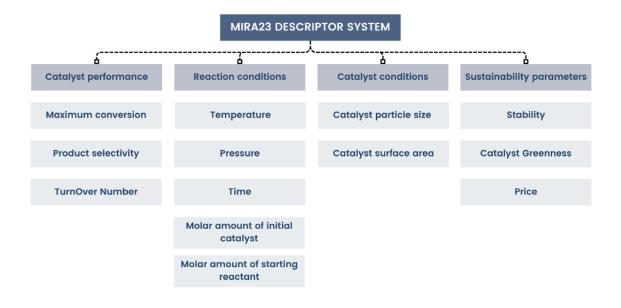


Figure T 2 Descriptor system of the model

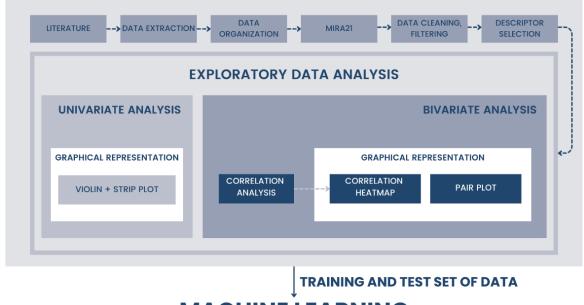
2. THE DESCRIPTOR SYSTEM AND THE UNIQUE WEIGHT FACTORS HAVE BEEN REVIEWED ON THE BASIS OF WHICH A GENERAL WEIGHT FACTOR HAS BEEN PROPOSED, MAKING BOTH THE RANKING AND CLASSIFICATION CLEAR.

During PhD work, the descriptor system was revised by Exploratory Data Analysis. Based on the evaluation, version MIRA23 was created by slightly modifying the MIRA21 model. The weighting of all parameters was validated.

The system is suitable for the characterization of catalysts for hydrogenation of nitrobenzene, and it can be applied on any catalytic system based on the test^{16,17}.

3. MIRA21(23) DATA- AND CATALYST RATING SYSTEM, TOGETHER WITH MACHINE LEARNING, CAN BE USED TO CREATE NEW CATALYST DESIGN STRATEGIES.

The steps taken in doctoral thesis provide a new catalyst design strategy. The first step is the application of the MIRA model, then the Exploratory Data Analysis and the application Machine Learning methods.



MACHINE LEARNING

Figure T 3 Catalyst design strategy

4. For specific active metals, it is possible to determine the concentration range to achieve a conversion level higher than a certain level.

Based on the examination of the parameter system of the developed multivariate method, chemical information was produced with the help of Exploratory Data Analysis. In the analysis of joint plots, groups of different compositions and performance characteristics can be distinguished depending on the composition of the catalyst. Based on Figure T4, in case of iron active component, a conversion of more than 70% can be achieved at a concentration of 20-22%.

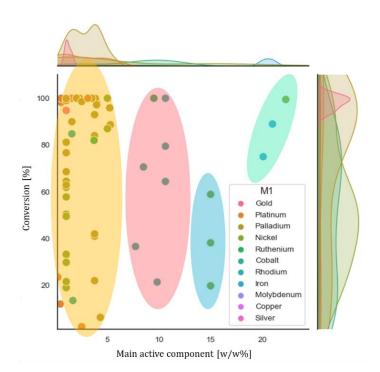


Figure T 4 Joint plot of Active component concentration (M12 w/w%) and Conversion classified by catalyst active component(M1)

5. IN CATALYST RESEARCH, SUSTAINABILITY PARAMETERS WERE FIRST DEFINED AS DESCRIPTOR INDICATORS OF CATALYSTS USING MATHEMATICAL PROCEDURES.

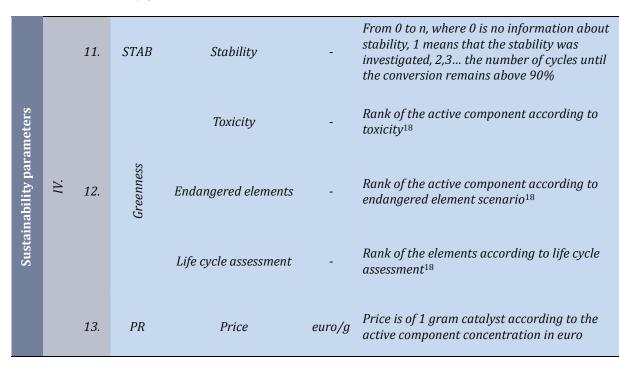


Table T 1 Sustainability parameters

5. SUMMARY

JOURNEYING FROM TRADITIONAL LITERATURE RESEARCH TO MACHINE LEARNING

The phrase best summarizes the last four years of my travel during my doctoral work. My doctoral research commenced with the development of aromatic nitro compounds catalytic hydrogenation processes and the exploration of corresponding hydrogenation catalysts. Among the traditional literature research, numerous inquiries arose, leading to the initiation of data collection.

The process of data collection evolved into a comprehensive catalyst database. Data comparison transformed into the MIRA (Miskolc Ranking) model, while data analysis transitioned an Exploratory Data Analysis. Subsequently, data utilization progressed into Machine Learning.

The catalyst database currently contains over 15,000 data points, inclusive 450 distinct catalyst combinations and experimental hydrogenation tests on aromatic nitrogen compounds. This extensive catalyst library comprises scientific sources and their quality, publication dates, the composition of the catalyst, and the parameters that characterize the catalyst.

These parameters, along with the database itself, are the basis of the MIRA model. Miskolc Ranking methodology embodies a multi-step approach for identifying novel patterns, potentially useful and interpreted in data collected for the selected catalytic reaction. This functional and practical mathematical framework aids in precise catalyst characterization and comparison. Employing a 15-parameter system, catalysts were characterized via model, the data was standardized and weighed, and then created a single quantifiable data using mathematical formulas. The catalysts were ranked and classified according to the MIRA number obtained so far to facilitate comparison. This model extends to catalysts used in hydrogenation reactions of nitrobenzene and dinitrotoluene.

A comprehensive review of the model was carried out through Exploratory Data Analysis. The data set was examined, carried out data cleaning and data filtering. Based on correlation analysis, modifications were made to the model, and an expanded database was established for validation. Using the MIRA model, best catalyst combinations were determined. Exploratory data analysis also produced additional chemical insights, revealing patterns within parameters pairs and catalyst compositions. Completed data sets have been transferred to machine learning applications and form training and testing sets for various parameter predictions.

In essence, extracting valuable chemical information from data leads to profound acquisition of knowledge. Personally, this journey has provided knowledge of multivariate data analysis and promoted understanding of artificial intelligence applications, thus enriching my knowledge of catalyst hydrogenation of aromatic nitro compounds.

Furthermore, my doctoral research has led to fruitful results. The establishment of MIRA highlighted the shortcomings of contemporary publishing practices. As a result, members of the Institute of Chemistry have aligned themselves with the standard MIRA descriptor systems in their publications.

The MIRA model also extends to a successful method of literature research, validated through education implementation. University students use the system t identify process weaknesses, which leads to valuable insight into the application of the model. Work is currently underway to use this model for the data base of methanol production from carbon dioxide.

Cooperation between the Institute of Chemistry and computer scientists aims to exploit artificial intelligence for chemical applications. I look forward to the continuation of this collaborative effort.

Overall, my work summarizes the latest research findings and introduces novel catalyst design approach, based on data-driven research. It also provides tools for the development of industrial catalysts, bridges the gap between the results of scientific research and the practical application through the MIRA model. The consideration of sustainability parameters clearly shifts catalyst qualification towards application. This work serves as an example of data analysis and the integration of artificial intelligence to optimize industrial processes.

6. LIST OF PUBLICATIONS

PUBLICATIONS RELATED TO THE SUBJECT OF THE DISSERTATION

- Alexandra Jakab-Nácsa, Emőke Sikora, Ádám Prekob, László Vanyorek, Milán Szőri, Renáta Boros Zsanett, Károly Nehéz, Martin Szabó, László Farkas, Béla Viskolcz, Comparison of Catalysts with MIRA21 Model in Heterogeneous Catalytic Hydrogenation of Aromatic Nitro Compounds, *MDPI Catalysts*, 2022, 12 (5), 467, https://doi.org/10.3390/catal12050467, IF 3.9
- Alexandra Jakab-Nácsa, Viktória Hajdú, László Vanyorek, László Farkas, Béla Viskolcz, Overview of Catalysts with MIRA21 Model in Heterogeneous Catalytic Hydrogenation of 2,4-Dinitrotoluene, *MDPI Catalysts*, 2023, 13(2), 387, https://doi.org/10.3390/catal13020387, IF 3.9
- 3. Alexandra Jakab-Nácsa, Attila Garami, Béla Fiser, László Farkas, Béla Viskolcz, Towards Machine Learning in Heterogeneous Catalysis—A Case Study of 2,4-Dinitrotoluene Hydrogenation, *MDPI International Journal of Molecular Sciences*, 2023, 24 (14), 11461, https://doi.org/10.3390/ijms2 41411461, IF 5.6

FURTHER PUBLICATIONS

 Viktória Hajdu, Alexandra Jakab-Nácsa, Gábor Muránszky, István Kocserha, Béla Fiser, Tibor Ferenci, Miklós Nagy, Béla Viskolcz, László Vanyorek, Precious-Metal-Decorated Chromium(IV) Oxide Nanowires as Efficient Catalysts for 2,4toluenediamine Synthesis, *International Journal of Molecular Science*, 2021, 22(11), 5945,

https://doi.org/10.3390/ijms22115945, IF 6.2

 Alexandra Jakab-Nácsa, Dávid Stomp, László Farkas, George Kaptay, Large NaCl-Effect on the Decomposition Rate of Chlorate Ions in HCl-Containing Brine Solutions and Its Consequences for Chlor-Alkali Industry, Periodica Polytechnica – Chemical Engineering, 2021, 65 (2),238-242, https://doi.org/10.3311/PPch.14634, IF 2022 1.3

PRESENTATION RELATED TO THE DISSERTATION

- Borsodi Vegyipari Napok *Rangsorolhatók-e a kémiai reakciók katalizátorai?* 17 November 2021, Miskolc, Hungary
- XXVIII. Nemzetközi Vegyészkonferencia Katalizátorok összehasonlíthatósága a MIRA21 modell alapján 27-29 October 2022, Nagyvárad, Romania
- **3.** International Conference on Chemical Engineering *Classification of Catalysts with MIRA21 Model in Heterogeneous Catalytic Hydrogenation of Aromatic Nitro Compounds* March 20-21, 2023, Rome, Italy

POSTERS RELATED TO THE DISSERTATION

- 26th International Congress of Chemical and Process Engineering Utilization of laboratory testing and a mathematical model for the purpose of selecting a heterogeneous hydrogenation catalyst
 21-25 August, 2022, Prague, Czech Republic
- 4. MKE Nemzeti Vegyészkonferencia Aromás nitrovegyületek katalitikus hidrogénezésére alkalmas katalizátorok fejlesztése az iparban 10-12 July, 2023, Eger, Hungary

FURTHER POSTERS

- 12th European Congress of Chemical Engineering Wastewater Treatment Optimization of Nitration of Aromatics 15-19 September, 2019, Florence, Italy
- 26th International Congress of Chemical and Process Engineering Cooperation of cities and local companies for climate change adaptation 21-25 August, 2022, Prague, Czech Republic
- 4. MKE Nemzeti Vegyészkonferencia Ipari szürkevíz visszaforgatási lehetőségek vizsgálata a LIFE projekt keretén belül

10-10 July, 2023, Eger, Hungary

7. **R**EFERENCES

- Medford AJ, Kunz MR, Ewing SM, Borders T, Fushimi R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* 2018;8(8):7403-7429. doi:https://doi.org/10.1021/acscatal.8b01708
- Ess D, Gagliardi L, Hammes-Schiffer S. Introduction: Computational Design of Catalysts from Molecules to Materials. *Chem Rev.* 2019;119(11):6507-6508. doi:https://doi.org/10.1021/acs.chemrev.9b00296
- 3. Morgenthaler S. Exploratory data analysis. *Wiley Interdiscip Rev Comput Stat.* 2009;1(1):33-44. doi:10.1002/WICS.2
- A Practical Introductory Guide to Exploratory Data Analysis | datos.gob.es. https://datos.gob.es/en/documentacion/practical-introductory-guideexploratory-data-analysis. Accessed April 17, 2023.
- Zavyalova U, Holena M, Schlögl R, Baerns M. Statistical analysis of past catalytic data on oxidative methane coupling for new insights into the composition of highperformance catalysts. *ChemCatChem*. 2011;3(12):1935-1947. doi:10.1002/cctc.201100186
- Schmack R, Friedrich A, Kondratenko E V., Polte J, Werwatz A, Kraehnert R. A metaanalysis of catalytic literature data reveals property-performance correlations for the OCM reaction. *Nat Commun.* 2019;10(1). doi:10.1038/s41467-019-08325-8
- Yang Q, Skrypnik A, Matvienko A, Lund H, Holena M, Kondratenko E V. Revealing property-performance relationships for efficient CO2 hydrogenation to higher hydrocarbons over Fe-based catalysts: Statistical analysis of literature data and its experimental validation. *Appl Catal B Environ*. 2021;282:119554. doi:10.1016/J.APCATB.2020.119554
- 8. Fedorov A, Linke D. Data analysis of CO2 hydrogenation catalysts for hydrocarbon production. *J CO2 Util*. 2022;61:102034. doi:10.1016/J.JCOU.2022.102034
- 9. Keith JA, Vassilev-Galindo V, Cheng B, et al. Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems. *Chem Rev.*

2021;121(16):9816-9872. doi:10.1021/acs.chemrev.1c00107

- Ras EJ, McKay B, Rothenberg G. Understanding catalytic biomass conversion through data mining. *Top Catal*. 2010;53(15-18):1202-1208. doi:10.1007/s11244-010-9563-z
- Rodríguez-Pérez R, Bajorath J. Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics. *Sci Rep.* 2021;11(1):1-9. doi:10.1038/s41598-021-93771y
- 12. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/S41586-020-2649-2
- 13. Waskom M. seaborn: statistical data visualization. *J Open Source Softw.* 2021;6(60):3021. doi:10.21105/JOSS.03021
- Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng.* 2007;9(3):90-95. doi:10.1109/MCSE.2007.55
- 15. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.; 2017.
- Jakab-Nácsa A, Hajdu V, Vanyorek L, Farkas L, Viskolcz B. Overview of Catalysts with MIRA21 Model in Heterogeneous Catalytic Hydrogenation of 2,4-Dinitrotoluene. *Catal 2023, Vol 13, Page 387.* 2023;13(2):387. doi:10.3390/CATAL13020387
- Jakab-Nácsa A, Garami A, Fiser B, Farkas L, Viskolcz B. Towards Machine Learning in Heterogeneous Catalysis-A Case Study of 2,4-Dinitrotoluene Hydrogenation. *Int J Mol Sci.* 2023;24(14). doi:10.3390/IJMS241411461
- Bystrzanowska M, Petkov P, Tobiszewski M. Ranking of Heterogeneous Catalysts Metals by Their Greenness. *ACS Sustain Chem Eng.* 2019;7(22):18434-18443. doi:10.1021/acssuschemeng.9b04230

8. Köszönetnyilvánítás/Acknowledgements

Nagy tisztelettel és hálával szeretném kifejezni köszönetemet Prof. Dr. Viskolcz Béla úrnak, a Kémia Intézet igazgatójának, aki számomra nem csupán egy egyszerű témavezető volt, hanem inkább egy iránytű a tudományos kutatások labirintusában. Az általa nyújtott szakmai segítség és iránymutatás felbecsülhetetlen értéket képvisel számomra. Mély hálával tartozom inspiráló beszédeiért, melyek a legnehezebb pillanatokban is erőt adtak, és segítettek kitartani a kutatás során. Őszinte köszönettel adózom elképesztő menedzseri készségeiért, melyek rendszerezetté és gördülékennyé tették a munkafolyamataimat.

Nem hagyhatom figyelmen kívül Farkas László, Manager Technology Support, ipari konzulensem kiemelkedő szerepét és befolyását az utam alakulásában. Ő volt az, aki rábírt arra, hogy ne csak a kényelmes mederben haladjak, hanem vállaljam fel a doktori képzés kihívásait. Ahogy a doktori képzés kanyargó útjain haladtam, mindvégig mellettem állt, értékes gondolatokkal és mélyreható tanácsokkal látott el. Mindig rendelkezésemre állt, türelmesen válaszolt kérdéseimre, meghallgatta aggodalmaimat, és bátorított arra, hogy ne féljek a változástól vagy a nehézségektől. Hálával és köszönettel tartozom neki mindazért, amit értem tett.

Külön köszönettel tartozom Garami Attilának, az Energia-, Kerámia- és Polimertechnológiai Intézet adjunktusának, aki az adatelemzés és a mesterséges intelligencia módszereinek alkalmazásával támogatta kutatásomat. Világossá tette továbbá számomra az összetettebb problémákat és megnyitotta az ajtót azok megértéséhez. Ő volt az, akinek a segítségével az elméleti alapoktól a gyakorlati alkalmazásig sokat tanulhattam.

Hálával tartozom a Kémiai Intézet katalitikus hidrogénezéssel foglalkozó csoportjának, akik hozzájárultak a doktori munkám tárgyát képező MIRA21 modell létrehozásához. Köszönöm, Vanyorek László, Sikora Emőke, Prekob Ádám és Hajdú Viktória segítségét és együttműködését. Köszönöm Dr. Fiser Bélának és Dr. Szőri Milán egyetemi munkatársaknak, akik támogattak és értékes észrevételeket tettek.

Köszönettel tartozom Dr. Nehéz Károlynak, az Informatikai Intézet igazgatójának, valamint Szabó Martin tanársegédnek a MIRA adatbázis webalapú fejlesztésében való

közreműködésért. Hálásan köszönöm továbbá Dr. Bánhidi Olivér, címzetes egyetemi tanár támogatását és online tanóráit, amelyek bevezettek az adatelemzés világába.

Az oktatók és nem oktató munkatársak a Kémiai Intézetben kulcsfontosságúak voltak az utam során. Mindig segítőkészen álltak rendelkezésemre, legyen szó akár technikai kérdésekről, akár szakmai támogatásról. Hálával tartozom értük!

Nem tudom eléggé kifejezni, milyen hálás vagyok a férjemnek. Mindvégig mellettem állt ezen az úton, türelmesen meghallgatta az aggodalmaimat és örömömet, és mindig biztatott a kitartásra. Ő mindig ott volt az oldalamon, hogy erőt adjon, amikor szükségem volt rá, és örömömre osztozzon az eredményeimben. A férjem szeretete és támogatása a legnagyobb kincs az életemben.

Az én két kisgyermekemnek is hálával tartozom, amiért megértéssel fogadták, amikor nem tudtam velük lenni, és támogattak engem azzal, hogy maguk is erőt adtak. Az ő ártatlan örömeik és szeretetük folyamatos inspirációt nyújtanak, és mindig emlékeztetnek arra, hogy miért is vállaltam mindezt. A családomnak mély hálával tartozom mindazért, amit értem tettek. Az ő segítségük nélkül jóval bonyolultabb lett volna az utam.

Végül, de nem utolsósorban, hálával tartozom a Technológiai Támogatás összes munkatársának, akik rugalmas együttműködésükkel, készséges segítségükkel és hozzáértésükkel hozzájárultak a sikeres eredmények eléréséhez.

Hálásan köszönöm Purzsa Tamás Vice President Úrnak, hogy lehetővé tette számomra, hogy elkezdhessem Ph.D. képzésemet munkám mellett.

I am deeply grateful to Mr. Zhao Nan for enabling me to successfully complete my doctoral studies.

I appreciate the opportunity provided by Wanhua-BorsodChem to conduct my Ph.D. Study.

Prepared with the professional support of Doctoral Student Scholarship Program of the Co-operative Doctoral Program of the Ministry of Innovation and Technology financed from the National Research, Development, and Innovation Fund. The project with identification number 2020-1.1.2-PIACI-KFI-2020-00121 is implemented with the support of the Ministry of Innovation and Technology from the National Research, Development, and Innovation Fund.