

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Cyber-physical oriented chemical process for green and sustainable production

Vinicius Viena Santana

DISSERTATION PRESENTED TO THE FACULTY OF ENGINEERING OF
THE UNIVERSITY OF PORTO FOR THE DEGREE OF DOCTOR IN
CHEMICAL AND BIOLOGICAL ENGINEERING



Doctoral Program in Chemical and Biological Engineering

Supervisor: Idelfonso Bessa dos Reis Nogueira (NTNU, Norway)

Co-Supervisor: Ana Mafalda Ribeiro (LSRE - LCM, ALICE, FEUP, UPorto,
Portugal)

Co-Supervisor: Christopher Rackauckas (Julia Lab, MIT, USA)

April 11, 2024

Cyber-physical oriented chemical process for green and sustainable production

Vinicius Viena Santana

Doctoral Program in Chemical and Biological Engineering

April 11, 2024

Resumo

Esta tese examina a combinação de modelos mecanicistas e orientados por dados em engenharia química, focando na aplicação dentro da engenharia de sistemas de processos. Um dos objetivos é melhorar as ferramentas e métodos para analisar sistemas descritos por equações diferenciais. Primeiro, validando novas técnicas de modelagem onde os dados são escassos e o conhecimento do domínio é de moderado a alto; e depois, explorando como dados abundantes podem aprimorar e refinar modelos mecanicistas. Neste sentido, esta tese está dividida em duas partes distintas.

A Parte I aborda o desafio de modelar processos de separação baseados em adsorção com dados limitados, mas com um conhecimento substancial do processo. O Capítulo 1 aplica Redes Neurais Informadas pela Física (PINNs) para simular a cromatografia em coluna para a purificação de anticorpos monoclonais, alcançando resíduos próximos do estado da arte e demonstrando robustez contra rigidez com estratégias avançadas de treino. O Capítulo 2 introduz uma abordagem de modelagem híbrida para a recuperação de lítio de baterias, utilizando redes neurais para capturar a resistência à transferência de massa, mostrando desempenho superior sobre modelos tradicionais em validações experimentais. O Capítulo 3 supera as limitações da modelagem híbrida em sistemas simples, estendendo-a para equações diferenciais parciais, melhorando a previsão da cromatografia em coluna de fase líquida substituindo a cinética de adsorção por uma Rede Neural Artificial (ANN). Esta abordagem produz previsões mais precisas do que modelos completamente mecanicistas. O Capítulo 4 aprofunda-se na numérica de modelos híbridos para equações diferenciais parciais, estabelecendo diretrizes para a Diferenciação Automática (AD) em modelos rígidos e demonstrando a capacidade da ANN de extrapolar o comportamento de adsorção a partir de dados de curva de ruptura com ruído gaussiano, com técnicas de regressão simbólica revelando modelos polinomiais que se assemelham de perto aos modelos cinéticos verdadeiros.

A Parte II muda o foco para cenários abundantes em dados, mas com escassez de conhecimento do domínio, sublinhando a identificação de modelos substitutos e a integração de aprendizado de máquina com modelos mecanicistas. Esta seção utiliza dados de várias fontes, incluindo modelos de ordem reduzida e previsões de aprendizado de máquina, para enfrentar desafios na engenharia de perfumes e na estimativa da pressão de vapor. O Capítulo 5 foca na formulação ótima de perfumes. Introduce um método para formular perfumes de forma eficiente usando modelos de aprendizado profundo treinados com dados gerados por um modelo fenomenológico, reduzindo significativamente o tempo computacional para a previsão do comportamento dinâmico – os modelos de aprendizado profundo desenvolvidos exigem um tempo de CPU consideravelmente baixo para prever o comportamento dinâmico completo da liberação do perfume – 0,16 milissegundos contra 6 segundos para o modelo de primeiro princípio. O Capítulo 7 apresenta um modelo de aprendizado de máquina para integrar a aprendizagem por transferência com um nó de

viés indutivo especializado baseado na equação de Antoine para aprimorar a previsão da pressão de vapor - uma propriedade muito desejada e importante não só na engenharia de perfumes, mas em vários outros campos. A abordagem mostrou superar as alternativas ao explorar o viés indutivo e usar a aprendizagem por transferência através de incorporações gráficas. Para concluir a tese, o Capítulo 6 introduz um método inovador para classificar e prever a intensidade do odor em perfumaria. Ele combina aprendizado de máquina científico com modelagem mecanística para aprimorar o design de fragrâncias. O método utiliza atribuição de pesos probabilísticos baseada em resultados de classificadores de aroma, reconhecendo a natureza subjetiva da classificação de aromas. Também incorpora um quadro de análise de incerteza para melhorar a adaptabilidade e a confiabilidade do modelo. A metodologia inclui um modelo de difusão de perfume para a evaporação e dispersão de moléculas de fragrância, um modelo de percepção de odor usando Valor de Odor para a quantificação da intensidade e uma quantificação de incerteza para analisar parâmetros e previsões do modelo. Os resultados demonstram uma abordagem viável para o desenho robusto de perfumes, onde elementos probabilísticos são combinados para prever as probabilidades de pertencimento a famílias olfativas em misturas fragrantes multicomponentes.

Abstract

This thesis examines the combination of mechanistic and data-driven models in chemical engineering, focusing on the application within process systems engineering. One goal is to improve the tools and methods for analyzing systems described by differential equations. First, by validating new modeling techniques where data is scarce, and domain knowledge is from moderate to high; second, by exploring how abundant data can enhance and fine-tune mechanistic models. In this sense, this thesis is divided into two distinct parts.

Part I addresses the challenge of modeling adsorption-based separation processes with limited data but substantial process knowledge. Chapter 1 applies Physics-Informed Neural Networks (PINNs) to simulate nonlinear column chromatography for purifying monoclonal antibodies, achieving near state-of-the-art residuals and demonstrating robustness to stiffness. Chapter 2 introduces a hybrid modeling approach for lithium recovery from batteries, employing neural networks to capture mass transfer resistance, showing superior performance over traditional models in experimental validations. Chapter 3 overcomes the constraints of hybrid modeling in simple systems by extending it to partial differential equations, enhancing the prediction of nonlinear liquid-phase column chromatography by substituting adsorption kinetics with an Artificial Neural Network (ANN). This approach yields more accurate predictions than fully mechanistic models. Chapter 4 delves into the numerics of hybrid models for PDEs, establishing guidelines for Automatic Differentiation (AD) in stiff models and demonstrating the ANN’s ability to extrapolate adsorption behavior from noisy breakthrough data, with symbolic regression techniques revealing polynomial models that closely resemble the true kinetic models.

Part II shifts the focus to scenarios abundant in data but sparse in domain knowledge, underscoring the identification of surrogate models and the integration of machine learning with mechanistic models. This section draws upon data from various sources, including reduced-order models and machine-learning predictions, to address challenges in perfume engineering and vapor pressure estimation. Chapter 5 focuses on the optimal formulation of perfumes. It introduces a method for efficiently formulating perfumes using deep learning models trained on data generated by a phenomenological model, significantly reducing computational time for dynamic behavior prediction – the developed DNN models demand considerably low CPU time to predict the full dynamic behavior of perfume release – 0.16 ms against 6s for the first principle model. Chapter 6 presents a machine learning model for integrating transfer learning with a specialized inductive bias node based on the Antoine equation for enhanced vapor pressure prediction - a very desired and important property in perfume engineering and several other fields. The approach surpassed alternatives by exploiting inductive bias and using transfer learning via graph embeddings. To close the thesis, Chapter 7 introduces an innovative method for classifying and predicting odor intensity in perfumery. It merges scientific machine learning with mechanistic modeling to improve fragrance design. The method uses probabilistic weight assignment based on scent

classifier outputs, acknowledging the subjective nature of scent classification. It also incorporates an uncertainty analysis framework to enhance model adaptability and reliability. The methodology includes a perfume diffusion model for fragrance molecule evaporation and dispersion, an odor perception model using Odor Value for intensity quantification, and an uncertainty quantification for analyzing model parameters and predictions. The results demonstrate a feasible approach towards robust perfume design where probabilistic elements are composed together to predict membership probabilities of olfactory families in multicomponent fragrant mixtures.

Acknowledgments

First and foremost, I would like to extend my deepest gratitude to my family, who have been unwavering in their support throughout this journey. To my mother, Rita, and my father, Luis, your love and encouragement have been the foundation upon which I've built my academic and personal life. To my brother, Victor, thank you for always being there when I needed a listening ear or a word of advice.

To my loving wife, Beatriz, I am immensely thankful for the incredible moments we've shared and your patience, especially during those long nights when work consumed me. Your support has been nothing short of extraordinary.

I am particularly indebted to my main supervisor, Idelfonso, whose friendship and guidance have been pivotal for this thesis and my academic journey. Your wisdom, friendship, and mentorship have been invaluable to me. I will always be grateful for what you have done for me.

A special thank you to my co-supervisor, Prof. Ana Mafalda, for her kind words, continuous support, and unwavering motivation. Your contributions to my academic life have been substantial.

To Chris Rackauckas, my other co-supervisor, I owe a debt of gratitude for guiding me through the learning processes in programming in Julia and SciML concepts. Your support and expertise have been indispensable.

Lastly, but by no means least, I want to thank my friends Erbet and Carine for their unconditional friendship and teamwork. Your support has made this journey far less daunting.

I also want to extend my heartfelt thanks to the folks in room 304—Mariana, Rita, Cristiana, and Paulo—for the unforgettable conversations and moments at FEUP. You have enriched my life in ways that words can't adequately express.

Last but not least. I want to thank FCT and the MIT Portugal program for the financial and administrative support that led to the conclusion of this thesis.

This work was financially supported by the doctoral Grant (reference PRT/BD/152850/2021) financed by the Portuguese Foundation for Science and Technology (FCT) and with funds from State Budget, under MIT Portugal Program; UIDB/50020/2020 and UIDP/50020/2020 (LSRE-LCM), and LA/P/0045/2020 (ALiCE), funded by national funds through FCT/MCTES (PIDDAC).



Vinicius Viena Santana

*“The important thing is not to stop questioning
Curiosity has its own reason for existing.”*

Albert Einstein

List of Publications

This thesis has culminated in six peer-reviewed publications in high-impact journals and one submitted manuscript under review. Of these, I served as the first author on five and the corresponding author on the remaining two. In both manuscripts where I was the corresponding author, my primary role was to apply machine learning techniques to the given case study. The first authors took charge of mechanistic modeling and data provision. It is important to highlight that the first and corresponding authors equally contributed to the manuscripts in implementation and conceptualization.

- **Chapter 1** was published as the cover article for its respective journal issue, an honor indicative of its contribution to the field.
- **Chapter 4** features a hybrid model that also received recognition through its acceptance as a formal model submission at the prestigious International Conference on Machine Learning (ICML), 2023 [URL]

Furthermore, select findings from this work were presented in poster format at the 2022 MIT Portugal Conference and received the *Best Work in Data Science* accolade. [URL]

Moreover, this thesis catalyzed substantial international collaborations across several continents and research institutions, effectively bridging research ecosystems between Europe and the Americas. Notably involved institutions encompass the University of Porto, the Federal University of Rio de Janeiro, the Norwegian University of Science and Technology, and the Massachusetts Institute of Technology.

Publications in Journals

1. Chapter 1: Santana, V. V., Gama, M. S., Loureiro, J. M., Rodrigues, A. E., Ribeiro, A. M., Tavares, F. W., Barreto, A. G., and Nogueira, I. B. R. A first approach towards adsorption-oriented physics-informed neural networks: Monoclonal antibody adsorption performance on an ion-exchange column as a case study. *ChemEngineering*, 6, 2022. ISSN 2305-7084. doi: 10.3390/chemengineering6020021. URL <https://www.mdpi.com/2305-7084/6/2/21>
2. Chapter 2: Regufe, M. J., Santana, V. V., Ferreira, A. F. P., Ribeiro, A. M., Loureiro, J. M., and Nogueira, I. B. R. A hybrid modeling framework for membrane separation processes: Application to lithium-ion recovery from batteries. *Processes*, 9(11), 2021b. ISSN 2227-9717. doi: 10.3390/pr9111939. URL <https://www.mdpi.com/2227-9717/9/11/1939>
3. Chapter 3: Nogueira, I. B. R., Santana, V. V., Ribeiro, A. M., and Rodrigues, A. E. Using scientific machine learning to develop universal differential equation for multicomponent adsorption separation systems. *The Canadian Journal of Chemical*

- Engineering*, 7 2022. ISSN 1939-019X. doi: 10.1002/CJCE.24495. URL <https://onlinelibrary.wiley.com/doi/full/10.1002/cjce.24495>
4. Chapter 4: Santana, V. V., Costa, E., Rebello, C. M., Ribeiro, A. M., Rackauckas, C., and Nogueira, I. B. Efficient hybrid modeling and sorption model discovery for non-linear advection-diffusion-sorption systems: A systematic scientific machine learning approach. *Chemical Engineering Science*, 282:119223, 2023. ISSN 0009-2509. doi: <https://doi.org/10.1016/j.ces.2023.119223>. URL <https://www.sciencedirect.com/science/article/pii/S0009250923007790>
 5. Chapter 5: Santana, V. V., Martins, M. A., Loureiro, J. M., Ribeiro, A. M., Rodrigues, A. E., and Nogueira, I. B. Optimal fragrances formulation using a deep learning neural network architecture: A novel systematic approach. *Computers and Chemical Engineering*, 150:107344, 4 2021. ISSN 00981354. doi: 10.1016/j.compchemeng.2021.107344
 6. Chapter 6: Santana, V. V., Rebello, C. M., Queiroz, L. P., Ribeiro, A. M., Shardt, N., and Nogueira, I. B. Puffin: A path-unifying feed-forward interfaced network for vapor pressure prediction. *Chemical Engineering Science*, 286:119623, 2024. ISSN 0009-2509. doi: <https://doi.org/10.1016/j.ces.2023.119623>. URL <https://www.sciencedirect.com/science/article/pii/S000925092301179X>
 7. Chapter 7: Vinícius V. Santana, Ana M. Ribeiro, Carine M. Rebello, Erbet Costa, Chris Rackauckas and Idelfonso B.R. Nogueira. Advancing Odor Classification Models Enhanced by Scientific Machine Learning and Mechanistic Model: Probabilistic Weight Assignment for Odor Intensity Prediction and Uncertainty Analysis for Robust Fragrance Classification. *Manuscript under review in Industrial and Engineering Chemistry Research*.

Contents

Resumo	i
Abstract	iii
Introduction	1
Motivation and Relevance	1
Outline and Objectives	2
I Physics-Informed Machine Learning in Chemical Engineering	6
1 A First Approach Towards Adsorption-Oriented Physics-Informed Neural Networks	7
1.1 Abstract	7
1.2 Introduction	7
1.3 Methodology	9
1.3.1 Simulation Scenario	10
1.3.2 Physics-Informed Neural Networks	11
1.4 Results	13
1.5 Conclusions	15
2 A hybrid modeling framework for membrane separation processes: Application to the Lithium ions recovery from batteries	17
2.1 Abstract	17
2.2 Introduction	18
2.2.1 Case Study: Lithium ions recovery from batteries	19
2.3 Materials and Methods	20
2.3.1 Hybrid model	20
2.3.2 Data Collection	22
2.3.3 Hybrid Modeling	23
2.3.4 Parameter Estimation	24
2.4 Results	25
2.5 Conclusions	32
3 Using Scientific Machine Learning to Develop Universal Differential Equation for Multicomponent Adsorption Separation Systems	33
3.1 Abstract	33
3.2 Introduction	33
3.3 Methodology	35

3.3.1	Fully mechanistic model	35
3.3.2	Universal Differential Equation in column chromatography context .	37
3.3.3	Simulation	38
3.4	Results	39
3.5	Conclusions	47
4	Efficient hybrid modeling and sorption model discovery for non-linear advection-diffusion-sorption systems: A systematic scientific machine learning approach	48
4.1	Abstract	48
4.2	Introduction	49
4.3	Methodology	52
4.3.1	Data set	52
4.3.2	Hybrid model	53
4.3.3	Model’s numerical solution and gradient calculation	55
4.3.4	Sparse Regression	57
4.4	Results and Discussion	61
4.4.1	UDEs’ training and test sets performance	61
4.4.2	Sparse Regression	65
4.4.3	Symbolic Regression	68
4.5	Conclusions	71
II	Model Acceleration and Scale Integration with Data-Driven Modeling in Chemical Engineering	72
5	Optimal fragrances formulation using a deep learning neural network architecture: a novel systematic approach.	73
5.1	Abstract	73
5.2	Introduction	73
5.3	Methodology	76
5.3.1	The perfume diffusion process and model	76
5.3.2	Odor intensity model	78
5.3.3	System Identification	79
5.3.4	Perfume performance optimization	85
5.4	Results and Discussion	86
5.4.1	Data Acquisition	86
5.4.2	Training and evaluation	90
5.4.3	Perfume performance optimization	93
5.5	Conclusions	95
6	PUFFIN: A Path-Unifying Feed-Forward Interfaced Network for Vapor Pressure Prediction	97
6.1	Abstract	97
6.2	Introduction	97
6.3	Methodology	100
6.3.1	Graph Neural Networks and Transfer Learning	101
6.3.2	Generic Descriptor: Morgan Fingerprints	103
6.3.3	PUFFIN: Path-Unifying Feed-Forward Interfaced Network	104

6.3.4	PUFFIN: Assessment of transfer learning efficacy	105
6.3.5	Vapor Pressure and Boiling Point Data Sets	106
6.3.6	Hyperparameter Identification	107
6.4	Results	108
6.4.1	Molecular Embedding—Estimation From Boiling Point	108
6.4.2	Molecular Embedding Without Domain Knowledge Node—Vapor Pressure Prediction	109
6.4.3	Molecular Embedding With Domain Knowledge Node (PUFFIN)— Vapor Pressure Prediction	110
6.4.4	Morgan Fingerprints—Vapor Pressure Prediction	114
6.4.5	Interpretability remarks	115
6.5	Conclusions	118
7	Advancing Odor Classification Models Enhanced by Scientific Machine Learning and Mechanistic Model: Probabilistic Weight Assignment for Odor Intensity Prediction and Uncertainty Analysis for Robust Fra- grance Classification	119
7.1	Abstract	119
7.2	Introduction	120
7.3	Methodology	123
7.3.1	Perfume diffusion model	123
7.3.2	Odor intensity perception model	126
7.3.3	Odor classification	127
7.3.4	Uncertainty quantification and propagation	127
7.3.5	Integration of models for robust fragrance odor classification	131
7.4	Results	133
7.4.1	Quantifying uncertainties of diffusion coefficient and vapor pressure .	133
7.4.2	Quantifying uncertainties of odor threshold	135
7.4.3	Uncertainty propagation	136
7.4.4	Integration of models for robust fragrance classification	137
7.5	Conclusions	142
	Conclusions and Future Directions	143
	References	146
A	Chapter 4	160
A.1	Fully mechanistic model	160
A.2	UDE Training and test set performance - Sips isotherm and Langmuir isotherm	162
A.3	Sparse regression - Sips isotherm	168
B	Chapter 7	170
B.1	Average concentration and quantile ribbons	170

List of Figures

1	Modeling landscape in chemical engineering, depicting data availability, domain knowledge, and corresponding modeling strategies.	3
1.1	Schematic representation of the antibody adsorption by ion-exchange	11
1.2	Physics-informed Neural Network building process.	12
1.3	Train loss history for numerical adsorption experiment.	14
1.4	MoL and PINNs solution’s comparison for adsorption experiment.	15
2.1	Schematic representation of an FSSLM.	20
2.2	Schematic representation of the membrane transport profile: solute (S), metal-carrier complex (C), and carrier (E)	21
2.3	Schematic representation of the proposed framework.	24
2.4	Lithium experimental set 1 - Hybrid model prediction (H – Model) versus the phenomenological model prediction (R – Model) and their respective parity graphics	26
2.5	Lithium experimental set 2 - Hybrid model prediction (H – Model) versus the phenomenological model prediction (R – Model) and their respective parity graphics.	27
2.6	Experimental sets 3 and 4 - Hybrid model prediction vs. the phenomenological model prediction and their respective parity graphics.	29
2.7	Experimental sets 4 and 5 - Hybrid model prediction vs. the phenomenological model prediction and their respective parity graphics	30
2.8	Experimental sets 6 and 7 - Hybrid model prediction vs. the phenomenological model prediction and their respective parity graphics.	31
3.1	Illustration of UDE building and fitting procedure	38
3.2	Training loss history.	41
3.3	Train and test breakthrough curves for POH – UDE vs Phenomenological comparison.	42
3.4	Train and test breakthrough curves for ProPro – UDE vs Phenomenological comparison.	43
3.5	Train and test breakthrough curves’ error for POH – UDE vs Phenomenological comparison.	43
3.6	Train and test breakthrough curves for ProPro – UDE vs Phenomenological comparison.	44
3.7	Parity plots for breakthrough curves’ train and test error– UDE vs Phenomenological comparison (a) Propanol (POH). (b) n-Propyl Propionate (ProPro).	45

3.8	Comparison of UDE and fully-mechanistic simulation profile for the second breakthrough curve. (a) Propanol (POH). (b) n-Propyl Propionate (ProPro).	46
3.9	ANNEQ competitive adsorption isotherm for the binary 1-Propanol and n-Propyl Propionate and experimental data over Amberlyst-46 at 313 K. . . .	46
4.1	Hybrid models types - adapted from von Stosch et al. (2014)	50
4.2	Schematic representation of the proposed hybrid model.	54
4.3	The decision-making process for identifying hybrid models.	57
4.4	Comparison between direct sparse/symbolic regression and UDE approach for mass transfer kinetic identification on PDEs	60
4.5	Training and test performance	62
4.6	Training and test performance	63
4.7	Training and test performance	64
4.8	Sparse regression results for the best number of active terms (3) - Langmuir isotherm + LDF kinetics at three positions in the bed.	66
4.9	Results for Langmuir isotherm and LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial. .	66
4.10	Sparse regression results for the best number of active terms (4) - Langmuir isotherm + improved LDF kinetics at three positions in the bed.	67
4.11	Results for Langmuir isotherm and improved LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.	67
4.12	Sparse regression results for the best number of active terms (6) - Langmuir isotherm + Vermeulen's kinetics at three positions in the bed.	67
4.13	Results for Langmuir isotherm and Vermeulen's kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.	67
4.14	Taylor expansion of uptake rate for Vermeulen's model and Langmuir isotherm - Mechanistic model	69
4.15	Taylor expansion of uptake rate for improved LDF model and Langmuir isotherm - Mechanistic model	70
5.1	Illustration of the device for perfume release and propagation study.	76
5.2	Diagram of system identification steps	80
5.3	Illustration of NOE predictor	82
5.4	Overall predictor structure of perfume release and propagation.	84
5.5	Distributions of mole fraction inputs.	88
5.6	Time evolution of odorant's intensities at five distances from the liquid perfume source for experiment number 56.	89
5.7	DNN models validation.	92
5.8	DNN models validation for tonalide.	92
5.9	Odor intensity radar plot for optimized perfume performance at different time instants.	94
5.10	Odor intensity plot for (Teixeira et al., 2013a)'s work perfume formulation .	95
6.1	Schematic diagram of the methodology presented in this work	102
6.2	Parity plot of graph neural network (GNN) prediction of boiling point compared to the ground truth in the test set	109

6.3	Comparative analysis: parity plots and kernel density estimation of residuals, including mean and variance for each model in the test dataset.	112
6.4	(a) Parity plot for GCN with Antoine reinforcement on vapor pressure prediction. (b) Residuals' Kernel Density Estimation for LogVP predictions by the same GCN model	113
6.5	Extrapolation assessment of all models for three test set components. (a) Acetone, (b) Octyl-Benzene and (c) Butanol.	117
7.1	Schematic representation of the steps involved in the odor perception model for fragrance mixtures. Adapted from Teixeira (2011).	121
7.2	Scheme of the simulated liquid and gas phases of the perfume system. Adapted from Teixeira et al. (2009a)	124
7.3	Diagram for the methodology of robust fragrance classification (olfactory families are generic).	132
7.4	Histogram of the error of diffusion coefficients in air.	133
7.5	Histogram of the error of ethanol's saturation temperature	134
7.6	Evolution of the expected value of geraniol concentration close to the perfume source and 14900s after the release	136
7.7	Mean gas phase concentrations and 99% credible intervals for all components at $z = 100\text{cm}$	137
7.8	Estimated probabilities of membership for each ingredient in the fragrant mixture	138
7.9	Probability density estimation of 6 olfactory families intensities at $z = 0.05\text{ m}$, $t = 4\text{ s}$	139
7.10	Average intensity distributions for each odor family, presented in ascending order of intensity - $z = 0.05\text{ m}$, $t = 4\text{ s}$	139
7.11	Radar chart depicting the top olfactory families based on their relative frequency of exhibiting the highest intensity at $z = 0.05\text{ m}$, $t = 4\text{ s}$	140
7.12	Average intensity distributions for each odor family, presented in ascending order of intensity - $z = 0.2\text{ m}$, $t = 600\text{ s}$	141
7.13	Radar chart depicting the top olfactory families based on their relative frequency of exhibiting the highest intensity at $z = 0.2\text{ m}$, $t = 600\text{ s}$	141
A.1	Training and test performance	162
A.2	Training and test performance	163
A.3	Training and test performance	164
A.4	Training and test performance	165
A.5	Training and test performance	166
A.6	Training and test performance	167
A.7	Sparse regression results for the best number of active terms (3) - Sips isotherm + LDF kinetics at three positions in the bed.	168
A.8	Results for Sips isotherm and LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.	168
A.9	Sparse regression results for the best number of active terms (4) - Sips isotherm + improved LDF kinetics at three positions in the bed.	168
A.10	Results for Sips isotherm and improved LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.	169

A.11	Sparse regression results for the best number of active terms (6) - Sips isotherm + Vermeulen's kinetics at three positions in the bed.	169
A.12	Results for Sips isotherm and Vermeulen's kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.	169
B.1	Mean gas phase concentrations and 99% credible intervals for all components at $z = 0.05$ m	170
B.2	Mean gas phase concentrations and 99% credible intervals for all components at $z = 50$ cm	171

List of Tables

1.1	Parameters to define the neural network for the adsorption experiment. . . .	13
2.1	Details for the experimental data collected in the literature and used to train and validate the hybrid model.	23
3.1	Initial concentrations, c_0 , and feed concentration, c_{in} , for each step performed and solid concentration, q^* , in equilibrium with the feed concentration in the breakthrough experiments.	39
3.2	Fixed bed parameters and feed flow rate.	40
4.1	Test <i>in-silico</i> data set	53
4.2	ANNs' hyperparameters	61
4.3	Sparse regression learned polynomials.	66
4.4	Symbolic regression learned polynomials.	68
5.1	Mole fraction ranges.	87
5.2	Hyperparameter search space.	90
5.3	Results of hyperparameter tuning with random grid search.	91
5.4	Weights for the proposed objective function to optimize perfume performance.	93
6.1	Architecture summary of identified graph neural network (GNN) for predicting boiling point.	109
6.2	Architecture summary of the identified feed forward neural network for predicting vapor pressure from embeddings.	110
6.3	Architecture summary of the identified feed-forward neural network for predicting vapor pressure from embeddings with domain knowledge reinforcement (the Antoine node).	111
6.4	Summary of test set performance for the three approaches evaluated based on Mean Squared Error (MSE) metric.	114
6.5	Architecture summary of identified feed-forward neural network for predicting vapor pressure from Morgan fingerprints.	115
6.6	Comparison of coefficients predicted from PUFFIN and from literature. . . .	115
7.1	Antoine parameters and standard deviations for selected substances. . . .	135

Acronyms

AI Artificial Intelligence.

ANN Artificial Neural Network.

BFGS Broyden–Fletcher–Goldfarb–Shanno.

BIC Bayesian Information Criterion.

DL Deep Learning.

DNN Deep Neural Network.

FF Flavor and Fragrance.

FSSLM flat sheet-supported liquid membranes.

HFSLM hollow fibers-supported liquid membranes.

JIT Just-In-Time.

LDF Linear Driving Force.

LHS Latin Hypercube Sampling.

LIBs lithium-ion batteries.

LSTM Long Short-Term memory.

mAb Monoclonal Antibody.

NARX Non-Linear AutoRegressive Model with Exogeneous Inputs.

NOE Non-Linear Output Error.

ODE Ordinary Differential Equation.

PDE Partial Differential Equation.

PINN Physics-Informed Neural Network.

POH Propanol.

ProPro n-Propyl Propionate.

PSO Particle Swarm Optimization.

QSOR Quantitative Structure-Odor Relationship.

SciML Scientific Machine Learning.

SINDy Sparse Identification of Nonlinear Dynamics.

SLMs supported liquid membranes.

UDE Universal Differential Equation.

Introduction

Motivation and Relevance

The traditional approach to scientific exploration follows a step-by-step process, starting from gathering data, then analysis, and finally constructing theories that encapsulate new-found understanding. However, in recent times, the integration of advanced computational tools for evaluating intricate models and data-driven strategies for optimizing and refining models has significantly boosted and expedited the scientific discovery process. This integration has given rise to what is referred to as the "third" and "fourth" paradigms of scientific exploration and engineering innovation.

Historically, the majority of scientific research effort has been on data collection. However, this existing paradigm is rapidly being stretched to its limits because of dramatic growth in scientific data. As the size and complexity of data increase, the likelihood of missing opportunities for breakthroughs also will escalate and may ultimately delay progress. To address this challenge, many areas of science are moving toward more data-driven techniques that ultimately aim to augment the need for prior assumptions and meticulously tailored hypotheses with massive data collections. There are many potential benefits from engaging in machine learning in computational science, including in chemical engineering science. Machine learning methods are particularly useful for discovering correlations in high-dimensional data and, thus, can be useful in analyzing computational results.

In chemical engineering, the application of scientific machine learning is relevant due to its potential to optimize complex processes and improve decision-making in the industry. The ability of machine learning to handle large amounts of data and make informed decisions, in combination with real-time capabilities, provides chemical engineers with a powerful tool to improve efficiency and reduce operating costs. Additionally, the ability to simulate and predict the behavior of chemical processes reliably in real-time can lead to the development of more sustainable and environmentally friendly processes and pave the way toward Industry 4.0. Overall, integrating scientific machine learning into chemical engineering science is crucial for the continued advancement and optimization of the industry.

Still, despite the clear opportunities, numerous challenges remain in adopting machine learning in scientific applications, especially in chemical engineering science, namely how to

efficiently integrate machine learning in complex chemical engineering models (distributed models in separation and reaction engineering) conditioned on the available information and knowledge? How can we make this integration robust, i.e., excel in extrapolation and have a clear interpretation? This work addresses these questions.

To address these questions systematically, this thesis is structured into two distinct parts, each tailored to specific scenarios determined by the interplay between the quantity of available data and the depth of understanding regarding the process under consideration.

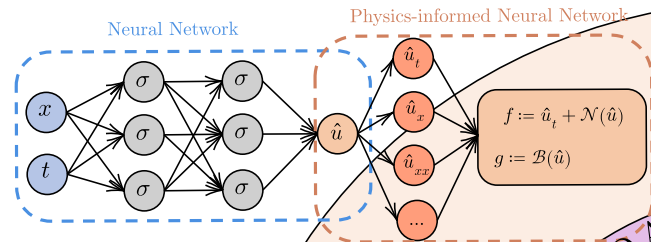
The division of this thesis into two parts serves as a deliberate framework, accommodating the diverse landscape of data abundance and domain knowledge in chemical engineering. In the first part, scenarios where low data abundance is accompanied by a solid foundation of process knowledge are explored. This work explores ways to construct precise and interpretable models that integrate data-driven and mechanistic models for enhancing predictive accuracy when conventional approaches might fall short. Conversely, the second part is dedicated to situations where data is abundant, but process knowledge is minimal or unnecessary, necessitating more highly parametrized models to extract useful relationships.

Adopting this dual-part structure ensures a comprehensive exploration of data-driven and physics-informed modeling techniques across data availability and process understanding. This approach provides valuable insights and practical solutions for various chemical engineering science applications, where the balance between data quantity and prior knowledge often dictates the most effective modeling strategy.

Outline and Objectives

As previously outlined, this study is segmented into two principal parts, each focused on specialized computational frameworks within chemical engineering. This division is schematically depicted in Figure 1. In the diagram, the data availability circle and the physics knowledge circle have been designed to represent varying levels of each. Chapters are accordingly situated within these fluctuating scopes; for example, Physics-Informed Neural Networks are categorized under conditions of sparse data but rich physics-based understanding. The point at which the circles intersect serves as a unique scenario, capturing the intersection of physics-based and physics-free methods and data-rich and data-scarce situations.

Chapter 1



Chapters 2,3 & 4

Universal Differential Equations

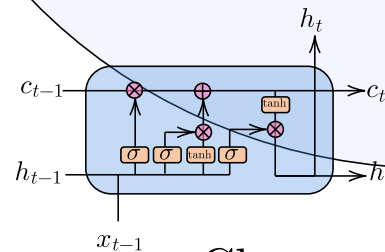
$$\begin{aligned} \frac{dx}{dt} &= \alpha x + \mathcal{N}\mathcal{N}_1(\theta, x, y) \\ \frac{dy}{dt} &= \lambda y + \mathcal{N}\mathcal{N}_2(\theta, x, y) \end{aligned}$$

Known Physics (orange) and Unknown Physics (blue) are indicated above the equations.

Chapter 7

Small Data + Big Data
Physics-Free + Physics-Based
(Scale Integration)

→ Deep Recurrent Neural Networks



Chapter 5 & 6

The initial part of this work (Part I) is dedicated to exploring concise integrations of mechanistic and data-driven models in scenarios marked by limited data yet bolstered by a strong foundation of process knowledge. Most case studies involve separation engineering, focusing on systems described by time-dependent differential equations (ordinary and partial). This section is structured into four chapters, each focusing on specific aspects.

Chapter 1 validates the effectiveness of the Physics-Informed Neural Network (PINN) technique in addressing the numerical challenges stemming from nonlinear liquid-phase column chromatography models. It also lays the groundwork for employing PINNs to tackle adsorption-based models.

Chapter 2 introduces a hybrid modeling framework for membrane separation processes, particularly lithium recovery from batteries.

Chapter 3 formulates a strategy for identifying hybrid models within nonlinear liquid-phase column chromatography, replacing adsorption kinetics and equilibrium with Artificial Neural Network (ANN).

Chapter 4 is dedicated to developing a kinetics-discovery strategy for nonlinear column chromatography. This strategy harnesses the power of the Universal Differential Equation (UDE) approach and employs sparse and symbolic regression techniques to enhance the interpretability of the universal approximators. This methodology is primarily applied in the context of single breakthrough experimental data.

The second part of this work, Part II, explores modeling strategies and applications within the field of chemical engineering, with a particular focus on scenarios characterized by abundant data and minimal domain knowledge. It is important to note that data abundance can manifest in various ways, not necessarily limited to experimental data, and does not necessarily signify large quantities. Abundant data here denotes the availability of informative and non-redundant information. In "traditional" chemical engineering science, data abundance generally arises when working with reduced-order or surrogate models, where data originates from a mechanistic model that is not sufficiently rapid for real-time applications. In this way, Chapter 5 delves into identifying surrogate models for perfume diffusion, which is subsequently used for optimal fragrance formulation.

It is important to emphasize that data-driven models with extensive datasets can also be effectively integrated with mechanistic models identified from limited datasets. However, their identifications are carried out separately. This approach has gained increasing prominence, particularly in addressing challenges related to scale integration, where phenomena described at the atomic or molecular level are connected to process-scale models through machine learning techniques. While this scenario occurs at the intersection of the data and physics knowledge scenarios, it remains relevant to the overarching goal of seamlessly integrating machine learning and data-driven models with physics-based approaches in chemical engineering.

Therefore, Chapters 6 and 7 build upon the open issues introduced in Chapter 5, specifically focusing on two crucial aspects. Firstly, it strikes through the challenge of

accurately estimating vapor pressure in Chapter 6, even for odorants that are not synthesized. Secondly, Chapter 7 explores the integration of machine learning algorithms for odor classification into the prediction of fragrance classification. Additionally, this chapter delves into the significant issue of handling uncertainties associated with these parameters. The approach in Chapter 7 falls within scale integration, as it incorporates a machine learning model capable of predicting molecular properties and odor olfactory families based on their chemical structure. These predictions are then integrated into the perfume diffusion model, providing a comprehensive framework for addressing these complex challenges.

Part I

Physics-Informed Machine Learning in Chemical Engineering

Chapter 1

A First Approach Towards Adsorption-Oriented Physics-Informed Neural Networks

1.1 Abstract

This study explores the application of Physics-Informed Neural Network (PINN) in solving complex numerical problems encountered in adsorption systems, specifically in purifying monoclonal antibodies within a fixed-bed column. By comparing the performance of PINN with traditional numerical methods, this work demonstrates the accuracy and potential of PINN for tackling intricate numerical challenges in chemical engineering.

1.2 Introduction

Solving Partial Differential Equation (PDE) is crucial in several fields. PDEs are essential within engineering, ranging from Civil Engineering to Chemical Engineering. They are a mathematical representation of changes in quantities in the observed reality. The more complex the relationship of the observed phenomena, the more complicated the PDEs, and consequently, the more difficult it is to obtain their solution.

A small amount of PDEs has analytical solutions. However, numerical methods can be used for several of these problems. Following the nature of the PDEs, one can choose from various combinations to obtain its numerical solution (finite elements discretization, variational approach, among others). This has been the *de-facto* for engineering and physics modeling for several years. Recently, this field was revolutionized by the advent of Physics-Informed Neural Network (PINN) (Karniadakis et al., 2021; Raissi et al., 2019).

Using Artificial Neural Network (ANN) to solve PDEs is not new. It was first presented by Lagaris et al. (1998). The main idea is to leverage the universal approximation capacity of ANNs to solve PDEs. However, the method remained underused for many years due

to technology limitations. This strategy could be employed only after the advent of Deep Learning and advanced computing techniques. Raissi et al. (2019) presented the first practical application of this concept.

PINN can address problems where the classical approach requires significant effort and specialized methods – stochastic, fractional, high-order PDEs or high-index differential algebraic equations (Raissi et al., 2019; Karniadakis et al., 2021). However, PINNs are based on conservation equations as constraints of a custom-designed loss function. Therefore, the design of the loss function plays a pivotal role in this technique. Following the peculiarities of the PDE, supervised learning must be adequately designed. Thus, the PINNs methods should be carefully evaluated considering the application domain. As an emerging field of study, it still has limitations. The PINN prediction is limited to the initial conditions assessed in the PDE solution.

Moreover, PINN is initially obtained through a learning problem, which is computationally heavy. On the other hand, it is a very recent field with several potential. For instance, once the system of PDEs is solved by PINNs, the resulting Artificial Neural Network can be used as a model. The trained PINN model can predict the system states in a fraction of milliseconds, an enormous advance in rigorous simulation. Hence, it must be explored in the application domain to evaluate its full potential. Therefore, this is one of the main motivations of this work, adapting this technique to the domain of chemical processes and assessing its potentialities and limitations.

Chemical engineering problems are modeled with conservation and constitutive laws that yield complex PDEs. The peculiarities of the phenomenological models that describe chemical processes usually take the classical numerical methods to their limits. Different lengths and time scales are present in several examples, challenging the PDEs systems' solution. For instance, a system not limited by diffusion might present abrupt changes in its behavior, leading to problems solving the phenomenological model that describes this system. This is usual in processes that leverage adsorption to promote separation and/or reaction.

This work provides the first efforts in adsorption systems to apply PINNs techniques to solve the numerical problem resulting from these systems' constitutive and conservative laws. We study the peculiarities of these systems from the point of view of Physics-Informed Neural Network (PINN). Thus, we provide initial guidelines to address adsorption-based models using PINNs.

An adsorption analysis of a therapeutic Monoclonal Antibody (mAb) is presented as a case study. It is known that chemical components and thermal stress lead to the formation of protein aggregates during mAbs production. A high purity concentration has to be achieved if the aim is the use for a medicinal treatment (Nicoud et al., 2015). Ion-chromatography columns are applied for the purification step to separate the mAbs from their impurities and protein variants due to the difference of electrostatic affinity between them and the ion-exchange stationary phase. For simulating these systems, the electrostatic

information must accurately describe the adsorption/desorption. Furthermore, simulating the adsorption behavior of mAbs in a column is a PDE problem, in which it is needed to solve a set of equations that contains the phenomenological representation of a fixed-bed system. As an adsorption problem, it presents the above described challenges to its numerical solution.

Thus, this work evaluates the application of PINNs to address the solution of a phenomenological model of a chromatographic column to purify a mAb protein using binding and elution strategy, as applied by Lee et al. (2021). It is important to note that adsorption and desorption usually yield different behavior, which is a source of problems for the numerical solution of these systems. Therefore, evaluating the proposed PINNs structure is a good case study. This work paves the way to handle more complex problems in chemical engineering, namely, reactor modeling, gas-phase separation processes (Pressure/Temperature Swing Adsorption), and continuous chromatography (Simulated Moving Bed).

1.3 Methodology

The column chromatographic process is usually modeled as a fixed bed by writing differential mass conservation laws in mobile and stationary phases. Solving these equations can predict the adsorbate concentration at any time and position. Considering that transport occurs predominantly in the axial direction (z), constant axial dispersion coefficient (D_{ax}) the dimensionless mass balance in the mobile phase (bulk and boundaries) can be written as:

$$\frac{\partial c_i^*}{\partial t^*} + \frac{1 - \varepsilon}{\varepsilon} \frac{c_{i,0}}{q_{i,0}} \frac{\partial y_i}{\partial t^*} + \frac{\partial c_i^*}{\partial z^*} = \frac{1}{Pe} \frac{\partial^2 c_i^*}{\partial z^{*2}}; z \in (0, 1) \quad (1.1)$$

$$\frac{\partial c_i^*}{\partial z^*} = Pe(c_i^* - c_{in}^*); z = 0, t \geq 0 \quad (1.2)$$

$$\frac{\partial c_i^*}{\partial z^*} = 0; z = 1, t \geq 0 \quad (1.3)$$

$$c_i^* = 0 \text{ (adsorption)}; z^* \in (0, 1), t^* \geq 0 \quad (1.4)$$

$$y_i = \frac{K' c_i^*}{(1 + (K' - 1)c_i^*)}; K' - 1 = K c_{i,0} \quad (1.5)$$

$$q_{i,0} = \frac{q_{\max} K c_{i,0}}{1 + K c_{i,0}} \quad (1.6)$$

$$c_i^* = \frac{c_i}{c_{i,0}}, z^* = \frac{z}{L}, t^* = \frac{tu}{L}, Pe = \frac{uL}{D_{ax}} \quad (1.7)$$

where z^* is the dimensionless axial position, t^* is the dimensionless time, c_i^* is the dimensionless concentration, y_i is the stationary phase dimensionless concentration in equilibrium with c_i^* (mobile phase specie concentration), Pe is the axial Peclet number, ε is the

bed porosity, u is the interstitial velocity, L is the column length, $c_{i,0}$ is a reference mobile phase concentration (inlet or feed), $q_{i,0}$ is a reference stationary phase concentration, is the dimensionless concentration at initial condition, $q_{\max}$ the maximum adsorption capacity of the stationary phase, K' is the constant separation factor, and K is the dimensionless adsorbate-adsorbent interaction parameter given by:

$$K = \frac{A_s}{1 - \varepsilon} \int_0^\infty \left[\exp \left(-\frac{W(I, pH, h)}{k_b T} \right) - 1 \right] dh \quad (1.8)$$

Where W is the binding free energy, A_s is the accessible surface area per unit of packed bed volume ($4.2 \times 10^7 \text{ m}^{-1}$) (Guélat et al., 2012), I is the ionic strength, h is the distance between the Monoclonal Antibody (mAb) and the stationary phase surfaces, k_b is the Boltzmann constant, and T is the temperature.

The adsorption equilibrium constant is a thermodynamic property, a function of W (potential of mean force, i.e., PMF), related to the adsorption binding affinity. Notably, electrostatic interaction is the leading property designating biomolecule surface interactions. The literature (de Souza Gama et al., 2018; Gama et al., 2021) reports how to calculate the PMF from the electrostatic potential obtained by solving a modified Poisson-Boltzmann (PB) equation in a different set of coordinates. The modification in the PB equation allows the PMF to be a function of the solution conditions, i.e., temperature, ionic strength, salt type, and pH. It is beyond the purpose of this work to show how to solve the modified PB equation. Thus, we applied the exact calculation details reported in the literature (Gama et al., 2021). In this way, the Langmuir isotherm (Equation 1.5) is linked with a thermodynamic parameter (Equation 1.8), becoming a function of the mAb and stationary phase's surface properties and the solution conditions.

1.3.1 Simulation Scenario

A simulation of an adsorption-based separation is considered here to assess the PINN performance. Hence, the adsorption of a mAb with known charge density taken from Guélat et al. (2010) over YMC BioPro SP ($10 \mu\text{m}$) where the solid phase is initially protein-free is evaluated. The liquid phase is at a constant pH of 5.9, and ionic strength 100 mM, and a feed stream at a concentration of 15 g/L is fed at the column inlet. Figure 1.1 describes the antibody protein ion-exchange adsorption schematically. The figure presents the ions and the protein/adsorbent surfaces charged, depicting the system mechanisms.

Under these conditions, the adsorption equilibrium constant (K , from Equation 1.5) was estimated as 4.004 by solving the modified PB equation, as previously mentioned (de Souza Gama et al., 2018; Gama et al., 2021). The chosen value for K aims to represent an intermediary condition in protein adsorption. In a typical protein separation scenario (where high efficiency is required), K should be much higher during adsorption and less during desorption. However, the scope of this work is comparing traditional numerical methods with a new approach (PINNs), and the chosen value for K fits this purpose

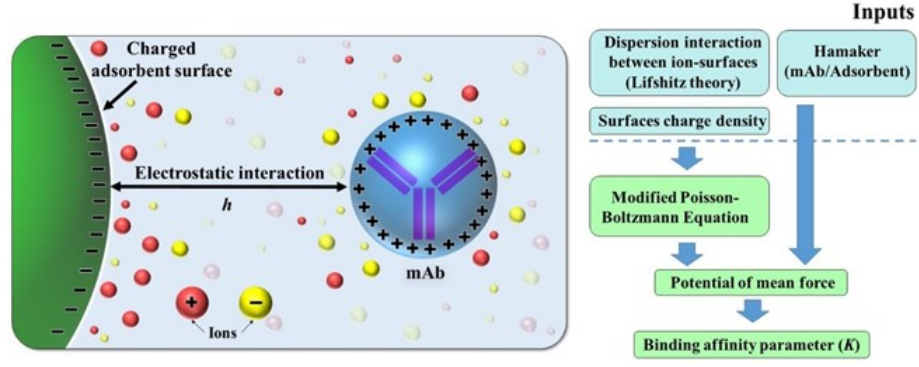


Figure 1.1: Schematic representation of the antibody adsorption by ion-exchange

well. Moreover, the maximum adsorption capacity ($q_{\max}$) is set as 125 g/L for the referred protein-adsorbent pair. We used the following hydrodynamic parameters reported in the literature (Guélat et al., 2012) as reference for mAb performance in ion-exchange columns, namely: bed porosity (ε), column length, column diameter, injection volume, and axial Peclet number (Pe) were set to be 0.53, 8.5 cm, 0.5 cm, 1.2 cm³, and 500, respectively.

The solution of the fixed bed model via Method of Lines (MoL) was implemented in gPROMs Model Builder advanced process simulator with 60 finite elements and 3 collocation points per element. The resulting differential-algebraic system of equations was solved with the default DASOLV solver and automatic index reduction algorithm available in gPROMs. The Physics Informed Neural Network was built and trained in the DeepXDE (Lu et al., 2021) version 0.14.0.

1.3.2 Physics-Informed Neural Networks

Given a generic space (x), a time-dependent nonlinear partial differential equation in the form:

$$u_t + \mathcal{N}(u) = 0; x \in \Omega, t \in [0, T] \quad (1.9)$$

and boundary conditions with nonlinear differential operator $\mathcal{B}[\cdot]$:

$$\mathcal{B}(u, x, t) = 0 \text{ on } \partial\Omega \quad (1.10)$$

where $u(t, x)$ denotes the latent solution of the PDE, u_t is the time derivate of u and $\mathcal{N}[\cdot]$ is the nonlinear differential operator. One can approximate the unknown PDE solution $u(x, t)$ by a deep neural network $\hat{u}(x, t, \theta)$ with trainable parameters θ and derive the called Physics-Informed Neural Network (Raissi et al., 2019) $f(\hat{u}, x, t)$ and $g(\hat{u}, x, t)$ as:

$$f(\hat{u}, x, t) = \hat{u}(x, t, \theta) + \mathcal{N}(\hat{u}(x, t, \theta)) \quad (1.11)$$

Then, one can train the neural network $\hat{u}(x, t)$ to satisfy the $f(\hat{u}, x, t)$, $\mathcal{B}(\hat{u}(x, t, \theta))$ and initial condition constraints. In practice, a set of points inside the space-time domain ($\mathcal{T}_f$), another set in the boundaries ($\mathcal{T}_b$), and at initial condition ($\mathcal{T}_i$), are sampled and used to calculate the overall loss function $\mathcal{L}(\theta, \mathcal{T})$ as:

$$\mathcal{L}(\theta, \mathcal{T}) = w_f \mathcal{L}(\theta, \mathcal{T}_f) + w_b \mathcal{L}(\theta, \mathcal{T}_b) + w_i \mathcal{L}(\theta, \mathcal{T}_i) \quad (1.12)$$

where,

$$\mathcal{L}(\theta, \mathcal{T}_f) = \frac{1}{|\mathcal{T}_f|} \sum_{x,t \in \mathcal{T}_f} |f(\hat{u}, x, t)|^2 \quad (1.13)$$

$$\mathcal{L}(\theta, \mathcal{T}_b) = \frac{1}{|\mathcal{T}_b|} \sum_{x,t \in \mathcal{T}_b} |g(\hat{u}, x, t)|^2 \quad (1.14)$$

$$\mathcal{L}(\theta, \mathcal{T}_i) = \frac{1}{|\mathcal{T}_i|} \sum_{x,t \in \mathcal{T}_i} |\hat{u}(x, t) - u(x, 0)|^2 \quad (1.15)$$

Figure 1.2 depicts the Physics-Informed Neural Network (PINN) building process to find the solution of differential equations.

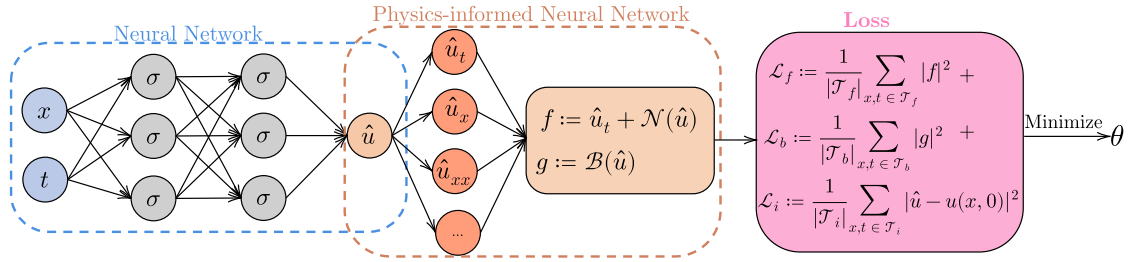


Figure 1.2: Physics-informed Neural Network building process.

Hence, considering the problem of protein adsorption in a fixed-bed column, we propose the loss functions:

$$\mathcal{L}_f = \frac{1}{|\mathcal{T}_f|} \sum_{x,t \in \mathcal{T}_f} \left\| \frac{\partial \hat{u}}{\partial t^*} + \frac{1-\varepsilon}{c_0 \varepsilon} \frac{\partial}{\partial t^*} \left(\frac{K \hat{u} c_0}{1 + \frac{K \hat{u} c_0}{q_{\max}}} \right) + \frac{\partial \hat{u}}{\partial z^*} - \frac{1}{Pe} \frac{\partial^2 \hat{u}}{\partial z^{*2}} \right\|^2 \quad (1.16)$$

$$\mathcal{L}_{bl} = \frac{1}{|\mathcal{T}_{bl}|} \sum_{x,t \in \mathcal{T}_{bl}} \left\| \frac{\partial \hat{u}(0, t, \theta)}{\partial z^*} - Pe (\hat{u}(0, t, \theta) - c_{in}^*) \right\|^2 \quad (1.17)$$

$$\mathcal{L}_{br} = \frac{1}{|\mathcal{T}_{br}|} \sum_{x,t \in \mathcal{T}_{br}} \left\| \frac{\partial \hat{u}(1, t, \theta)}{\partial z^*} \right\|^2 \quad (1.18)$$

$$\mathcal{L}_i = \frac{1}{|\mathcal{T}_i|} \sum_{x,t \in \mathcal{T}_i} \left\| \hat{u}(x, 0, \theta) - c_0^* \right\|^2 \quad (1.19)$$

By minimizing the sum of Equations 1.16 to 1.19 with an appropriate optimizer or a combination of optimizers, one can find the neural network $\hat{u}(z, t, \theta)$ parameters θ that minimize the loss, leading to the solution of the PDE.

1.4 Results

This section provides the numerical experiment to assess the PINNs performance. The proposed solution is compared with the current state-of-the-art numerical method (MoL). Usually, these models are solved using MoL with axial position discretized with orthogonal collocation on finite elements.

For the numerical adsorption experiment, the feed-forward architecture was employed. The hyperparameters, i.e., layers and neurons, were searched via random grid search. The best architecture, i.e., with minimum training error, had six layers with 80 neurons per layer. Layer-wise, each layer's locally adaptive hyperbolic tangent activations proposed in (Jagtap et al., 2020) were used to improve convergence speed. The loss function weights were equal to 1.0 for all terms except for the boundary condition term at axial position $z=0$. For this term, the corresponding weight (W_{bcl}) was adaptively changed during training, starting at 1×10^{-4} and finishing at 1.0 due to the associated high loss at the beginning of the training compared to other terms. An additional term was included in the loss function, namely, the derivative of the PDE residuals with respect to both dimensionless time and position (t^*, x^*) weighted by 1×10^{-4} as it was previously shown to improve training loss (Jagtap et al., 2020).

A total of 2500 collocation points were used for the PDE residual ($\mathcal{T}_f$), 800 for both boundary conditions ($\mathcal{T}_b$), and 500 for the initial condition ($\mathcal{T}_i$). The points were sampled using the Latin Hypercube Sampling strategy (McKay et al., 1979). Then, the loss function was minimized by running 25000 iterations of Adam (Kingma and Ba, 2015) algorithm at a 3.0×10^{-4} learning rate with early stopping (5000 iterations patience) followed by L-BFGS-B (Nocedal and Wright, 2006) until convergence.

Table 1.1 summarizes the neural network hyperparameters for the numerical adsorption experiment.

Table 1.1: Parameters to define the neural network for the adsorption experiment.

Parameter	Description	Value
N_l	Number of layers	6
N_n	Number of neurons per layer	80
$w_f, w_{bcl}, w_{bcr}, w_i$	Weights of terms in the loss function	[1.0, 1.0, adaptive, 1.0]
$\mathcal{T}_f, \mathcal{T}_b, \mathcal{T}_i$	Collocation points of PDE, boundary, and initial condition	[2500, 800, 500]
α	Learning rate of Adam algorithm	3×10^{-4}

Figure 1.3 displays the training history on a semi-log scale. It shows that the loss starts decreasing quickly and monotonically, then stagnating at around 0.1. When the L-BFGS-B

algorithm is used, a pronounced drop in the loss is reached until it reaches the final loss (order of 10^{-4}).

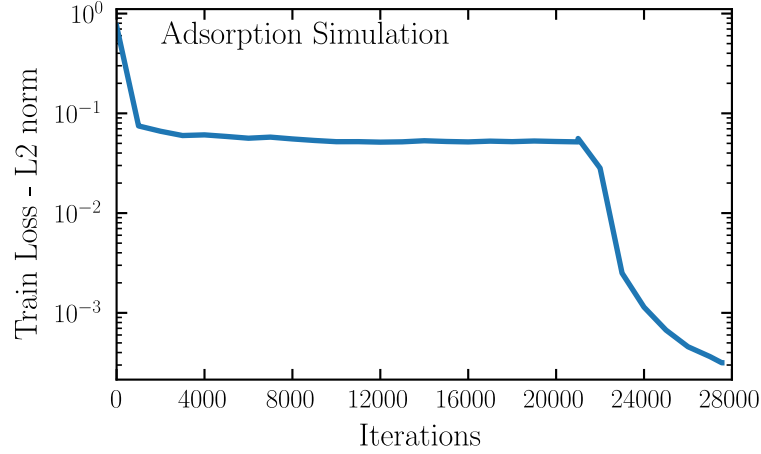
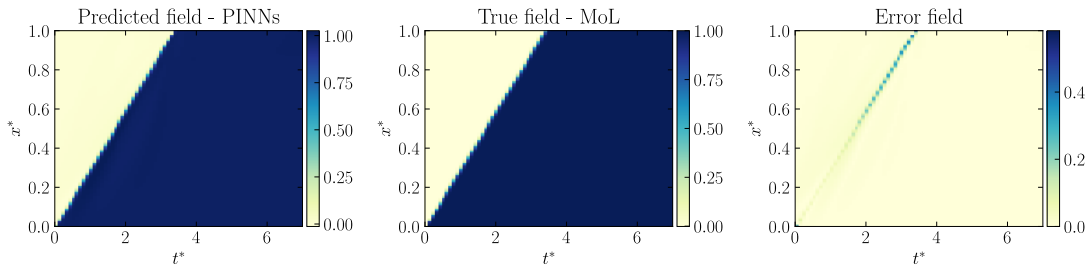


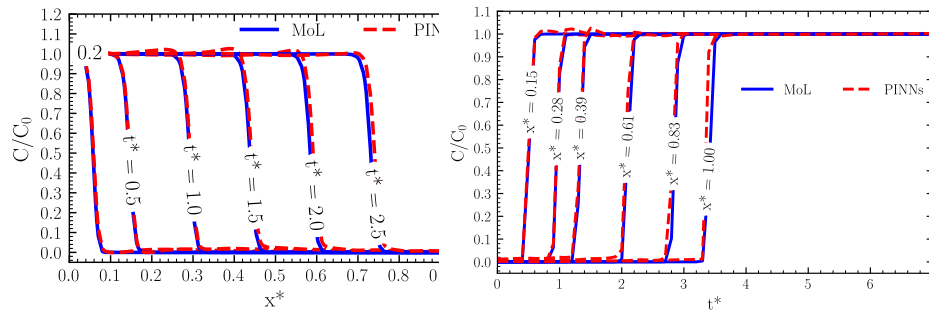
Figure 1.3: Train loss history for numerical adsorption experiment.

Figure 1.4 compares the MoL and PINNs solutions from 3 different perspectives. The first (1.4a) shows the solutions and error fields as a color map in the dimensionless space-time (x^*, t^*) domain, i.e., for every tuple (x, t) , the solution of the PDE can be read by matching the corresponding color at the color bar in the figures' right-hand size. The rightmost figure shows the absolute difference between the solution field obtained by PINN and by MoL. It can be seen that the PINN could learn the PDE solution satisfactorily with the largest errors (mean absolute error) found in the breakthrough region. In this line, the inlet concentration reaches these points. Figure 1.4b shows the concentration profile at six different dimensionless times. It is possible to see that at higher times and closer to the upper boundary ($x^* = 1$), the error is more pronounced; however, the profiles are satisfactorily learned by the PINN model. Figure 1.4c shows the history of the concentration at several axial positions. It shows that the concentration changes suddenly, characterizing a compressive concentration front propagation.

The numerical adsorption experiment is challenging to simulate for any numerical method. As known (Hornik et al., 1989), neural networks are universal approximators for C1 continuous functions, and fields with steep gradients as those seen in adsorption are challenging. Even so, the PINN approximates the PDE solution satisfactorily (mean absolute error calculated from ground truth of less than 1%). It is important to highlight that the PINNs are replacing the numerical method to solve the PDE system that composes the mechanistic model of the adsorption separation. Hence, it is preferred to compare its solution with a classical numerical method, the MoL, rather than experimental data. Furthermore, the model used as a reference was already experimentally validated in the literature. The validation against experimental data can be done in future applications of this method in other systems, which would be an important development of this work.



(a) Solution field.



(b) Solution profile

(c) Solution history

Figure 1.4: MoL and PINNs solution's comparison for adsorption experiment.

Therefore, we certify that the PINNs can efficiently solve the numerical problem within a continuous domain. This also allows us to compare the computational effort of trained PINNs against the numerical solution by a classical numerical method. For instance, the MoL solved the model in 37s. While the PINNs, once trained, provide a prediction in a fraction of a second. The comparison was made in an Intel Core i7-7500U CPU @ 2.70GHz 2.90 GHz. This difference is significant in an optimization context, where hundreds of thousands of model evaluations are done. In this way, the PINNs might employ mechanistic-informed models online, such as real-time optimization, process control, and optimization. This can unlock several new possibilities in the field of Process Systems Engineering.

1.5 Conclusions

This work presents one of the first attempts to solve numerical problems arising from adsorption models with physics-informed neural networks. This is a recent field of study that needs further studies to explore its full potential. This work provides the basis for this direction in chemical engineering, specifically in adsorption-based systems.

One key aspect to consider when comparing PINN with traditional numerical solvers is that the former does not require special treatment for issues arising in methods such as MoL: a careful choice of spatial discretization scheme, index reduction methods, and

appropriate time integrators for stiff systems. In a PINN solver, space-time points are sampled straightforwardly, and the problem is solved when the residuals reach an acceptable value (L2 norm in the order 1×10^{-4} for conservation laws is enough for most practical applications). Moreover, the PINN can accept additional independent variables in the input layer apart from spatial and time, such as PDE, geometrical parameters, and initial and boundary conditions. Then, inference-time predictions can be delivered in a fraction of a second at any position and time within the training domain for several system configurations. Once the PINN is known to solve the PDE describing adsorption, it can be safely used for parameter estimation with simultaneous solution of the PDE and assimilation from sensor data and time-demanding tasks with traditional numerical solvers.

Therefore, the results of this work demonstrated that the PINNs are efficient in addressing the numerical solution of fixed-bed protein separation by adsorption. Both monoclonal antibody adsorption were calculated using the first-principles model solved using the physical-informed networks. The next step to explore this technique would be to evaluate more complex processes, such as continuous chromatography. From the current state of the literature and the present results, it is possible to envision the significant impact of the PINNs on the chemical systems phenomenological modeling.

Chapter 2

A hybrid modeling framework for membrane separation processes: Application to the Lithium ions recovery from batteries

2.1 Abstract

Hybrid modeling occupies a unique position at the intermediate point of two contrasting modeling paradigms and data scenarios: data-driven and physics-based modeling on one end and small data and big data on the other. It harnesses the advantages of valuable domain knowledge and mechanistic understanding to compensate for small data sets in the scientific domain. This synergy between data-driven insights and physics-based principles enhances the precision and interpretability of models. Compared with the Physics-Informed Neural Network approach, hybrid modeling enables the seamless integration of comprehensive data with underlying physical laws, resulting in more accurate and robust solutions for complex problems in various domains, including chemical engineering.

This chapter kicks off the part of hybrid modeling through a framework for membrane separation processes where lithium from batteries is recovered. This is a pertinent problem nowadays as lithium batteries are popularized in hybrid and electric vehicles. The hybrid model is based on an Artificial Intelligence (AI) structure to model the mass transfer resistance of several experimental separations in the literature. It is also based on a phenomenological model representing the transient system regime. An optimization framework is designed to perform the AI model training and simultaneously solve the Ordinary Differential Equation (ODE) system representing the phenomenological model. The results demonstrate that the hybrid model can better represent the experimental validation sets than the phenomenological model alone. This strategy opens doors for further investigations of this system.

2.2 Introduction

Phenomenological modeling is essential for systems design, development, and optimization. Usually, phenomenological modeling is done by applying conservation laws, capable of representing the known phenomenology through one or a system of differential equations. Though consolidated and reliable, this approach requires more significant computational effort to provide the results. The more complex the system is, the more complex the model itself gets. On the other hand, the modeling of first principles is usually associated with system parameters, such as mass transfer coefficient, reaction rate, etc. These parameters are generally embedded into a constant that needs to be estimated through experimental data. Thus, how these parameters are added to the conservation laws is usually associated with simplifications that allow the model to be identifiable. This limiting factor will be reflected in the model's precision and reliability. Empirical modeling is an alternative to phenomenological modeling; it is a data-oriented strategy that extracts the system information and stores it in the model structure. The data-driven empirical modeling has been explored in chemical engineering to solve several problems in this field (Narayanan et al., 2021b,a; Nogueira et al., 2017). This strategy efficiently identifies reliable models with a low computational effort and can be easily applied in online applications. On the other hand, empirical modeling requires a significant amount of data to be identified due to its limited extrapolation capacity.

The advances in both fields, phenomenological and empirical modeling, led to hybrid models, namely hybrid models based on Artificial Intelligence (Narayanan et al., 2021b,a; Mogk et al., 2002). This strategy leverages both strategies' advantages to build a more complex and reliable model without increasing the associated computational effort. In Nagrath et al. (2004), a hybrid modeling framework for chromatographic process optimization is presented. The authors highlight the ability of the proposed strategy to explore several different design scenarios without additional computational effort. Still, in the chromatographic separation field, Narayanan et al. (2021b) presents a study regarding a phenomenological model hybridization level. The authors point to the gains in prediction accuracy and extrapolation ability as the main advantages of this strategy.

Furthermore, the referred work indicates that these solutions are more likely to be implemented in practical applications. Even though the potential of hybrid modeling is emphasized in the literature, its application to address chemical engineering problems is still limited. For instance, in the membrane separation field, it would have the potential to address issues related to identifying a proper strategy to model the system mass transfer resistances. The membrane separation is composed of a series of elements that promote a driving force between the membrane sides; therefore, promoting the separation. These elements are translated into mass transfer resistance coefficients that should be estimated to obtain a phenomenological system model. As an alternative, a global mass transfer coefficient can be defined to simplify the model and the system identification. In this

scenario, the hybrid modeling approach presents the potential to address this problem to provide more versatility and flexibility to the membrane models.

Hence, this work proposes a hybrid modeling framework for membrane separation processes. As a case study, it presents the recovery of lithium from batteries, which is a pertinent case in the current scenario where lithium batteries are popularized in hybrid and electric vehicles. The proposed framework uses Artificial Intelligence (AI) structures to model the mass transfer resistance of several experimental separations in the literature. The AI models are incorporated into the phenomenological membrane model. An optimization strategy trains the AI model and simultaneously solves the Ordinary Differential Equation (ODE) that represents the phenomenological model.

2.2.1 Case Study: Lithium ions recovery from batteries

Solvent extraction, membrane separation, adsorption, and electrochemical methods are the most common techniques used for lithium separation from lithium-ion batteries (LIBs). Liquid membrane techniques, mainly supported liquid membranes (SLMs), have been increasingly studied for treating and purifying solutions in hydro-metallurgical industries (Benzal et al., 2004; Leon et al., 2016; Wang et al., 2015). SLM can be designed in several configurations: flat sheet-supported liquid membranes (FSSLM) and hollow fibers-supported liquid membranes (HFSLM) (Benzal et al., 2004; Leon et al., 2016; Wang et al., 2015; Peydayesh et al., 2013).

The mass flux of a species is defined as the amount of matter passing through an area unit and during a time unit. This flux depends on various factors, such as choice and concentrations of carrier and solvents, choice and type of membrane support, and volumes of aqueous phases. To estimate or represent the experimental data of separations, several parameters must be measured experimentally, such as permeabilities and mass transfer coefficients. However, in addition to the costs related to the experimental work, if experimental data is unavailable or challenging to obtain, it is essential to have a preview or an estimation for the solute transference. This way, to assess the process dynamic behavior, a phenomenological model, which can be complex, is used, and its simulation can require great computational time and effort.

For the industry, simplifying this model can help predict mass transfer in a simple way, which is proposed in the work reported by Regufe et al. (2021a). The simplified model proposed estimating the global ionic mass transfer resistance with only a few experiments. However, the prediction needs to use the experimental data of the feed and receiving phases. This assessment in the real-time process is challenging, which is a barrier to using a phenomenological model. Therefore, it is necessary to find alternative models that can provide accurate real-time information about the process dynamics.

In this context, the present chapter proposes a flat sheet-supported liquid membrane

(FSSLM) for lithium recovery as a case study. FSSLM is the most usual and most straightforward type of liquid membrane supported in microporous solids. The membrane is positioned between two sections: a section where the solution is fed and another for the stripping solution; these sections are stirred by mechanical agitation. This unit is depicted in Figure 2.1, and its main advantage is simple assembling. On the other hand, it presents a low surface area as the main drawback.

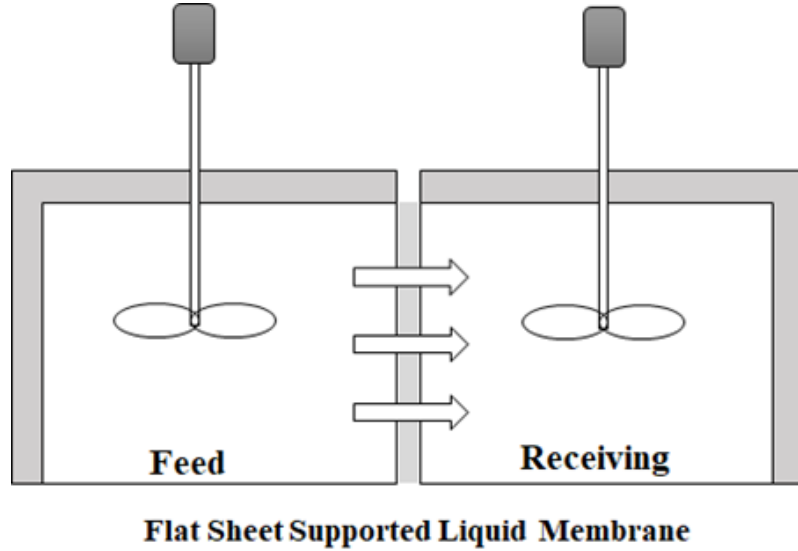


Figure 2.1: Schematic representation of an FSSLM.

In the next section, the proposed hybrid ODE-ANN model is presented. This novel strategy is compared to this system's phenomenological model to understand better the methodology proposed here. Furthermore, the objective function of the system identification step is presented together with the parameters estimation strategy.

2.3 Materials and Methods

2.3.1 Hybrid model

The phenomenology involved in a liquid membrane separation comprises the aqueous feed phase, the organic phase (membrane), and the aqueous receiving phase. The existing literature provides the first principle models representing the transport mechanism through the supported liquid membrane containing different carriers.

Regufe et al. (2021a) presents a phenomenological model for the present case study. In the referred work, the authors developed a simplified strategy to identify phenomenological models for membrane separation using a few experimental data for model validation. Regufe et al. (2021a)'s method identifies two time constants, one for the receiving phase and another for the feed phase, to improve the model accuracy. This strategy was demonstrated to have accuracy close to more detailed phenomenological modeling strategies.

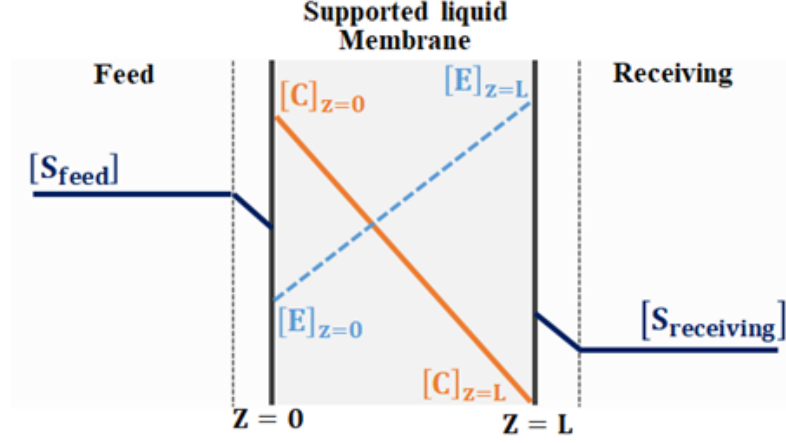


Figure 2.2: Schematic representation of the membrane transport profile: solute (S), metal-carrier complex (C), and carrier (E)

Therefore, the phenomenological model proposed by Regufe et al. (2021a) that considers an overall mass transfer coefficient was used as a base. A schematic representation of the membrane transport profile is depicted in Figure 2.2.

This transport phenomenon comprises a solute, S , that reacts with the carrier, E , at the feed-membrane interface, forming the complex species, C . The complex species diffuses through the liquid membrane until it reaches the membrane-strip interface, where it reacts with the stripping solution. Then, the concentration gradient will promote a driving force to conduct this solute in the strip solution, diffusing towards the bulk strip solution.

This system, extraction of a metal ion by FSSLM, is portrayed by the following equilibrium, which is valid at both interfaces (feed/membrane and membrane/receiving):



where S is the solute, E is the carrier at the membrane interfaces, C is the complex species, X is the component in the carrier exchanged with S (if the solute is an ion, then X is denominated as counter-ion), and α and β are the stoichiometric coefficients.

Assuming ideal behavior, the system dynamic behavior can be described, as proposed by Regufe et al. (2021a), as:

$$-\frac{dS_f}{dt} = \phi_f(V_f, A, S_f)(S_f - S_{f,i}) \quad (2.2)$$

where ϕ_f is an empirical model that describes the mass transfer coefficient in the feed-membrane interface, S_f and $S_{f,i}$ are respectively the solute/metal concentrations at the feed phase and the feed-membrane interface, $S_{f,0}$ is its corresponding initial concentration, V_f is the feed phase volume, and A is the effective interfacial area.

It is essential to highlight that, unlike a phenomenological model, the proposed strategy uses an empirical model (Artificial Neural Network) to portray the film diffusion transport

mechanism. Usually, this phenomenon is described by a single parameter, which is also implemented here to provide a comparison source to the proposed approach.

To estimate S_r^* , it is used:

$$S_r^* = S_r \left(\frac{X_f}{X_r} \right)^\beta \quad (2.3)$$

where X_r represents the component in the carrier exchanged with S_r , X_f represents the component in the carrier exchanged with S_f (when the solute is an ion, X is denominated as counter-ion), and β represents the stoichiometric coefficient.

The accumulation in the membrane phase is neglected, and it is imposed the conservation in both S and X species:

$$V_f(S_{f0} - S_f) = V_r(S_r - S_{r0}) \quad (2.4)$$

$$V_f(X_f - X_{f0}) = V_r(X_{r0} - X_r) \quad (2.5)$$

The stoichiometric exchange at the feed/membrane interface gives:

$$\beta(S_{f0} - S_f) = (X_f - X_{f0}) \quad (2.6)$$

The mass balance for the receiving phase is computed by:

$$\frac{dS_r}{dt} = \phi_r(V_r, A, S_r)(S_{r,i} - S_r) \quad (2.7)$$

where V_r is the receiving phase volume, S_r , $S_{r,0}$ and $S_{r,i}$ represent the solute concentration at the receiving phase and the membrane-receiving interface. As in Equation 2.7, the concentration in the receiving phase is conditioned by the stagnant film in the interphase; at this point, the film diffusion is the transport mechanism. As previously mentioned, the present work leverages the potential of the ANN model to model this phenomenon, building a hybrid model precisely.

Then, the conservation of counter-ions in the receiving phase is given by:

$$X_r = X_{r,0} - \beta(S_r - S_{r,0}) \quad (2.8)$$

where X_r is the component in the carrier exchanged with S_r .

The model was implemented in Python and validated using experimental data from the literature, as will be detailed in the following sections.

2.3.2 Data Collection

FSSLM has gained attention in lithium recovery from spent lithium-ion batteries. This way, developing models to represent the transport mechanism through the supported liquid membranes is essential to reduce experimental test work and decrease costs. Besides

lithium, studying the transport behavior of other metal ions or solutes in wastewater treatment is crucial. In this context, experimental data on separating and extracting metal ions from aqueous phases using FSSLM was collected to train and validate the models. The details of each data set used are presented in Table 2.1. Each training set requires an identification procedure for a specific ion.

Thus, the ANNs will represent the mass transfer for a particular ion given its inputs, as presented in Equations 2.2 and 2.7.

Table 2.1: Details for the experimental data collected in the literature and used to train and validate the hybrid model.

Data set	Characteristics
Training and Validation sets I	Reference: (Bhatluri et al., 2015) Extraction of two heavy metals, lead(II) and cadmium(II) Feed and receiving phases: 250 cm ³ Area: 8.0 cm ² Carrier-solvent: sodium salt of Di-2-ethylhexylphosphoric acid (D2EHPA)
Training and Validation sets II	Reference: (Zaheri et al., 2017; Zante et al., 2019) Extraction of lithium Feed and receiving phases: 250 cm ³ Area: 8.0 cm ² Carrier-solvent: Cyanex272
Training and Validation sets III	(Zante et al., 2020) Extraction of cobalt and nickel Feed and receiving phases: 250 cm ³ Area: 17.8 cm ² Carrier-solvent: tri(hexyl)tetradecyl phosphonium chloride
Training and Validation sets IV	Reference: (Zante et al., 2019) Transport of Ge(IV) Feed and receiving phases: 60 cm ³ Area: 4.5 cm ² Carrier-solvent: Alamine 336

2.3.3 Hybrid Modeling

With the hybrid model proposed, the next step is to address the issues related to the model identification. As it comprises phenomenological and empirical models, its identification should comprise its peculiarities. The phenomenological model is a system of ODE equations solved by a conventional Runge-Kutta algorithm. However, the phenomenological part depends on the empirical one, which requires an identification step. Therefore, a framework for this model identification needs to be established. The framework proposed here is based on the work of Narayanan et al. (2021b).

A scheme of this framework is provided in Figure 2.3, depicting an overview of the employed strategy. It consists of an initial step of initialization of the empirical model parameters. These parameters must be estimated based on the experimental data collected in the literature. The solution of the empirical model provides the mass transfer information to the phenomenological model, which is solved by the numerical method. Thus, a

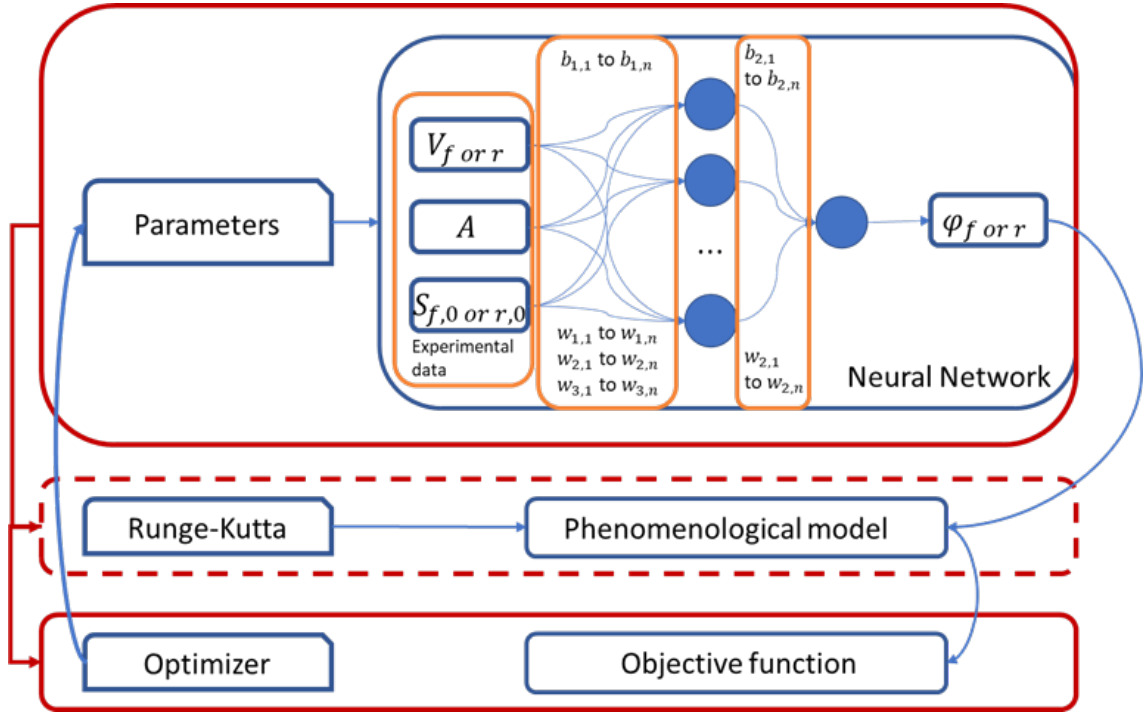


Figure 2.3: Schematic representation of the proposed framework.

predicted value of system concentrations is given, which is used to compute the parameters estimation objective function. The particle Swarm optimizer calculates the new set of parameter values, sending it back to the Neural Network. This procedure is repeated until the objective function is minimized and the model-required precision is achieved.

2.3.4 Parameter Estimation

As shown in Figure 2.3, parameter estimation is represented in the last layer of the proposed identification framework, where the optimizer minimizes the objective function. Therefore, an objective function was designed to address the proposed problem. The objective function adopted here is based on the difference between the predicted and experimental concentrations. However, it is essential to note that the ANN models, whose parameters will be estimated, do not predict the concentrations. Instead, they predict the mass transfer effects in the system, as previously described. In this way, the solution of the parameter estimation problems requires the complete resolution of the hybrid model. This strategy goes against the usual approach in ANN model identification, where the ANN models are identified independently and are meant to predict the main property straight. Therefore, this step needs to be done simultaneously with the solution of the complete hybrid model. Hence, the objective function that should be solved to proceed with the parameter estimation is defined as a least-square problem:

$$\min_{\lambda} F_{\text{obj}} = \sum_{j=1}^{NS} \sum_{i=1}^{NE} [(S_r - S_r^e)^2 + (S_f - S_f^e)^2] \quad (2.9)$$

$$\text{s.t. : } S_f = f(\phi_f(V_f A S_{f,0})), \quad (2.10)$$

$$S_R = f(\phi_r(V_r A S_{r,0})). \quad (2.11)$$

where NE is the total number of experimental points in the experimental set, NS is the number of experimental sets. S_r^e and S_f^e represent the experimental solute concentrations at the receiving and feed phases, λ is the set of decision variables of the problem, in this case, the matrix of ANN's parameters to be estimated.

2.4 Results

Finally, the framework presented here can be applied in identifying and validating the hybrid model for the FSSLM process. Therefore, the optimization problem described in subsection 2.3.4 needs to be solved. It was solved using a Constrained Sliding Particle Swarm Optimization (CSPSO) proposed by Rebello et al. (2021) with 400 particles and 400 iterations. As in the present case, the CSPSO was applied to cope with multimodal problems. Therefore, following this trial-and-error strategy, an optimal number of neurons equal to 20 in one hidden layer was identified, activated by hyperbolic tangent functions.

The hybrid modeling strategy proposed here was applied to model the lithium recovery in an FSSLM process. The case study evaluated for lithium separation is a highly selective membrane for lithium cations, as previously reported in the literature. In Zante et al. (2020), the lithium transport across this supported liquid membrane is reported. The referred work demonstrates that lithium cations are selectively separated from aqueous solutions containing sodium, cobalt, and nickel ions in this membrane separation. This demonstrates this process's potential in recovering lithium from battery solutions. The experimental data used to validate the proposed strategy were obtained from Zante et al. (2020), presenting a liquid membrane with very high selectivity for separating lithium from sodium. These experimental data were obtained for a feed of LiCl (0.1 mmol/L), pH = 12, and a supported liquid membrane HFDOD (0.2 mol/L) and TOPO (0.1 mol/L).

The models were identified using the literature data, and validation was done against another independent literature data set. Figure 2.4 presents the models' validation with the sole phenomenological model prediction obtained from Regufe et al. (2021a) to compare the proposed approach. The phenomenological model uses two different time constants, one for the feed and another for the receiving, as presented in Regufe's work, to compare the strategies fairly.

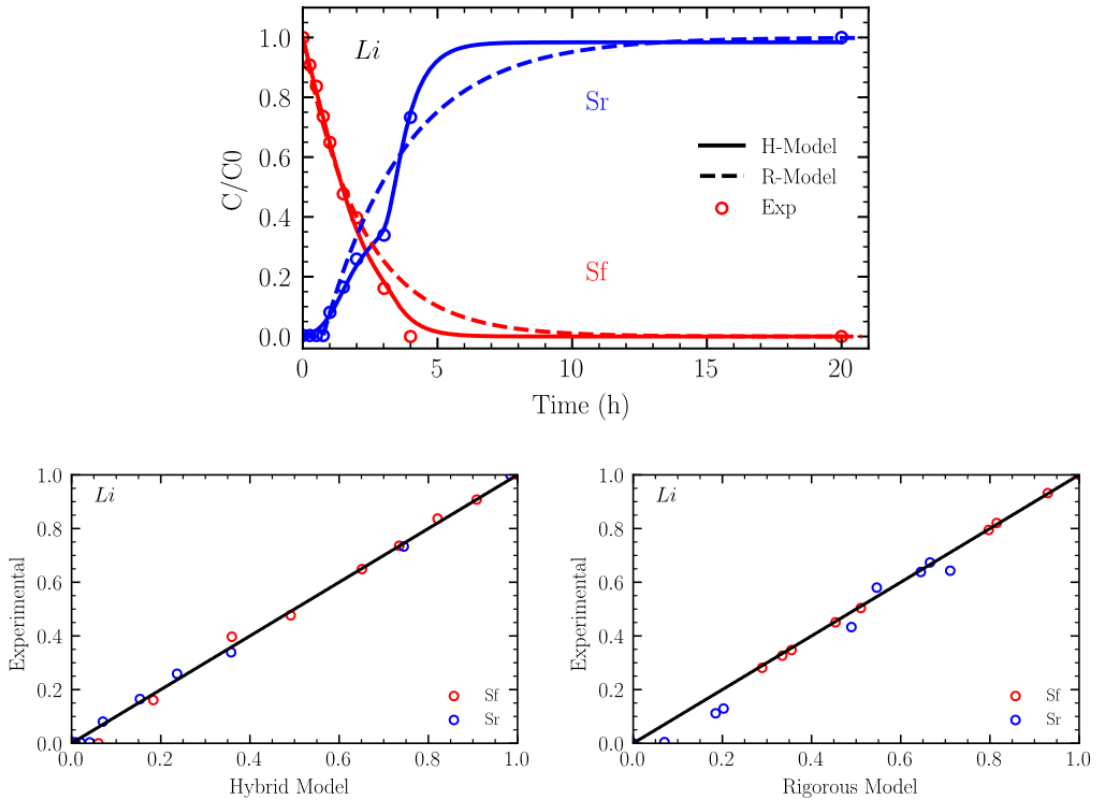


Figure 2.4: Lithium experimental set 1 - Hybrid model prediction (H – Model) versus the phenomenological model prediction (R – Model) and their respective parity graphics

Still, in Figure 2.4, it is possible to analyze the model residuals in the parity, which displays the hybrid model prediction versus the experimental data and the phenomenological model prediction versus the experimental data. The plots display point density along the diagonal line. It can be seen that the predictions from the hybrid model are randomly distributed around the diagonal line across the whole range of values. This is evidence that the residuals are random. Therefore, the model was satisfactorily identified. In contrast, the phenomenological model presents a nonrandom pattern within some regions. These points correspond to the transient regime of the process, pointed out to be a limitation in the phenomenological model.

Another experimental set was used to demonstrate the hybrid model prediction for the lithium separation, of which results are presented in Figure 2.5. This experimental data was obtained from Zante et al. (2020) for a feed phase composition of LiCl (619 mg/L) and receiving phase composition of Na₂CO₃ and NaHCO₃. This liquid membrane was impregnated with a mixture of hydrophobic ionic liquid 1-butyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide and tri-n-butyl phosphate as the carrier.

From Figure 2.5, it is possible to see that the proposed hybrid model can predict the dynamic behavior of this separation with precision. In contrast, the phenomenological model presents difficulties in predicting the initial transient state.

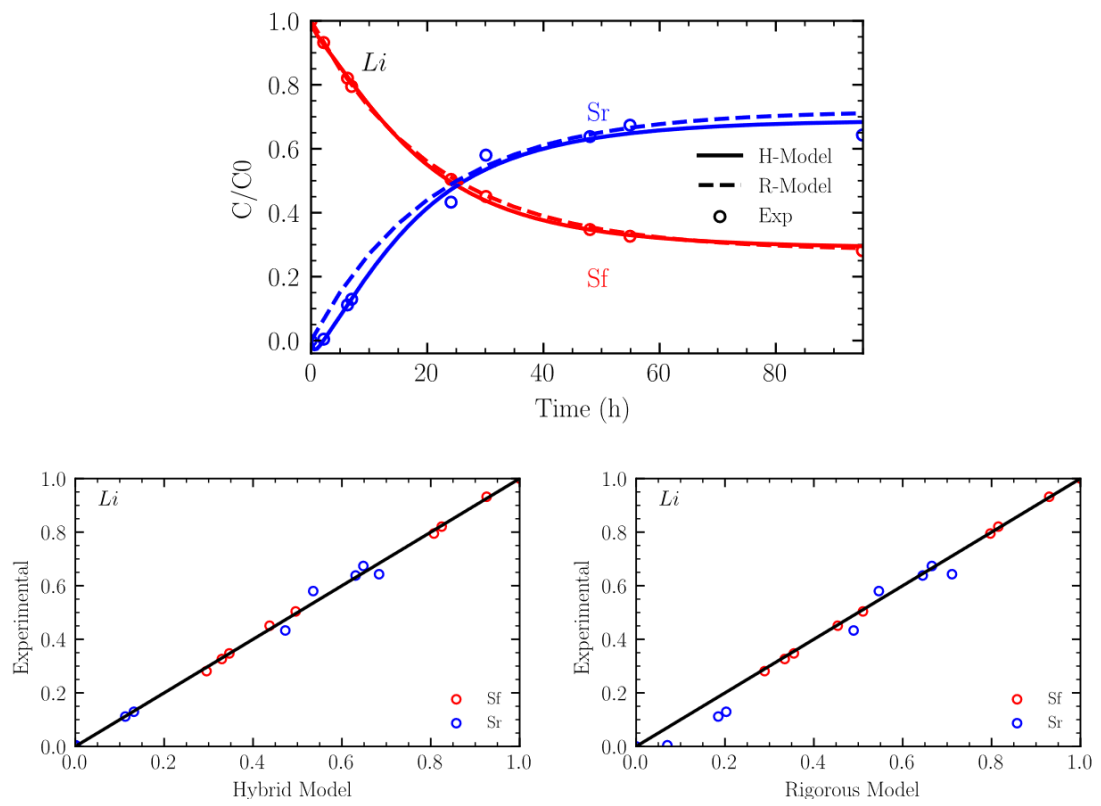


Figure 2.5: Lithium experimental set 2 - Hybrid model prediction (H – Model) versus the phenomenological model prediction (R – Model) and their respective parity graphics.

Still, in Figure 2.5, it is also possible to analyze the model residuals in the parity. It can be observed that the predictions from the hybrid model are randomly distributed around the diagonal line across the whole range of values. On the other hand, the phenomenological model presents a nonrandom pattern within some regions. As before, these points correspond to the transient regime of the process, reinforcing the limitations in the phenomenological model mentioned above.

The same hybrid modeling strategy was applied to other ion separations. This was done to evaluate the method's precision compared with the phenomenological approach in scenarios other than lithium recovery. The same analysis done in the lithium recovery case can be done with similar conclusions using this experimental validation set for different ions.

Figure 2.6, Figure 2.7, and Figure 2.8 present the models' prediction and the parity graphics for a series of 8 experiments. The hybrid model is presented as the H-model, and the phenomenological model is presented as the R-model. It is essential to highlight that the experimental validation sets are another independent system obtained from the literature. As it is possible to see in Figure 2.6, Figure 2.7, the hybrid model presents superior results and significant precision. In Figure 2.8, the hybrid model behavior is very similar to the phenomenological model.

The tendencies observed in the lithium case are also observed in Figure 2.6, Figure 2.7, and Figure 2.8. On the one hand, the proposed hybrid model can precisely describe the systems' dynamic transition. On the other hand, the phenomenological model presents limitations in predicting their dynamics. This demonstrates the potential of this synergy between phenomenological knowledge and artificial intelligence. The AI model adds more flexibility to the description of the system phenomenology, better describing the mass transfer resistance. One can analyze these results from the following perspective. While the solely phenomenological model needs to identify a parameter that can fully describe the mass transfer of the system, the AI tries to understand the mass transfer resistance and store this knowledge in a series of simple mathematical operations.

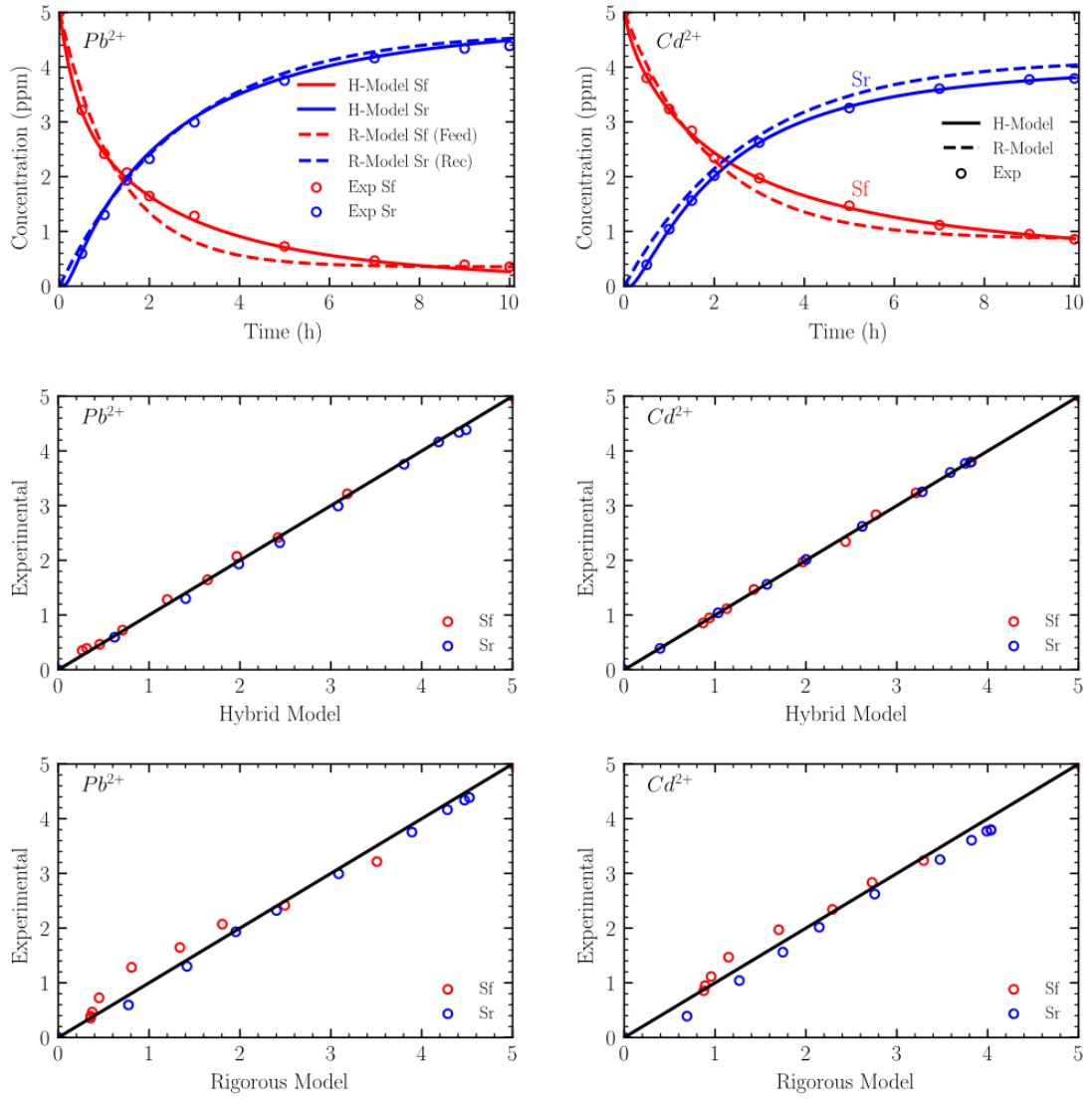


Figure 2.6: Experimental sets 3 and 4 - Hybrid model prediction vs. the phenomenological model prediction and their respective parity graphics.

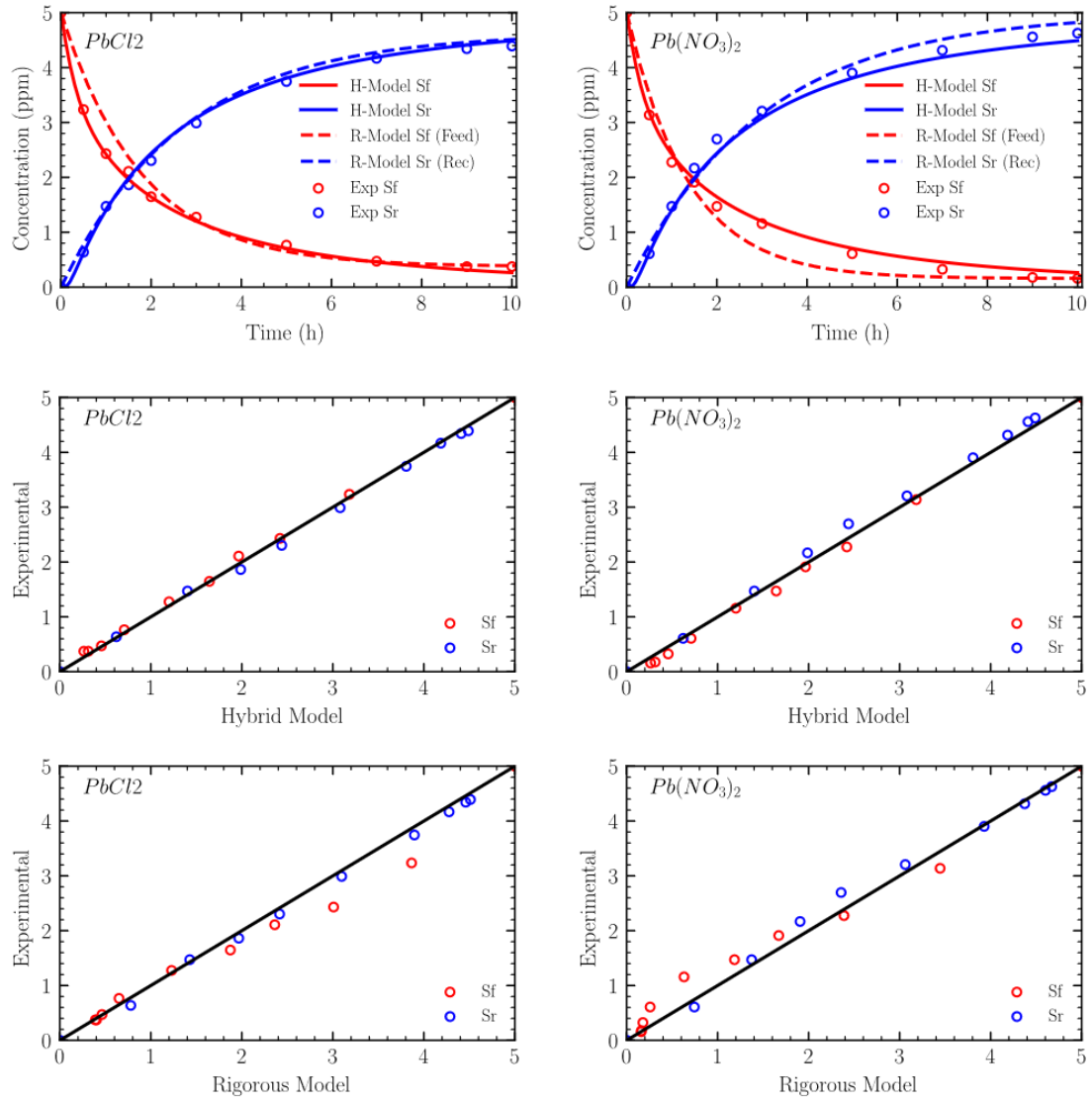


Figure 2.7: Experimental sets 4 and 5 - Hybrid model prediction vs. the phenomenological model prediction and their respective parity graphics

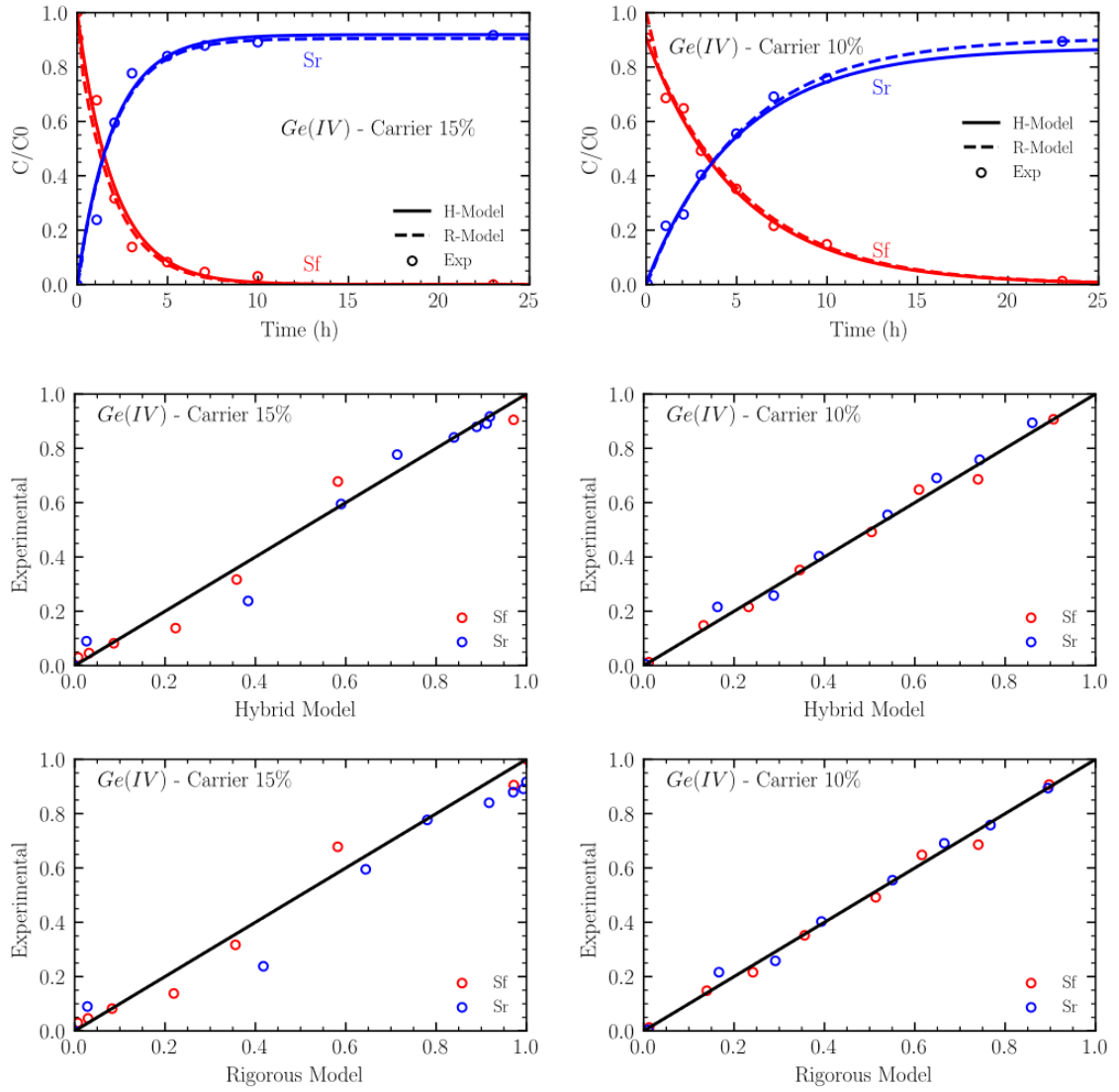


Figure 2.8: Experimental sets 6 and 7 - Hybrid model prediction vs. the phenomenological model prediction and their respective parity graphics.

2.5 Conclusions

This chapter proposed a hybrid modeling framework to build a complete dynamic model representing the flat-sheet supported liquid membrane (FSSLM) for lithium recovery. A phenomenological model was used to describe the system dynamic evolution. In contrast, a surrogate model was employed to represent the mass transport phenomena within the system. This approach was presented as an alternative to the traditional first-principle modeling that embeds the mass transfer phenomena into parameters that need identification.

The main advantage of the proposed framework is that it is possible to extract more information regarding the system and store it within an artificial neural network structure. As the results show, this strategy yields models that are more precise than the phenomenological model alone.

Furthermore, the hybrid strategy can leverage the flexibility of empirical models to understand better the system in the study, for example, the system behavior in possible scenarios of mass transfer resistance. Moreover, it can potentiate the studies regarding manipulating the process component to harness its recovery.

Chapter 3

Using Scientific Machine Learning to Develop Universal Differential Equation for Multicomponent Adsorption Separation Systems

3.1 Abstract

The previous chapter introduced hybrid modeling through a simple system described by a single ordinary differential equation - the first introductory approach in this field in this thesis. Due to limited computer power, hybrid modeling remained limited to small-scale systems for a long time. This chapter expands it to partial differential equations. It formulates a strategy for identifying hybrid models within nonlinear liquid-phase column chromatography, replacing adsorption kinetics and equilibrium with an Artificial Neural Network (ANN).

3.2 Introduction

Using models to describe systems and develop solutions is a longstanding tradition in engineering. These models serve as the foundation for modern engineering practice and are considered valuable assets. The most common approach to modeling physical systems is mathematical models, also known as mechanistic or phenomenological models. These models use established physical laws to accurately describe the behavior of a system within a wide range of conditions. Overall, using models in engineering is a crucial aspect of the field, and the development and utilization of accurate and robust models is an ongoing area of research and development.

In contrast with phenomenological modeling, there is the purely data-driven approach. This alternative has gained attention in the last 20 years due to the increasing amount of

data available and generated. Recent computer science developments have promoted data-driven modeling to a widespread level. Data-driven modeling is the dominant strategy in several fields due to its simplicity and computational cost. The universal approximation theory (Debao, 1993) points out that a large enough Artificial Neural Network model can describe any relationship within a data set. However, the data-driven approach is still auxiliary in fields where mechanistic models have been used for decades.

In chemical engineering, data-driven techniques have been utilized, but their role is primarily supplementary. According to research by Dobbela et al. (2021), the misuse of data-driven techniques in chemical engineering has hindered progress in solving problems within this field. However, it is important to note that the extensive knowledge of the physical systems in chemical engineering makes phenomenological models a reliable and desirable source of information. Additionally, the interpretability of these models is an important factor that has been a concern in the literature. Therefore, to effectively solve chemical engineering problems, it is crucial to balance data-driven techniques with the established knowledge of phenomenological models.

Phenomenological models in chemical engineering are constructed using well-established conservation laws. However, they also require a certain degree of heuristics to describe a specific system accurately. This level of uncertainty is represented by the model parameters and the sink/source term in the mass balance equation. For example, an adsorption model's sink/source term represents a porous matrix's adsorption and desorption phenomena. In this case, the sink/source term is typically defined by a heuristic or empirical model that describes the relationship between the equilibrium adsorbate concentrations in a given phase and the equilibrium adsorption amount in the solid phase, considering factors such as temperature and pressure.

When considering using models in chemical engineering, an important question arises: Why not combine the established knowledge of mechanistic models with the universal approximation capabilities of artificial intelligence (AI) models? By leveraging the strengths of both approaches in a single model, it is possible to utilize the potential of each fully. As previously discussed, phenomenological models rely on sink/source terms that empirical models often describe. These models and parameters are typically identified using experimental data or *ab initio* simulations. When data is available, it is in these situations that Artificial Neural Network (ANN) models are particularly effective.

Concerned with the abovementioned issues, in 1992, Psychogios and Ungar proposed a hybrid framework composed of a mechanistic model and ANN to develop a model of a fed-batch reactor. It is one of the first attempts to bridge these two worlds into a single framework. The ANN is encapsulated within a system of Ordinary Differential Equation (ODE) representing the mechanistic model in the referred work. The solution of the ODE is used to solve the ANN's learning function. Thus, the ODE solution and the ANN's training gradient are correlated, making a complex numerical problem. Perhaps due to the problem's complexity and the limited computing power, for the next 20 years, Psychogios

and Ungar’s work remained one of the most advanced levels of hybridization reported in the literature. Braake et al. (1998) presented a comprehensive study about the application of hybridization in ODE models of chemical processes.

Azevedo et al. (1997) clarified the most challenging issue in the hybrid modeling field, the correlation between the solutions of the mechanistic model and the learning law of the ANN model. The referred work presented an important study of how to address this problem. Unfortunately, computing power at that time was insufficient to provide a definitive solution in this field. However, the potential of hybrid models was demonstrated to enhance empirical models’ extrapolation capacities through mechanistic knowledge while providing interpretable results (Schweidtmann et al., 2021; Kahrs and Marquardt, 2007).

The potential of hybrid modeling is clear. In this sense, the concept of Universal Differential Equations is proposed in Rackauckas et al. (2020). It is inserted into the Scientific Machine Learning (SciML) ecosystem as a tool for mixing the information of physical laws and scientific models with data-driven machine learning approaches. The core idea is replacing unknown parts of a first principle model with a universal approximator, e.g., Artificial Neural Network (ANN). ANNs are universal nonlinear approximators and have been generalized well, even with sparse measurements embedded in hybrid models. Mixing physics knowledge with a data-driven approach can make modeling more flexible and allows the dropping of simplifying assumptions required to derive theoretical models and, ultimately, allow building models with higher predictive capacity using fewer experiments.

The increasing computing power has allowed researchers to revisit hybrid modeling with a different perspective. For instance, Narayanan et al. (2021b,a) could use a hybrid strategy within a system of PDE to describe the single-component adsorption in a fixed bed unit. The hybrid strategy provided notable results in modeling the system. The authors demonstrate that this strategy can overcome the problems of finding suitable mechanistic assumptions for describing the sink/source phenomena. Furthermore, it achieves a better description of the system with minimal experiments.

This study investigates using Universal Differential Equations (UDEs) to solve partial differential equations (PDEs) that describe adsorption separation processes. The strategy is applied to the competitive adsorption of multiple components to develop a hybrid model to describe the system. A series of multicomponent breakthrough experiments were used to identify the model. The study also evaluated the minimum amount of data required to provide a precise model and compared it with a fully mechanistic model reported in the literature for the same system.

3.3 Methodology

3.3.1 Fully mechanistic model

The present work uses a phenomenological model that describes the dynamic behavior of the concentration profile of a fixed bed adsorptive column as a reference case. The suit-

able mechanistic assumptions for the competitive adsorption phenomena of the n-propyl Propionate and Propanol were previously identified by Nogueira et al. (2019). The mechanistic premises are: Isothermal process; Plug flow model with axial dispersion; Negligible radial dispersion; Constant bed porosity and length; External mass transfer to the particle; Internal mass transfer negligible.

In this way, the conservation laws that described the adsorption phenomena within this system can be written as:

$$\frac{\partial c_i}{\partial t} = D_{ax} \frac{\partial c}{\partial z} \left(c_T \frac{\partial y_i}{\partial z} \right) - \frac{\partial (uc_i)}{\partial z} - \frac{1-\varepsilon}{\varepsilon} \frac{3}{r_p} k_{L,i} (q_i^* - q_i) \quad (3.1)$$

$$\frac{\partial q_i}{\partial t} = \frac{3}{r_p} k_{L,i} (q_i^* - q_i) \quad (3.2)$$

where D_{ax} , is the axial dispersion coefficient, ε is the bed porosity, q_i^* is the adsorbed concentration of compound i in equilibrium with c_i , y_i is the molar fraction of the compound i in the liquid phase, u is the interstitial fluid velocity, r_p is the particle radius, $k_{L,i}$ is the external mass transfer coefficient and c_T is the total molar concentration in the liquid phase, which is related to the molar volume $V_{M,i}$ of each compound and can be calculated as:

$$c_T = \frac{1}{\sum_i y_i V_{M,i}} \quad (3.3)$$

The boundary and initial conditions can be written as follows:

$$x = 0 : uc_{i,x=0} - D_{ax} c_T \frac{\partial y_i}{\partial z} = uc_{in} \quad (3.4)$$

$$x = L : c_T \frac{\partial y_i}{\partial z} \Big|_{x=L} = 0 \quad (3.5)$$

$$t = 0 : c_i(z \in (0, L)) = c_{i,0} \quad (3.6)$$

$$t = 0 : q_i(z \in (0, L)) = q^*(c_{i,0}) \quad (3.7)$$

As it is possible to see from Equations 3.1 and 3.2, the model comprises three explicit parameters and a sink/source term describing the adsorption phenomena. It is important to note that the sink/source term of the model is based on premises used to depict the adsorption equilibrium.

In Nogueira et al. 2019 work, Langmuir competitive adsorption was adopted to represent the liquid phase adsorption equilibrium. The Langmuir competitive adsorption model is represented as:

$$q_i^* = \frac{Q_{\max,i} K_i c_i}{1 + \sum_i^{NC} K_i c_i} \quad (3.8)$$

As it is possible to see from Equation 3.8, the model is composed of several parameters, $Q_{\max,i}$ is the saturation adsorption capacity, and K_i is the Langmuir adsorption constant for component i , which need to be estimated. Therefore, a series of experiments are usually

carried out to identify the final structure of the sink/source within the given experimental conditions.

Besides the adsorption equilibrium parameters, the mass transfer coefficient is another parameter that must be defined. In this case, two routes can be taken: experimental estimation or calculation through correlations. The referred work took the second route, making use of Wilson and Geankoplis correlations (Glueckauf and Coates, 1947):

$$Sh_p = \frac{1.09}{\varepsilon} (Re_p Sc)^{0.33} \text{ for } 0.0015 < Re_p < 55 \quad (3.9)$$

Where Sh_p , Re_p , and Sc respectively, the Sherwood, Reynolds, and Schmidt numbers, which in their turn, are computed by:

$$Sh_p = \frac{k_{L,i} d_p}{D_{i,\text{mix}}} \quad (3.10)$$

$$Re_p = \frac{\rho u d_p}{\eta} \quad (3.11)$$

$$Sc = \frac{\eta}{\rho D_{i,\text{mix}}} \quad (3.12)$$

where ρ is the fluid density, η is the fluid viscosity, d_p is the particle diameter $D_{i,\text{mix}}$ is the diffusivity coefficient of compound i in the mixture. The diffusivity coefficient can be calculated by the correlation of Perkins and Geankoplis (Ruthven, 1984).

3.3.2 Universal Differential Equation in column chromatography context

Universal Differential Equation (UDE) is derived from the universal approximator theory. They are proposed as differential equations defined in full or partly by a universal approximator (a parameterized function capable of representing any possible functions in some parameter size limit) (Hornik et al., 1989). Known approximators include piecewise polynomial functions, Fourier expansions, and Gaussian processes. On the other hand, common universal approximators in high dimensions include Artificial Neural Network (ANN).

Figure 3.1 illustrates the concept of a UDE with ANNs as a universal approximator. The differential equation comprises two parts – the part with known physics, i.e., with prior structural knowledge, and the part with unknown physics, which an ANN replaces. The ANN parameters can then be estimated using data, e.g., experiments, a loss function, and a suitable optimizer.

As presented in the last section, modeling the multicomponent equilibrium is essential when designing a chromatographic separation unit. The traditional approach usually involves fitting experimental data to single-component isotherms (Langmuir, Sips, BET). Then, multicomponent equilibrium is generally predicted using heuristic rules (competitive Langmuir) or consistent thermodynamic formulas such as Ideal Adsorption Solution Theory (IAST) for ideal solutions.

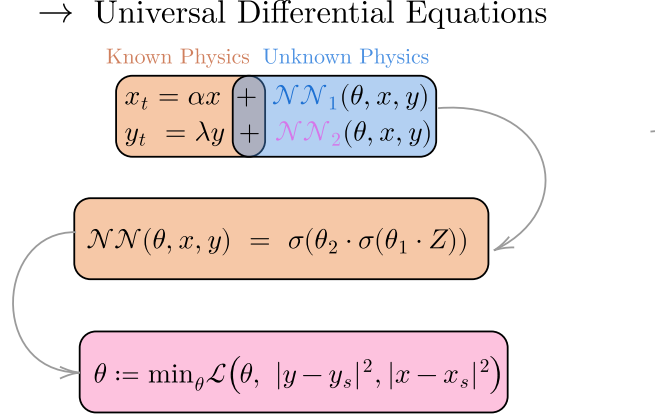


Figure 3.1: Illustration of UDE building and fitting procedure

These are just a few possibilities in many closed-form formulas for predicting multi-component adsorption equilibrium. Therefore, UDE suits well in this context if replacing closed-form (and somehow limited) multicomponent isotherms by ANNs. The models, in turn, can be trained directly from breakthrough experiments. Furthermore, the ANN can deeply extract the relationship the data represents, learning the adsorption equilibrium without being limited by a simplified mathematical form.

In this way, a UDE is proposed for the fixed bed model, and both mass transfer and adsorption equilibrium are replaced by ANNs similar to Narayanan et al.'s works. Therefore, Equations 3.1, 3.2, and 3.8 are rewritten as:

$$\frac{\partial c_i}{\partial t} = D_{\text{ax}} \frac{\partial c}{\partial z} \left(c_T \frac{\partial y_i}{\partial z} \right) - \frac{\partial (uc_i)}{\partial z} - \frac{1 - \varepsilon}{\varepsilon} \frac{3}{r_p} \text{ANN}_{\text{mass}}(\theta_{\text{mass}}, c_{1:N})(q_i^* - q_i) \quad (3.13)$$

$$\frac{\partial q_i}{\partial t} = \frac{3}{r_p} \text{ANN}_{\text{mass}}(\theta_{\text{mass}}, c_{1:N})(q_i^* - q_i) \quad (3.14)$$

$$q_i^* = \text{ANN}_{\text{eq}}(\theta_{\text{eq}}, c_{1:N}) \quad (3.15)$$

Therefore eliminating the need for the correlations described in equations 3.13 to 3.15. The remaining equations are kept unchanged. The parameters $(\theta_{\text{mass}}, \theta_{\text{eq}})$ are fitted by minimizing an appropriate error function proportional to the difference between experimental breakthrough data and predictions derived from the UDE solution.

3.3.3 Simulation

The proposed UDE was implemented in Julia Programming Language. The partial differential equations were transformed into a system of ODEs using the method of lines. The spatial domain was discretized using orthogonal collocation on finite elements with cubic Hermite polynomials as basis functions. The resulting ODEs was fused with the

ANN using the DiffEqFlux.jl library (Rackauckas et al., 2020), which counts with highly optimized implementations of gradient-based optimizers. A total of 20 elements (40 internal collocation points) were used. A Julia Native implementation of the third-order stiffly stable Rosenbrock (ROS3P) method was used to solve the resulting ODE. It is important to highlight that Julia presents a significant advantage for this work. Even with random initialization of weights, Julia’s native implementation of the third-order stiffly stable Rosenbrock method can solve the discretized differential-algebraic equations in milliseconds and accurately calculate gradients with forward-mode automatic differentiation.

3.4 Results

The proposed Universal Differential Equation (UDE) was implemented to solve the case study previously described. The ANNs replace the multicomponent adsorption of the non-reactive pair Propanol (POH) and n-Propyl Propionate (ProPro). Table 3.1 presents the feed stream’s time-varying concentration and the solid phase concentration in equilibrium with the feed concentration (calculated from breakthrough curves as presented in Nogueira et al. (2019)).

Table 3.1: Initial concentrations, c_0 , and feed concentration, c_{in} , for each step performed and solid concentration, q^* , in equilibrium with the feed concentration in the breakthrough experiments.

	1-Propanol – n-Propyl Propionate
c_0 (mol L ⁻¹)	13.23 - 0.00
c_{in} (mol L ⁻¹)	11.64 - 0.95
q^* (mol L ⁻¹)	8.05 - 0.55
c_0 (mol L ⁻¹)	11.64 - 0.95
c_{in} (mol L ⁻¹)	8.80 - 2.53
q^* (mol L ⁻¹)	7.97 - 0.59
c_0 (mol L ⁻¹)	8.80 - 2.53
c_{in} (mol L ⁻¹)	6.23 - 3.95
q^* (mol L ⁻¹)	7.26 - 0.97
c_0 (mol L ⁻¹)	6.23 - 3.95
c_{in} (mol L ⁻¹)	3.67 - 5.38
q^* (mol L ⁻¹)	5.80 - 1.77
c_0 (mol L ⁻¹)	3.67 - 5.38
c_{in} (mol L ⁻¹)	1.33 - 6.67
q^* (mol L ⁻¹)	3.60 - 2.99
c_0 (mol L ⁻¹)	1.33 - 6.67
c_{in} (mol L ⁻¹)	13.23 - 0.07
q^* (mol L ⁻¹)	8.10 - 0.48

Table 3.2 presents the fixed bed operating conditions and parameter values – geometrical (bed size L , area, A_c , and particle radius, r_p), and hydrodynamic (Peclet number, Pe ,

porosity, ε), feed flow rate (Q_f) and temperature (T).

Table 3.2: Fixed bed parameters and feed flow rate.

Parameters	
ε	4×10^{-1}
Pe	166
r_p (μm)	122.75
L (dm)	1.15
A_c (dm^2)	4.15×10^{-2}
T (K)	313
Q_f ($\text{mL} \cdot \text{min}^{-1}$)	7.88

We propose the traditional feed-forward neural network to replace the adsorption equilibrium and mass transfer coefficient. For adsorption equilibrium, each component is represented by an ANN composed of one dense layer with 7 neurons with hyperbolic tangent activation and an output layer with 1 neuron (output size) - 29 parameters for each.

A similar architecture is used for modeling the mass transfer coefficient, i.e., one neural network per component with one dense layer with 7 neurons. The only difference is the output layer, where a rescaled sigmoid activation was used as a hard constraint, so the outputs are bounded to be between 0 and 3. Overall, there were 116 parameters to fit. This architecture was chosen considering a tradeoff between model complexity and computational time – training the neural network involves solving a system of 168 stiff differential equations and backpropagating the gradient of cost with respect to the ANN weights. At the same time, the architecture cannot be too simple or fails to model the equilibrium behavior. One of the advantages of UDE is that the neural network sizes do not need to be too large as we bias the outputs toward ensuring physical consistency with differential equations.

To train the ANN, choosing the right amount of data is crucial. The choice must meet a tradeoff between a good fitting and a generalization error. In this work, 240 points derived from 6 binary breakthrough experiments were available– fitting the ANN with 5 breakthrough curves (train set) and assessing its generalization capacity on the remaining 1 (test set).

The ANN inputs were rescaled with the concentration ranges as follows:

$$X = \frac{c_i - c_{i,\min}}{c_{i,\max} - c_{i,\min}} \quad (3.16)$$

Where $c_{i,\min}$ and $c_{i,\max}$ are the minimum and maximum concentrations from the training set.

The cost function to be minimized is the Mean Squared Error (MSE) given by:

$$\mathcal{L}(\theta, c^{\text{exp}}, q^{*,\text{exp}}) = \sum_{i \in N_{\text{train}, c}} \frac{(c_i^{\text{exp}} - c_i^{\text{UDE}}(x = L, t = t_{\text{exp}}, \theta))^2}{N_{\text{train}}} + \sum_{i \in N_{\text{train}, q}} \frac{(q_i^{*,\text{exp}} - q_i^{\text{UDE}}(x = L, t = t_{\text{final}}, \theta))^2}{N_{\text{train}, q}} \quad (3.17)$$

The cost function aims to minimize the square difference between predicted and experimental concentration in the liquid phase at all experimental time steps t_{exp} , at the column outlet, L, summed with the prediction error in the solid phase concentration at all discretized axial positions and terminal experimental times (times where the system reached equilibrium). The proposed loss function was minimized using the Adaptive moment estimation algorithm (Kingma and Ba, 2015) with linear learning rate decay starting from 0.01 to 0.008 over 320 iterations. Figure 3.2 displays the training loss history over the iterations.

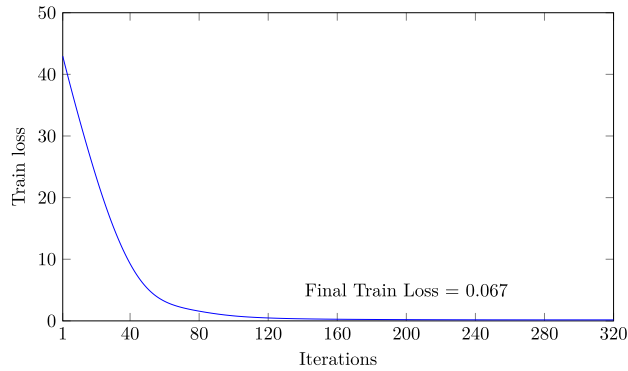


Figure 3.2: Training loss history.

Figure 3.3 and Figure 3.4 show the results for the fitted Artificial Neural Network (ANN) with five breakthrough experiments and the comparison with the fully-mechanistic model developed in Nogueira et al.'s work. As can be seen, the five breakthrough curves used for training show a good fitting for the training set and a good generalization error on the test set for both components. A mean squared error of $0.067(\text{mol/L})^2$ was obtained with the UDE for both components in the training set. Compared with the entirely mechanistic (phenomenological) model, the UDE presents a better generalization capacity, as seen clearly in the test set for ProPro.

To better visualize the predictive capacity of UDE and its comparison with the fully mechanistic model, several plots of the prediction error (predictions subtracted from experimental values) were used. Figure 3.5 and Figure 3.6 display the prediction error over time. In the training set, smaller errors are observed in the UDE in the third and fourth breakthrough curves.

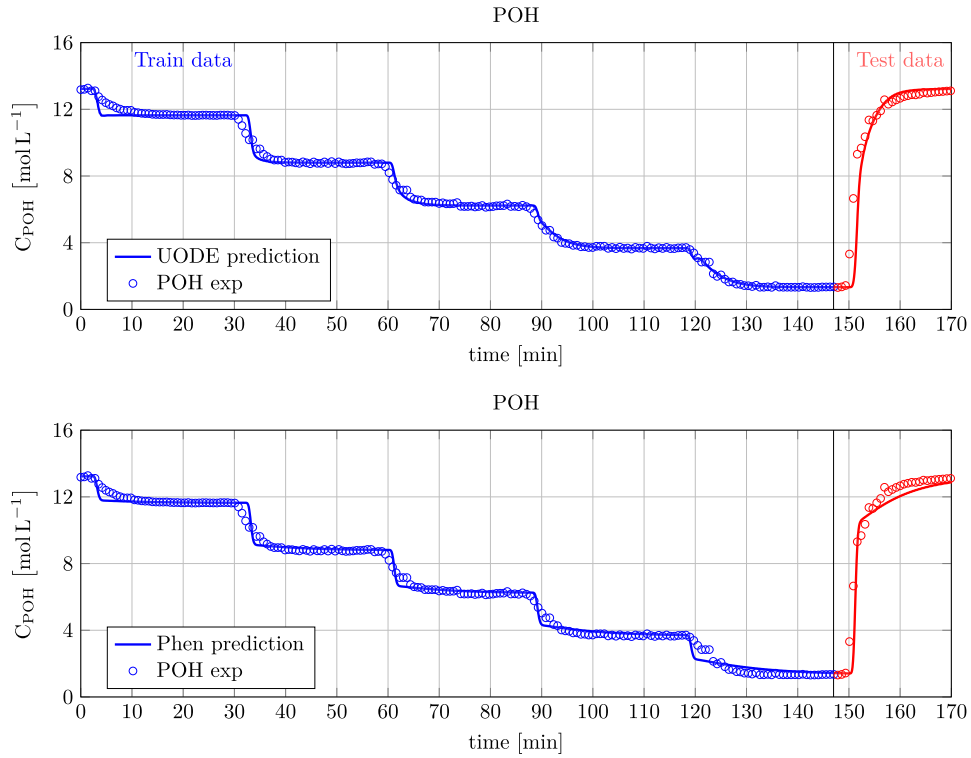


Figure 3.3: Train and test breakthrough curves for POH – UDE vs Phenomenological comparison.

Figure 3.7 displays another error analysis via parity plots. It can be seen that most UDE predictions lie near the diagonal line (zero error), and similar behavior is observed for the fully-mechanistic approach.

An additional comparison of the fully mechanistic approach and UDE simulation can be made by analyzing the concentration profiles during the change in the feed concentration. Figure 3.8 displays the profile snapshots during the second breakthrough curve for both components and models. It can be seen that for POH, the profiles are almost identical for both models. However, a pronounced difference is observed for ProPro, where UDE predicts less dispersive moving fronts.

Finally, the ANNEQ identified is tested in the reconstitution of the multicomponent competitive isotherms of the non-reactive pair ProPro and Propanol. This is a final test to evaluate the physical interpretability of the identified model. Figure 3.9 presents the ANN competitive adsorption isotherm plotted against the experimental data. The loading corresponds to a pair of concentrations ranging from almost 100% pure ProPro to 100% POH of the multicomponent isotherm.

As it is possible to see in Figure 3.9, the ANN isotherm can predict well the experimental points. Therefore, the obtained model can provide the predictions of the multicomponent isotherm with high precision. This shows the potential of UDE – introducing a robust inductive bias via conservation laws leads to minor generalization errors.

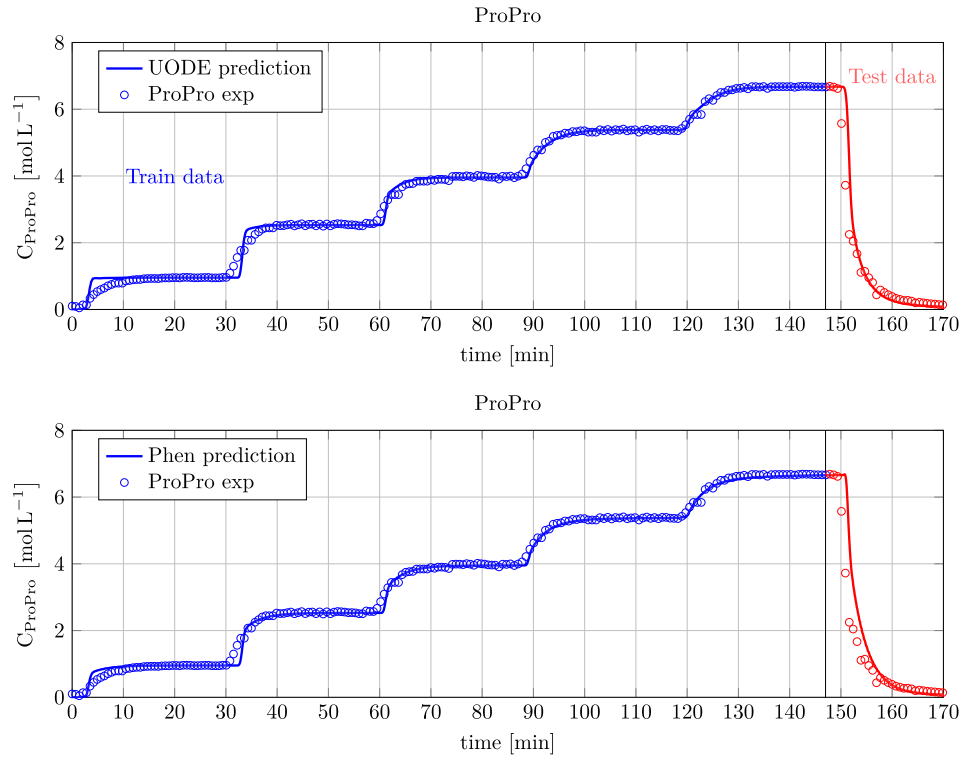


Figure 3.4: Train and test breakthrough curves for ProPro – UDE vs Phenomenological comparison.

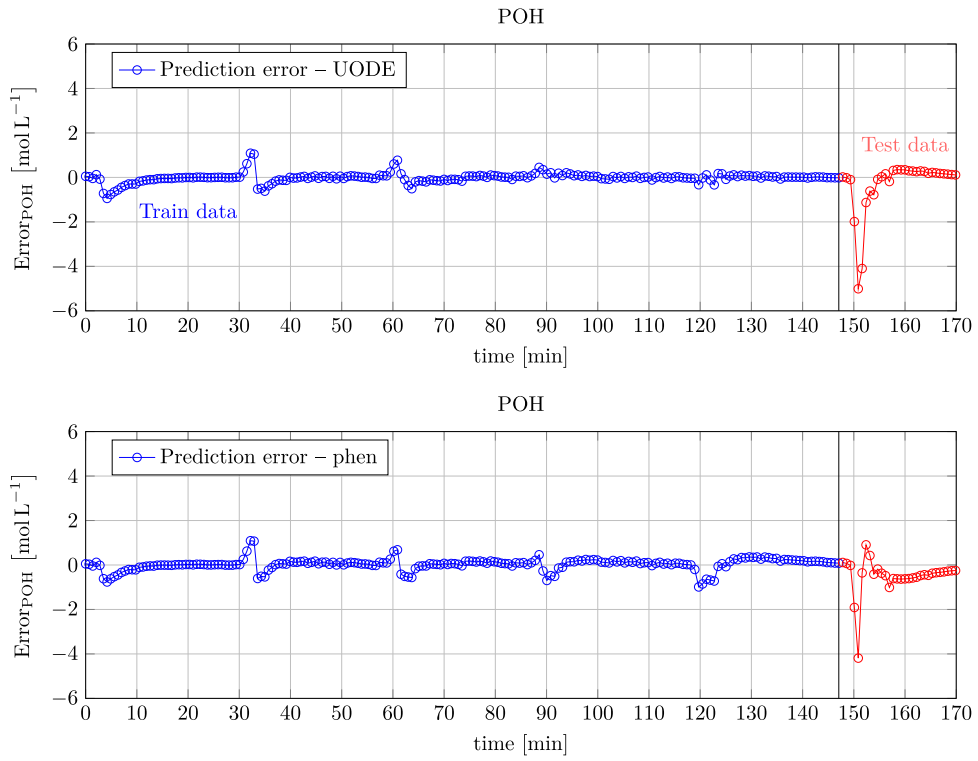


Figure 3.5: Train and test breakthrough curves' error for POH – UDE vs Phenomenological comparison.

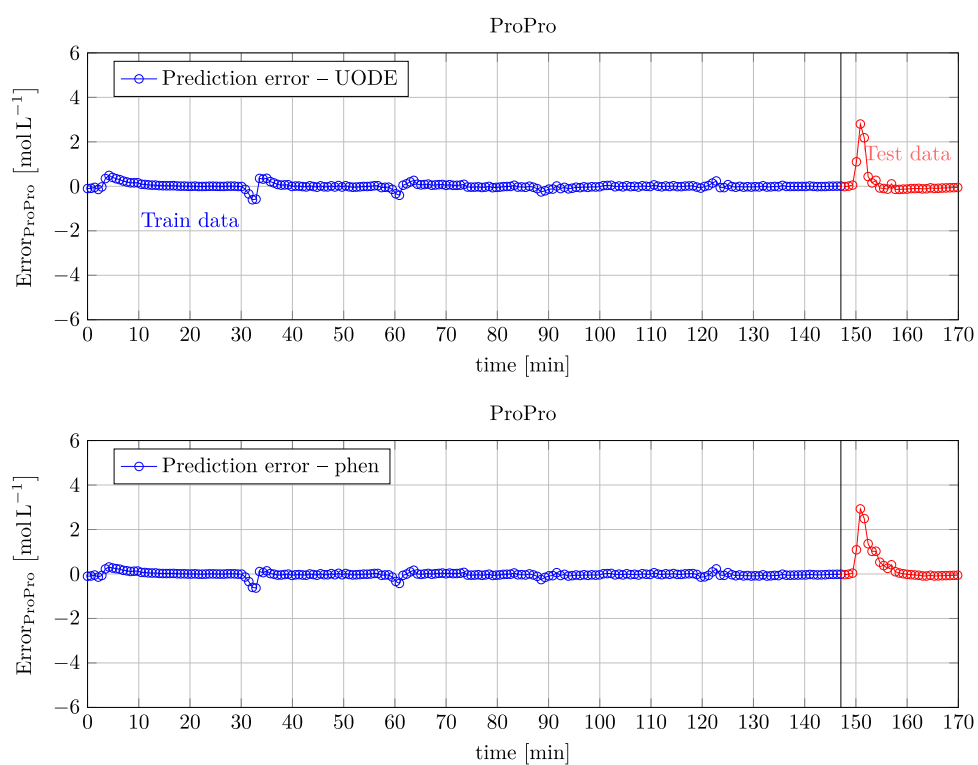


Figure 3.6: Train and test breakthrough curves for ProPro – UDE vs Phenomenological comparison.

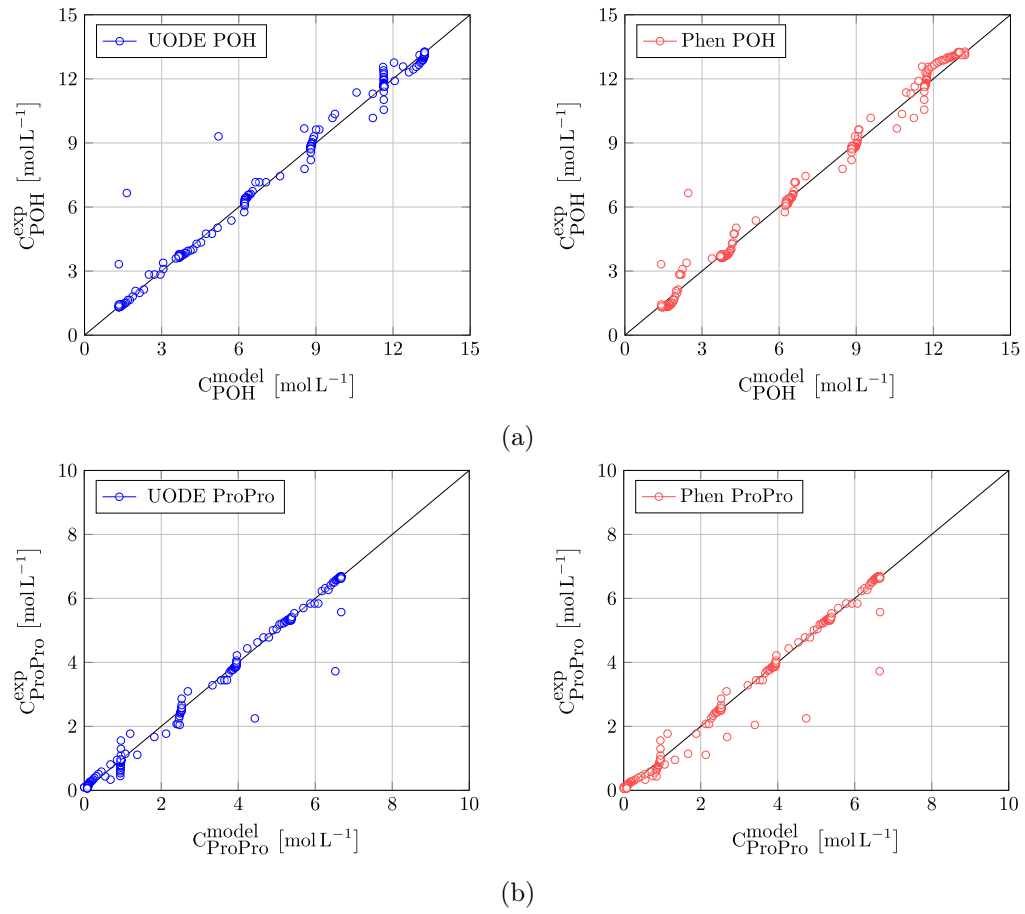


Figure 3.7: Parity plots for breakthrough curves' train and test error– UDE vs Phenomenological comparison (a) Propanol (POH). (b) n-Propyl Propionate (ProPro).

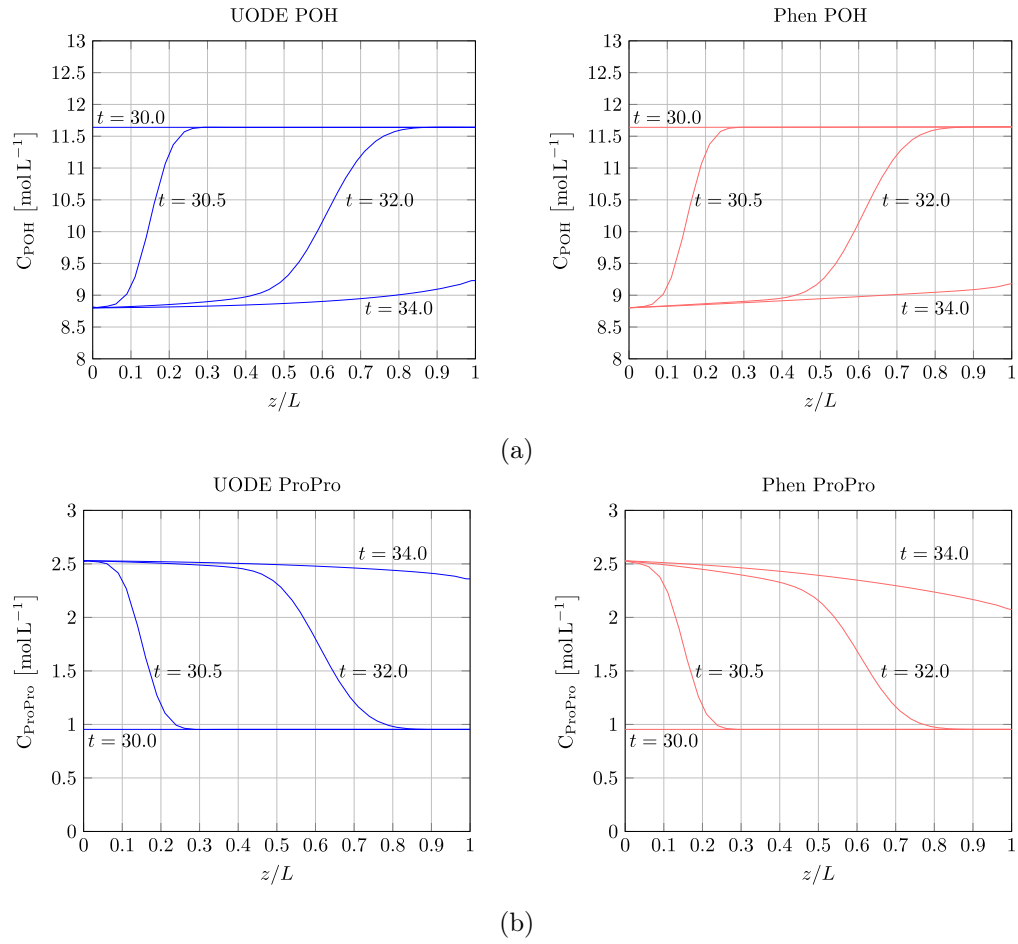


Figure 3.8: Comparison of UDE and fully-mechanistic simulation profile for the second breakthrough curve. (a) Propanol (POH). (b) n-Propyl Propionate (ProPro).

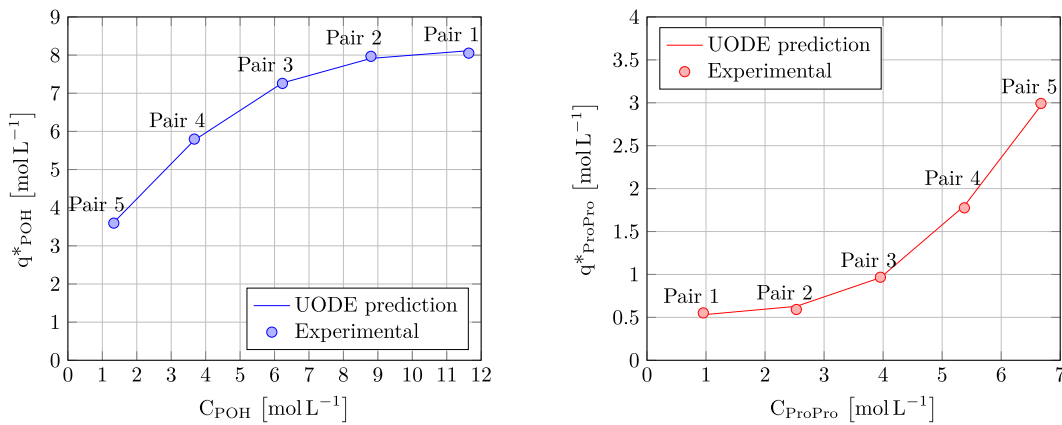


Figure 3.9: ANNEQ competitive adsorption isotherm for the binary 1-Propanol and n-Propyl Propionate and experimental data over Amberlyst-46 at 313 K.

3.5 Conclusions

This work presented the first approach towards developing Universal Differential Equations to solve multicomponent competitive adsorption systems. A fixed bed adsorption unit was modeled through the UDE approach and validated against experimental data. The developed methodology solved the system of stiff ordinary differential equations that represent the physical laws involved in the processes. At the same time, the ANN model was able to learn from experimental data. The developed model provided a precise prediction of the experimental data, providing lower prediction errors when compared with the fully-mechanistic model. Further tests demonstrated that the identified ANNs can predict well the competitive multicomponent isotherm with higher precision than the Langmuir model.

This is the first step in modeling adsorption separation in a multicomponent scenario. Furthermore, studies are required to leverage the full potential of the Universal Ordinary Differential Equations. However, the results are promising and might point to the future widespread use of this strategy in the adsorption field and several other chemical processes.

Chapter 4

Efficient hybrid modeling and sorption model discovery for non-linear advection-diffusion-sorption systems: A systematic scientific machine learning approach

4.1 Abstract

While gradient-based optimizers have been effectively applied in ANN-based hybrid advection-diffusion-sorption PDE models outlined in the previous chapter (chapter 3), they come with their challenges. Specifically, accurately and efficiently calculating the gradient poses a bottleneck. The rudimentary numerical differentiation method using finite differences suffers from roundoff errors that can significantly impair the performance of optimizers. As an alternative, Automatic Differentiation (AD) promises high-accuracy gradients but often at the expense of increased computational load and potential numerical instability, especially in non-linear advection-diffusion-sorption PDE contexts. This chapter extends general guidelines for employing AD in stiff models originating from PDE discretizations and navigates the trade-offs between stability and efficiency. It introduces a novel set of guidelines specifically designed for non-linear advection-diffusion-sorption PDEs, marking a significant stride in hybrid modeling for sorption processes.

4.2 Introduction

Mathematical modeling has been an essential part of science and engineering since its early stages. The use of mathematical models for problem-solving in chemical engineering has become increasingly important due to the exponential growth of computer power. Two main approaches have been used to model systems in chemical engineering: The mechanistic/classical and the empirical approaches (Rasmuson et al., 2014). The classical approach encompasses conservation equations, transport, and thermodynamic expressions and can accurately describe the problem. However, simplifications are often employed to make the problem computationally tractable. The empirical/data-driven models do not make any assumptions about the physics behind the observations, making them flexible. Nevertheless, they are suitable for use in a limited range of process conditions due to their limited extrapolation power (Sansana et al., 2021). In certain scenarios, simpler models might offer better performance, particularly when paired with representative training data obtained through a systematic Design of Experiments (DOE) approach. It is important to note that while they can potentially enhance performance, the hypotheses they bring to the table may increase associated uncertainties and degrade the model’s predictive capacity. Our proposed strategy frees the model from these simplifications and hypotheses, allowing the network to learn from both data and physics.

On the other hand, empirical models have gained substantial attention in the last few years with machine learning and deep learning. Simultaneously, it is becoming more accepted that mechanistic model errors caused by simplifications can be compensated with other forms of knowledge, e.g., empirical models (Azevedo et al., 1997). In this sense, hybrid models have received considerable attention in engineering (von Rueden et al., 2020; Pan et al., 2022). The “hybrid” term refers to a structure where empirical equations and physics knowledge (conservation laws, operation invariance, etc.) are merged. The merging process for engineering applications can take several forms, according to von Stosch et al. (2014). Figure 4.1 illustrates the types of hybrid models according to the authors. The black box represents data-driven models/universal approximations, and the white box mechanistic models are derived from the first principles. The authors divided hybrid models into 3 categories: parallel, serial A and B. Parallel models are suitable when the model structure is not sufficiently known. An empirical model compensates for non-explained prediction residuals. When the model structure is well known, serial A replaces a missing model term with a universal approximator and is expected to perform significantly better than the parallel, especially in extrapolation. Serial B uses mechanistic model predictions as inputs of an empirical model. However, it has found very few applications in chemical engineering. Sansana et al. (2021) have extended this classification and include surrogates as a separate class. Another noteworthy class of hybrid model is the approach called Physics-Informed Neural Networks (Raissi et al., 2019). It is conceptually distant from all described hybrid models as it uses physics to regularize Artificial Neural Network (ANNs)

training.

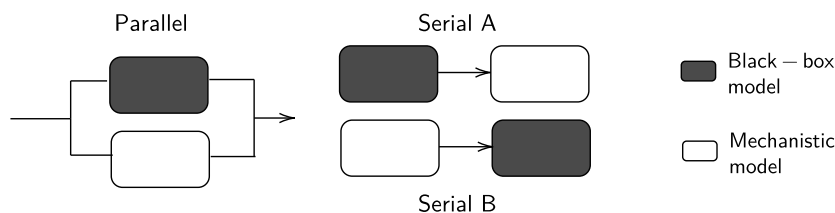


Figure 4.1: Hybrid models types - adapted from von Stosch et al. (2014)

Hybrid serial A models in chemical engineering may combine first-principles derived equations with a universal approximator to replace one or more terms of the equation - kinetics, transport, fluxes. When the underlying physics can be described by differential equations (ordinary, algebraic, or partial), the resulting mathematical object is named Universal Differential Equation (UDE) as described in Rackauckas et al. (2020). A universal approximator is a parameterized object representing any possible function in some parameter size limit (Rackauckas et al., 2020). In this sense, Artificial Neural Networks (ANNs) are particularly interesting as they can accommodate high-dimensional input space functions and be made differentiable with appropriate activation functions.

The most common applications of hybrid serial A models lie in bio(reaction) engineering, where the kinetic reaction law is replaced by an ANN (Azevedo et al., 1997; Zander et al., 1999). However, fitting ANNs' parameters as part of a differential equation poses a significant computational challenge (Azevedo et al., 1997). Therefore, applications have been limited to simple systems with few equations and poorly explored for more complicated problems, e.g., distributed-parameter representations such as fixed-bed reactors and chromatographic columns. Recently, Narayanan et al. (2021b,a) pioneered the field using hybrid serial A models in nonlinear chromatography where an ANN in a hybrid advection-diffusion-sorption PDE was trained. These works demonstrated the superior capacity of the hybrid model for predicting breakthrough observations and uncovering adsorption equilibrium information. However, despite being able to train the neural networks with Bayesian optimization, it is well known that derivative-free methods such as Bayesian optimization perform poorly in high-dimensional parameter spaces, e.g., artificial neural networks (Moriconi et al., 2019; Frazier, 2018).

In such scenarios, gradient-based or hessian-based optimizers are more appropriate and preferred. In fact, the state-of-the-art methods for training neural networks are based on gradients (Kingma and Ba, 2015). In this sense, the chapter 3 delved into the application of forward-mode automatic differentiation for training an ANN in a hybrid 1D advection-diffusion-sorption PDE. The work demonstrated that it was possible to identify a hybrid model with discrete events that outperformed the purely mechanistic approach. Another pioneering work in adsorption modeling is presented by Praditia et al. (2021), where a

Finite volume Neural Network (FINN) is presented for modeling subsurface contaminant transport as a non-linear diffusion-sorption process – fitting as a hybrid model by capitalizing finite volume structural knowledge and the flexibility of ANNs.

Despite being able to use gradient-based optimizers in Nogueira et al. (2022), several issues are still open in ANN-based hybrid advection-diffusion-sorption PDE problems. One key problem with using gradient-based optimizers is the ability to calculate the gradient itself accurately and efficiently. The simplest approach is numerical differentiation with finite differences. However, roundoff errors can become prohibitively high and degrade optimizers’ performance. The preferred approach is Automatic Differentiation (AD), where high-accuracy gradients can be obtained. On the other hand, they can be computationally demanding and numerically unstable for non-linear advection-diffusion-sorption PDE problems (Ma et al., 2021). A general guideline for choosing AD for stiff models resulting from PDE discretization and efficiency/stability trade-offs is presented by Ma et al. (2021). Nonetheless, stability and efficiency are system dependent and should be assessed for each problem. Here, a guideline for non-linear advection-diffusion-sorption PDE problems is provided - an important contribution for hybrid modeling in sorption processes.

In this sense, the present work demonstrates a feasible and efficient way for training hybrid non-linear advection-diffusion-sorption PDE problems using gradient-based optimizers via continuous adjoint sensitivity analysis using quadrature adjoint with JIT-compiled tape-based vector jacobian product, orthogonal collocation on finite element PDE discretization scheme, and a fixed-leading coefficient adaptive-order adaptive-time BDF method ODE integrator. As far as we are concerned, it is the first work assessing the numerical aspects of hybrid non-linear advection-diffusion-sorption PDE problems. Once the stability of the hybrid model identification approach is demonstrated, another open issue in the literature is finding ways to make ANNs’ predictions interpretable, i.e., identifying the typical functional forms that are expected within domains. One way to achieve it is by using symbolic or sparse regression on the ANNs’ output (Rackauckas et al., 2020). This approach reduces computational burden by avoiding solving differential equations during function space search (as in traditional symbolic regression involving differential equations). It allows finding a combination of simple functions that share similar properties with the trained ANN. Several studies employed either sparse or symbolic regression on chemical engineering (McKay et al., 1997; Cozad and Sahinidis, 2018; Narayanan et al., 2022). Still, as far as we know, none have used them for giving interpretability to trained hybrid non-linear advection-diffusion-sorption PDEs.

To demonstrate the contributions above, an in-silico data set is used. The data set was built by simulating breakthrough curves of a hypothetical single-component non-linear advection-diffusion-sorption system with two different isotherms (Langmuir and Sips) and three kinetic models – Linear Driving Force (LDF), Vermeulen’s model (Hai, 2004), and improved LDF (Li et al., 2018b). The present work can help engineers to efficiently introduce hybrid modelling in packed-bed separation for either improving predictive power or

discovering mass-transfer kinetics direct from breakthrough curve data.

The present work is organized as follows. section 4.3 presents the methodology, including the *in-silico* data set building process, hybrid model proposition, numerical aspects of the hybrid model solution and gradient calculation, and sparse regression details. section 4.4 presents the results for the universal differential equation training and test sets performance and sparse regression on the trained ANNs.

4.3 Methodology

4.3.1 Data set

An in-silico data set was created to demonstrate the aforementioned objectives by simulating column chromatography breakthrough curves (5.5 mg/L concentration). It was considered an additive Gaussian noise parametrized with a standard deviation compatible with real experimental data (5%) as referred to in the literature (Nogueira et al., 2022; Narayanan et al., 2021b,a). Two types of isotherms – Langmuir and Sips, three adsorption kinetic models – Linear Driving Force (LDF), quadratic, and improved LDF from Li et al. (2018b) and one sampling rate – 0.5 min^{-1} were investigated. The objective is to show the robustness of the proposed method to accommodate several kinetic models and isotherms. The reason for using 5.5 mg/L in the feed concentration was to cover a considerable part of the adsorption isotherm in the breakthrough curve, considering the case study as detailed below. The model equations and kinetic laws are in the Appendix A.

The hybrid model makes no assumptions about the adsorption kinetics but considers that the adsorption isotherm is known and is an input of the ANN. Moreover, all remaining parameters are considered known – Peclet number and porosity. This approach differs from Nogueira et al. (2022); Narayanan et al. (2021b,a) work where isotherms were unknown and fitted with breakthrough data. Here it is tested if the ANN can learn adsorption uptake from single noisy breakthroughs at column outlet and extrapolate in adsorption and desorption well.

Only the data at the column outlet was used as the training set, i.e., for fitting ANN's parameters. The reason is that data collection at positions aside from column outlets is not a usual practice in the field. Additionally, two test sets were created with simulation to assess the generalization capacity of the model, i.e., whether the model is general enough to extrapolate adsorption and to predict desorption correctly.

In the in-silico training data set, parameters such as $v = 5.1 \times 10^{-1} \text{ (dm/min)}$, $L = 2.0 \text{ (dm)}$, $Pe = 21.0$, and $\varepsilon = 0.5$ were used. Tests were conducted under six different configurations, combining two types of isotherms - Langmuir and Sips, with three different kinetics - LDF, Vermeulen's, and Improved LDF. For all configurations, the input concentration (c_{in}) was kept at 5.5 mg/L, observation time (t_{obs}) was set to 110 min, and the sampling frequency (f_{sample}) was maintained at 0.5 min^{-1} .

Table 4.1 shows the test set process conditions. In the test set, inlet concentrations are changed from 5.5 mg/L to 3.58 mg/L to simulate desorption, from 3.58 mg/L to 7.33 mg/L to simulate extrapolated adsorption for Langmuir isotherm, and changed from 5.5 mg/L to 0.75 mg/L to simulate desorption and from 0.75 mg/L to 9.33 mg/L to simulate extrapolated adsorption for Sips isotherm. The introduction of different steps in the inlet concentration to each isotherm aims to examine the proposed approach thoroughly. This strategy is designed to underscore the versatility and robustness of the method under a diverse set of conditions ranging from near zero to high inlet concentrations.

Table 4.1: Test *in-silico* data set

Parameters - $v = 5.1 \times 10^{-1}$ (dm/min), $L = 2.0$ (dm), $Pe = 21.0$, $\varepsilon = 0.5$				
Isotherm	Kinetic	c_{in} (mg/L)	t_{obs} (min)	f_{sample} (min ⁻¹)
Langmuir	LDF	5.5 $\rightarrow$ 3.58 $\rightarrow$ 7.33	270	0.5
Langmuir	Vermeulen's	5.5 $\rightarrow$ 3.58 $\rightarrow$ 7.33	270	0.5
Langmuir	Improved LDF	5.5 $\rightarrow$ 3.58 $\rightarrow$ 7.33	270	0.5
Sips	LDF	5.5 $\rightarrow$ 0.75 $\rightarrow$ 9.33	270	0.5
Sips	Vermeulen's	5.5 $\rightarrow$ 0.75 $\rightarrow$ 9.33	270	0.5
Sips	Improved LDF	5.5 $\rightarrow$ 0.75 $\rightarrow$ 9.33	270	0.5

The column geometrical and hydrodynamical parameters were chosen based on Li et al. (2018b) work. Even though the velocity affects the mass transfer coefficient for some situations where external mass transfer is dominant, this work did not aim to assess all specific model parts where outputs can vary. The main goal is to show that the proposed approach is a stable, reproducible, and efficient way of fitting hybrid models in non-linear advection-diffusion-sorption PDE problems. Extending the work to accommodate mass transfer flow-rate dependency is straightforward. It only requires including this variable at the input of the ANN and collecting data at several flow rates. All the methodology presented here still applies with such modifications.

4.3.2 Hybrid model

The proposed hybrid model inherits most of the structure of the traditional "linear driving force" mechanistic fixed bed model in chromatography (Minceva et al., 2015). It encompasses unsteady-state mass transfer between the fluid and solid phases. The model can be derived from a time-dependent 1-dimensional conservation equation in cylindrical geometry. It considers a plug flow, negligible pressure drop, isothermal operation, and uniform radial distribution with no variation in the superficial flow velocity throughout the bed. The hybridization occurs at the material balance that models mass transfer between solid and fluid phases. Thus, it can be written as:

$$\frac{\partial c}{\partial t^*} = -\frac{1-\varepsilon}{\varepsilon} \text{ANN}(q, q^*, \theta) - \frac{\partial c}{\partial x^*} + \frac{1}{Pe} \frac{\partial^2 c}{\partial x^{*2}}, \quad (4.1)$$

$$\frac{\partial q}{\partial t^*} = \text{ANN}(q, q^*, \theta), \quad (4.2)$$

$$\frac{\partial c(x^* = 1, \forall t)}{\partial x^*} = 0, \quad (4.3)$$

$$\frac{\partial c(x^* = 0, \forall t)}{\partial x^*} = Pe(c - c_{inlet}), \quad (4.4)$$

$$c(x^* \in (0, 1), t^* = 0) = c_0, \quad (4.5)$$

$$q(x^* \in (0, 1), t^* = 0) = q^*(c_0), \quad (4.6)$$

$$q^* = f(c, p), \quad (4.7)$$

in which, $t^* = tv/L$ is the dimensionless time, $x^* = x/L$ is the fixed bed dimensionless length, the axial distance from column inlet, c is the liquid phase concentration, q is the solid phase concentration, $q^*(c, p)$ is the solid phase equilibrium concentration (calculated from a known isotherm f with known parameters p), ANN is the artificial neural network parametrized by θ , ε is the bed porosity, v is the interstitial velocity, Pe is the Peclet number. In this work, $q^*(c, p)$, can be either the Langmuir isotherm $q^* = Qkc/(1 + kc)$ or the Sips isotherm $q^* = Qkc^\alpha/(1 + kc^\alpha)$. For all simulations, as shown in the Appendix A, the parameters of the isotherms were kept constant with parameters $k = 1.80$ L/mg, $Q = 55.54$ mg/g, $\alpha = 1.5$. Langmuir isotherm parameters were based on Li et al. (2018b) work, and the Sips isotherm was built so that equilibrium values were at the same magnitude as the Langmuir model.

Here, it is important to remark that q^* is assumed to be known (structure and parameters), which contrasts with any other previously reported approach. It is a reasonable assumption as equilibrium models are usually fit separately from kinetics. Adding the task of fitting an isotherm with sorption kinetic simultaneously complicates the problem unnecessarily. Figure 4.2 illustrates the proposed methodology. The isotherm calculation is one of the nodes in the computational graph located in the leftmost layer. This variable is then concatenated with the adsorbed amount and used as an input of the ANN that predicts the instantaneous uptake rate.

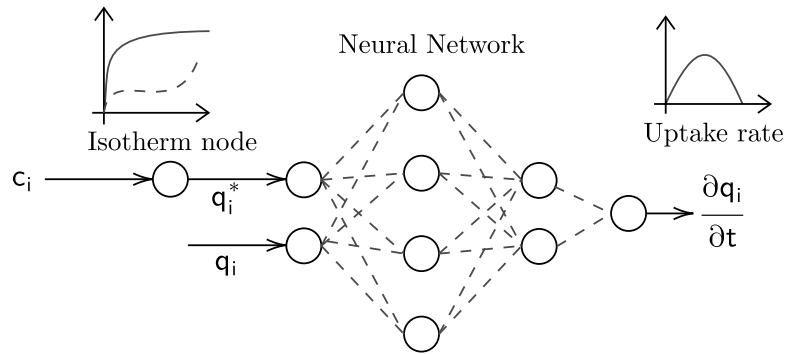


Figure 4.2: Schematic representation of the proposed hybrid model.

4.3.3 Model's numerical solution and gradient calculation

It is important to detail the numerical solution of the Partial Differential Equation (PDE) hybrid model as it is known that the stability and speed of gradient calculation are influenced by the discretization and numerical integration method (Rackauckas et al., 2020). In this sense, detailing all the steps to solve the non-linear advection-diffusion-sorption is paramount. The Partial Differential Equation (PDE) presented in Section 4.3.2 has to be solved numerically, and the resulting numerical solution is used for calculating sensitivities (gradients, Jacobians, Hessians). One approach for solving the presented model is discretizing the spatial domain and integrating the resulting Ordinary Differential Equation (ODE) in time with an appropriate numerical method – the popular method of lines. Many choices can be made in the discretization of the spatial domain. In the literature of fixed-bed chromatography, Orthogonal Collocation on Finite Elements (OCFEM) is preferred and recommended (Finlayson, 1974; Ma and Guiochon, 1991) as it achieves high accuracy in the presence of steep gradients in concentration profiles typically found in such problems.

This way, OCFEM is used with cubic Hermite polynomials (Ganaie et al., 2014) and zeros of shifted orthogonal Legendre polynomials as collocation points. Cubic Hermite polynomials are C^1 continuous at the connecting nodes in OCFEM formulation, considerably reducing the required number of equations in the discretization. Thus, the solution to the PDE can be expanded as a sum of cubic Hermite basis functions as:

$$y(u, t) = \sum_{i=1}^4 a_{i+2k-2}(t) H_i(u), \quad (4.8)$$

$$H_1(u) = (1 + 2u)(1 - u)^2, \quad (4.9)$$

$$H_2(u) = u(1 - u)^2 h_k, \quad (4.10)$$

$$H_3(u) = u^2(3 - 2u), \quad (4.11)$$

$$H_4(u) = u^2(u - 1)h_k, \quad (4.12)$$

u is the local coordinate inside each finite element, k is the finite element index, h_k is k^{th} finite element size, and $a(t)$ is the time-dependent coefficients solved in the ODE integration. A total of 42 evenly-spaced finite elements were used. The resulting system of ODE was then solved using a fixed-leading coefficient adaptive-order adaptive-time BDF method (FBDF) implemented in DifferentialEquations.jl with tolerances of 1×10^{-6} (Rackauckas and Nie, 2017). FBDF presented no instability and, by far, the fastest run time for solving the discretized PDE. The decision to employ 42 finite elements draws upon existing literature. It is common practice among scholars to use between 20 and 40 elements (Constantino et al., 2015; Nogueira et al., 2019) when discretizing the advection-diffusion-sorption partial differential equations with orthogonal collocation on finite elements. In

this study, the chosen quantity of finite elements represents the lowest number at which the solution to the PDE solution ceased to demonstrate significant variation.

Once the solution of the described model can be obtained numerically, the gradient calculation problem is formulated. As in any parameter estimation problem, a cost function has to be defined so the gradients can be calculated concerning the parameters and used in an optimizer. In this way, an L2-norm-based discrete cost function $G(\mathbf{u}, \theta)$, where $\mathbf{u}$ is the PDE solution vector, and θ is the ANNs' parameters, can be defined as:

$$G(\mathbf{u}, \theta) = \sum_{i=1}^N \|\mathbf{u}(t_i, \theta) - \mathbf{u}(t_i)^{obs}\|^2, \quad (4.13)$$

in which, $\mathbf{u}_i^{obs}$ are the vector of observations at each i^{th} point in time. In this work, $\mathbf{u}_i^{obs}$ corresponds to the liquid phase concentration at the column outlet, i.e., $c(x^* = 1, t_i)$. Thus, the gradient of the cost function consists in calculating $\frac{dG(\mathbf{u}, \theta)}{d\theta}$. It means that the solution of the ODE is part of the cost function. Calculating gradients of cost functions that contain the solution of ODEs has been investigated over many years (Ma et al., 2021). Continuous sensitivity analysis with forward-mode AD is efficient when the number of equations and parameters is small. However, the sum of parameters and equations may reach hundreds in hybrid models involving discretized partial differential equations and ANNs. In this scenario, continuous or discrete adjoint sensitivity analysis is preferred. In continuous adjoint sensitivity analysis, gradient calculation for ODEs consists in solving the problem (Rackauckas et al., 2020):

$$\frac{dG(\mathbf{u}, \theta)}{d\theta} = \frac{d}{d\theta} \int_{t_0}^T \left(\sum_{i=1}^N \|\mathbf{u}(t_i, \theta) - \mathbf{u}_i(t_i)^{obs}\|^2 \delta(t - t_i) - \lambda^* (M\mathbf{u}' - f(\mathbf{u}, \theta, t)) \right) dt \quad (4.14)$$

$$M^* \lambda' = -\frac{df^*}{du} \lambda - \frac{dg^*}{du} \quad (4.15)$$

$$\frac{dg}{du}(t_i) = 2(\mathbf{u}(t_i, \theta) - \mathbf{u}(t_i)^{obs}), \quad (4.16)$$

in which, M is the mass matrix, λ is the Lagrange multiplier vector, and $\frac{df^*}{du} \lambda$ is the vector-jacobian product, δ is the Dirac delta function. The $*$ notation means transpose, and $'$ means the time derivative.

The Julia SciMLSensitivity.jl library (Rackauckas et al., 2020) counts with many implementations of adjoints and intermediate calculations. There are 48 ways of solving the adjoint problem, each with different performance characteristics and limitations. However, Ma et al. (2021) shows that the quadrature adjoint with AD-jacobian vector product seeding outperforms any other method when the number of parameters + equations surpasses 100. In this way, the quadrature adjoint with Just-In-Time (JIT)-compiled reverse-mode AD vector-jacobian product with FBDF solver is used for solving Equations 4.14, 4.15, and 4.16 for the hybrid model presented in Section 4.3.2. It was the most performant and

stable method for solving the adjoint problem resulting from a quadratic loss function and hybrid PDE.

Figure 4.3 illustrates the decision-making process for training hybrid models in non-linear advection-diffusion-sorption PDE systems. The lines in pink indicate the one used in this work. First, a parameter initialization algorithm has to be chosen. Then, the stability of the primal solution of the discretized PDE has to be assessed. A decision on derivative-free or derivative-based optimizers has to be taken - if the cost function is differentiable, derivative-based optimizers are recommended. In derivative-based optimizers, adjoint sensitivity analysis is the most accurate and efficient way of calculating derivatives for large parameter space. However, the stability of the adjoint problem solution depends on the system under study. Here, a way to use the most efficient approach for fitting ANNs' weights and biases was found.

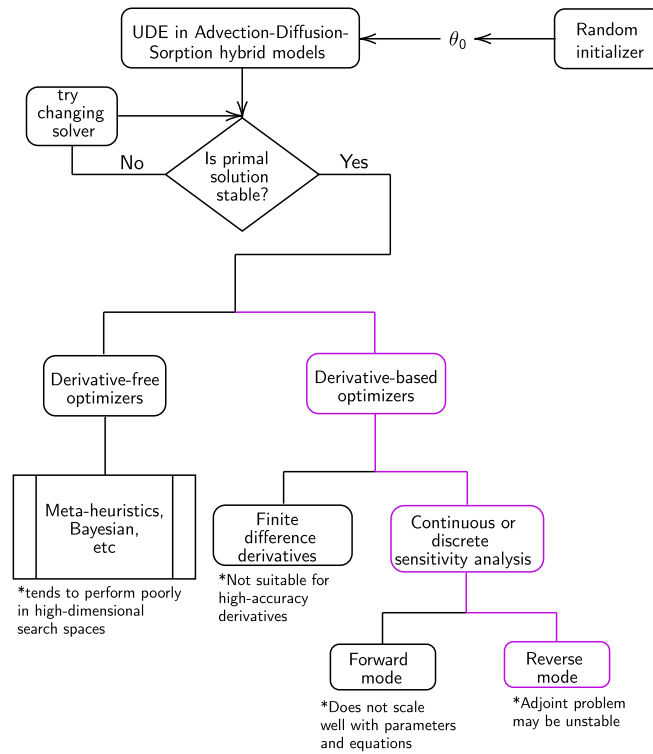


Figure 4.3: The decision-making process for identifying hybrid models.

4.3.4 Sparse Regression

Sparse regression is a technique used in scientific machine learning for identifying a small subset of important variables from a more extensive set of potential predictors. Sparse regression methods aim to find a sparse solution where only a subset of the predictors is included in the final model. It results in a more straightforward and interpretable model while maintaining good predictive performance. Several methods for performing sparse

regression are available, including Lasso (Least absolute shrinkage and selection operator), Ridge Regression, and SINDy (Sparse Identification of Nonlinear Dynamic systems) (Brunton et al., 2016).

In universal differential equations and hybrid modeling, sparse regression can recover the missing interaction term in the differential equation (Dandekar et al., 2020). When trained, the ANN should resemble the missing function in the differential equation. By applying sparse regression on the trained ANN, one can recover the missing function at a computational cost significantly slower than running it directly on the original PDE. Moreover, the popular Sparse Identification of Nonlinear Dynamics (SINDy) method, for instance, involves using a spline interpolation over data points. However, this approach requires a dense set of data points for accurate derivative estimates. In contrast, running sparse regression on the ANN does not require this, as it uses a trained neural network that can be continuously sampled. Assuming a trained ANN, the optimal parameter value $\hat{\theta}$ can be written as:

$$\hat{\theta} \leftarrow \arg \min_{\theta} G(\mathbf{u}, \theta) = \sum_{i=1}^N \|\mathbf{c}(t_i, \theta, x^* = 1) - \mathbf{c}_i^{obs}\|_2 \quad (4.17)$$

Then, to recover the missing interaction with sparse regression, the following problem is solved:

$$\hat{\Xi} \leftarrow \arg \min_{\Xi} \|\text{ANN}(q^*(x = L, t_i), q(x = L, t_i), \hat{\theta}) - \Theta \Xi\|_2 + \Phi \|\Xi\|_1 \quad (4.18)$$

Where Θ is the set of candidate basis functions (functions of q^* and q), Ξ is the parameters of the basis functions, Φ is the regularization parameter, and t_i is the time when the solution of the PDE is evaluated and used in the ANN.

On the one hand, even with their universal approximation ability, ANNs are not the best choice for elucidating the kinetics in advection diffusion-sorption systems. Sparse and symbolic regressions are commendable for their simplicity and interpretability. On the other hand, executing symbolic/sparse regression directly on partial differential equations, particularly highly stiff ones like the advection diffusion-sorption systems, poses important computational challenges. Each iteration equates to a parameter estimation task, each necessitating numerous forward and backward PDE solutions. This escalates computational time and risks deriving unstable dynamics in many regression attempts. Incorporating a universal approximator into the PDE using the Universal Differential Equation approach mandates a single, albeit complex, parameter estimation. This ANN parameter fitting was demonstrated empirically to maintain a stable dynamic system for such PDE. Post ANN training, symbolic and sparse regression can be applied more efficiently, handling merely algebraic equations without requiring hundreds of parameter estimation problems of stiff ODE solutions. Despite not being the best, this approach is a feasible path to elucidate the sorption kinetics directly from breakthrough data.

To better elucidate the difference in implementing direct sparse/symbolic regression in the original PDE, it is illustrated in Figure 4.4.

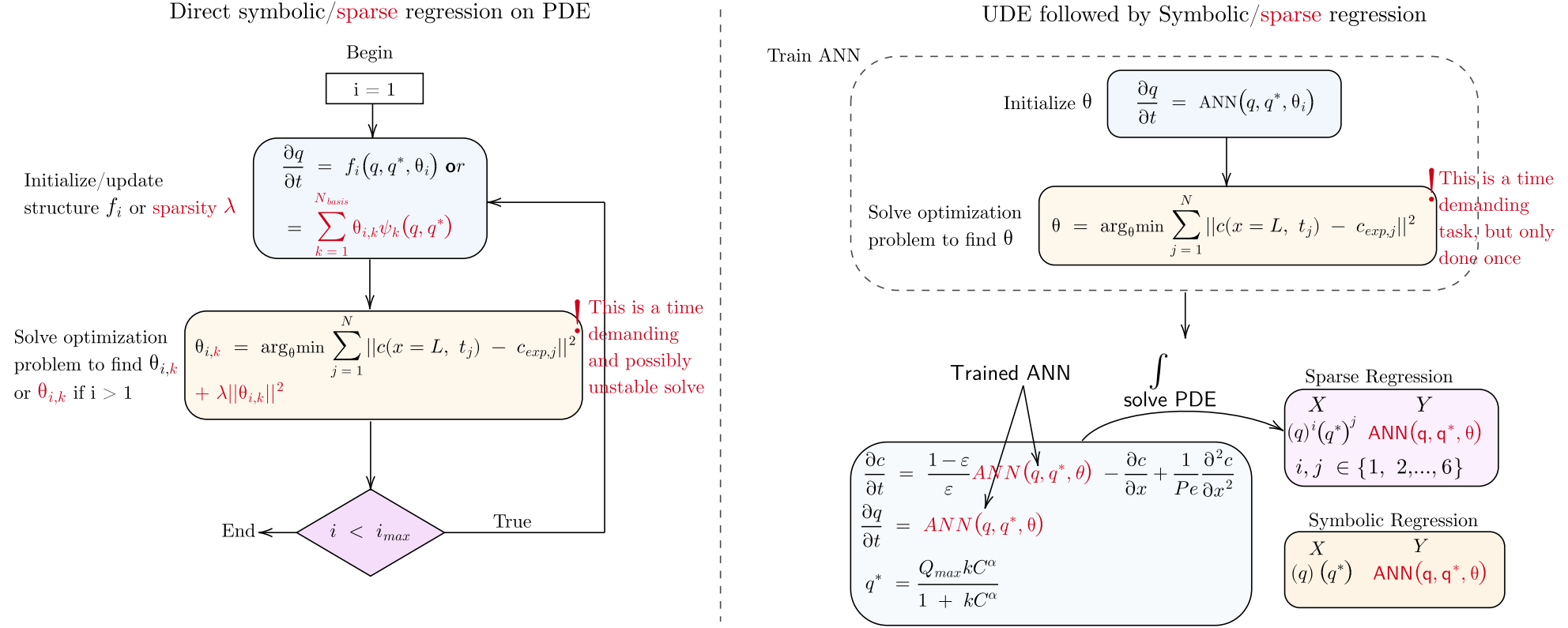


Figure 4.4: Comparison between direct sparse/symbolic regression and UDE approach for mass transfer kinetic identification on PDEs

4.4 Results and Discussion

4.4.1 UDEs' training and test sets performance

This subsection shows the training and test performance of the UDE-based hybrid model proposed for all conditions described in 4.3.1. The history of predictions and *in-silico* data points are displayed as the prediction error. Moreover, the training uptake rate is also displayed and compared to the ground truth, i.e., the noise-free training set uptake rate of the mechanistic model used.

Table 4.2 displays the hyperparameters of the training process, i.e., the number of layers and neurons for each of the cases shown above. The learning rate during optimization with ADAM started at 0.05 with an exponential decay every 20 iterations for all cases. After ADAM, the BFGS optimizer was used until convergence.

Table 4.2: ANNs' hyperparameters

Isotherm	Kinetic	Layers	Neurons per layer
Langmuir	LDF	1	17
Langmuir	improved LDF	1	17
Langmuir	Vermeulen's	2	10, 8
Sips	LDF	1	20
Sips	improved LDF	1	20
Sips	Vermeulen's	2	10, 8

As seen for the Langmuir isotherm and LDF kinetics in Figure 4.5, the UDE approach fits breakthrough training data well, i.e., with errors compatible with simulated noise and no apparent auto-correlation in time. Surprisingly, it also performs very well in the test set where desorption and adsorption from another steady state occur. The uptake rate is also close to the training noiseless ground-truth data. The same conclusions can be drawn for the Langmuir isotherm and improved LDF kinetics – Figure 4.6. However, for Vermeulen's model (Figure 4.7c), the uptake rate is underestimated despite the good fitting of breakthrough data. A key contributor to this underestimation is the notably pronounced sharp change in the uptake rate near the peak of Vermeulen's model, more so than in other models. It's worth noting that enhancing the sampling rate around this crucial inflection point could improve the model's predictive accuracy within this particular domain.

The results and conclusions for the Sips isotherm and all kinetics were similar to the ones obtained for Langmuir. For this reason, the results of it are displayed at the Appendix A.

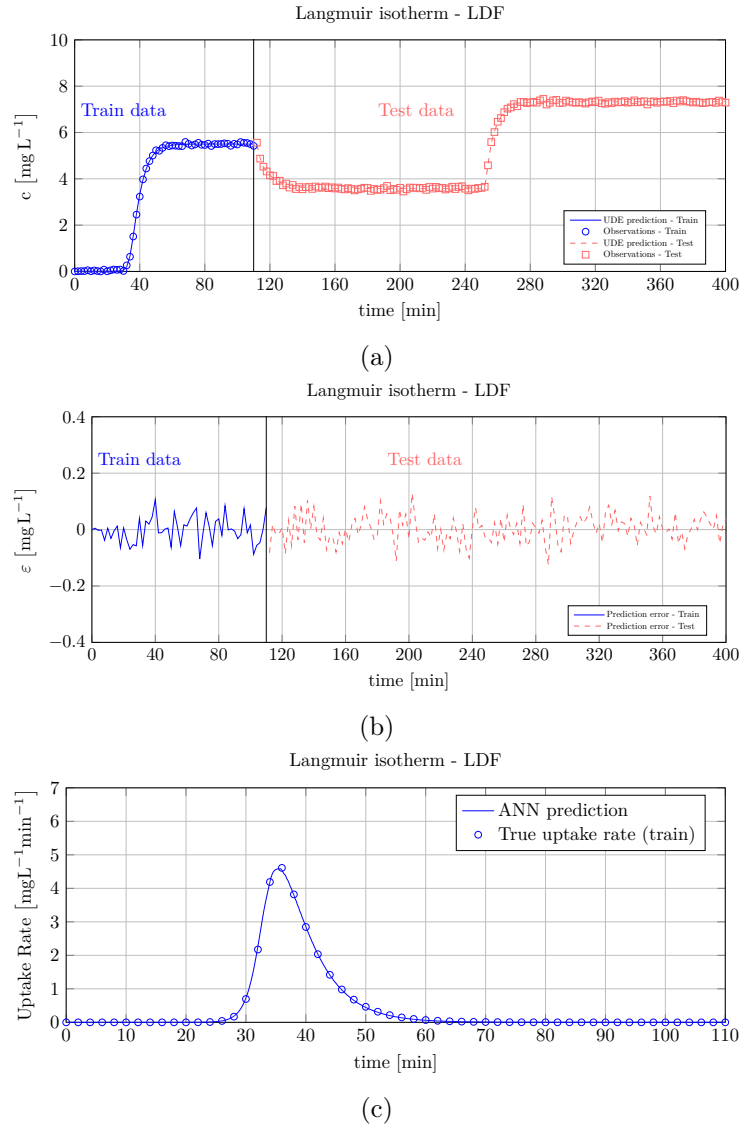


Figure 4.5: Results for Langmuir isotherm and LDF kinetics - 0.5 min⁻¹ sampling rate (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets. (c) UDE prediction of uptake rate ($x^* = 1$) compared to ground truth.

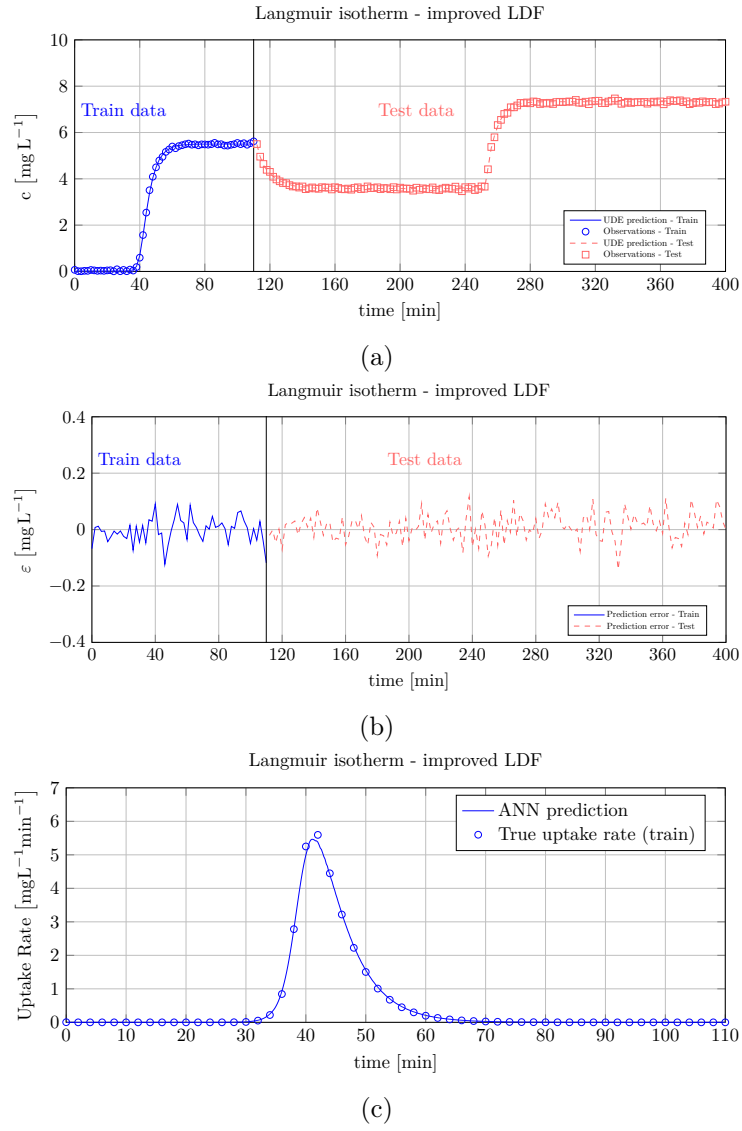


Figure 4.6: Results for Langmuir isotherm and improved LDF kinetics - 0.5 min⁻¹ sampling rate (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets. (c) UDE prediction of uptake rate ($x^* = 1$) compared to ground truth.

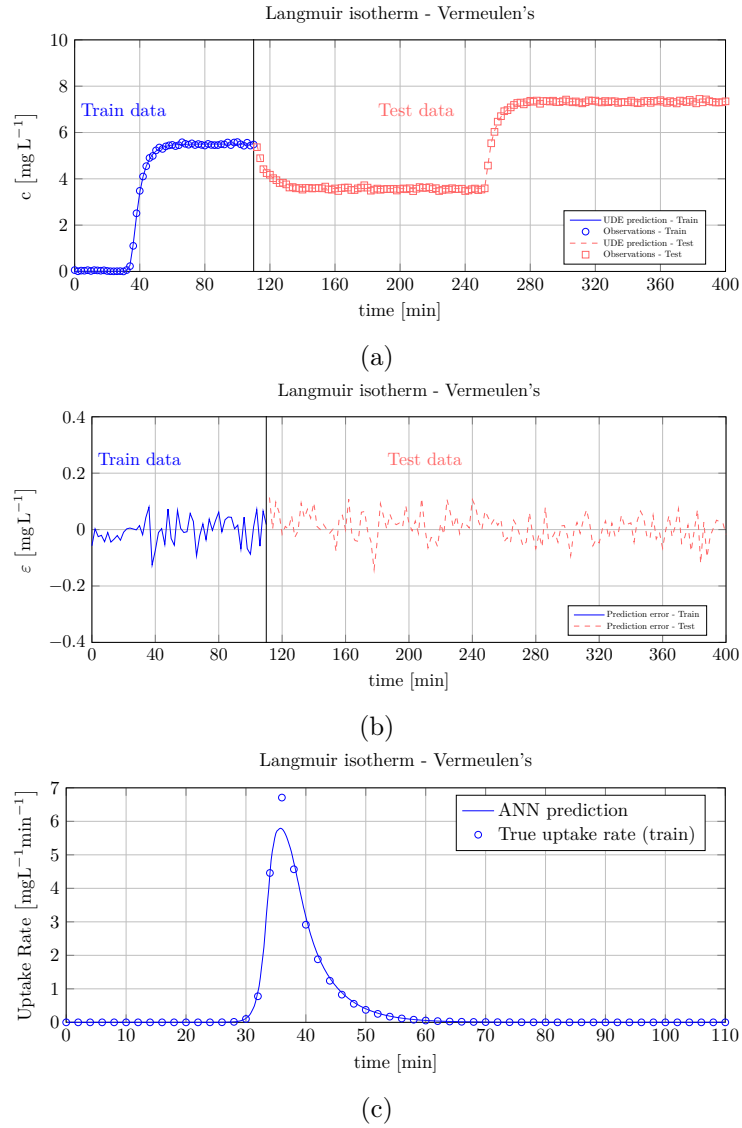


Figure 4.7: Results for Langmuir isotherm and Vermeulen's kinetics - 0.5 min⁻¹ sampling rate (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets. (c) UDE prediction of uptake rate ($x^* = 1$) compared to ground truth.

4.4.2 Sparse Regression

As seen in Section 4.4.1, the fitted ANN showed a reasonable train and test set performance for predicting breakthrough data. The uptake rates are also close to the ground truth. However, the neural network outputs are hard to interpret. In this sense, sparse regression can be used to understand the missing functions that the ANN approximates. Here, a set of basis polynomials up to sixth order are used to carry sparse regression using the SINDy algorithm, i.e., $\Theta \in q^{*i} \times q^j$ $i, j \in (0, 1, 2, 3, 4, 5, 6)$. That results in 49 basis to be combined in regression. The uptake rate at three different positions in the bed was used for carrying out the regression.

As sparse regression requires tuning the regularization parameter Φ , a grid search on Φ s in $\log_{10}$ space from -3.0 to 0.0 with 0.05 step was carried using ADMM algorithm in DataDrivenDiffEq.jl (JuliusMartensen et al., 2021) (an implementation of Lasso - Least absolute shrinkage and selection operator). The Bayesian Information Criterion (BIC) is compared for each Φ and used to decide which model is the best – the lower the BIC, the higher the chance that a model explains the observations. BIC is a metric that combines the model’s predictive capacity (squared residuals) and complexity (number of parameters). Here, it is tested if this approach allows reconstructing the known kinetics originating the data, i.e., LDF, improved LDF, and Vermeulen’s model. After running the sparse regression on the ANN, the parameters are readjusted using Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimizer over 1000 iterations.

It is important to note that the need to carry sparse regression on the ANN outputs appears when test set (extrapolation) performance is poor or ANNs’ interpretability is desired. However, it was demonstrated that the ANNs performed surprisingly well in extrapolation tasks. In this way, the attempt to recover interaction terms in the current work serves only the purpose of giving interpretability to ANNs’ outputs.

Table 4.3 summarizes the results of sparse regression for each case. It can be noted that the obtained polynomial resembles the original kinetics for the LDF kinetics. For improved LDF, interpreting the obtained polynomial is not straightforward due to an exponential term in the original interaction (Equation A.11). This is also true for Vermeulen’s due to the q term in the denominator. However, as shown in the next section, improved LDF and Vermeulen’s expression can be expanded using Taylor series and compared to the obtained polynomials.

Figure 4.8 shows that a good fitting is achieved for the Langmuir isotherm and LDF kinetics and that the dominant coefficients are the order 1 exponent, which agrees with the original interaction term (Table 4.3). After running sparse regression on the ANN, the obtained polynomial can be used in the original PDE for carrying predictions. Figure 4.9 shows that the breakthroughs are well predicted with this 3-term polynomial in the kinetics.

For the Langmuir isotherm and improved LDF kinetics, Figure 4.10 shows that a good

Table 4.3: Sparse regression learned polynomials.

Isotherm	Kinetic	True kinetics	Learned kinetics
Langmuir	LDF	$0.22q^* - 0.22q$	$-0.535 - 0.225q + 0.234(q^*)$
Langmuir	improved LDF	$0.22(q^* + 0.2789q^*e^{\frac{-q}{2q^*}} - q)$	$-0.549 - 0.221q + 0.278(q^*) - 0.000212qq^*$
Langmuir	Vermeulen's	$0.22\frac{q^{*2}-q^2}{2.0q}$	$-1.253 - 0.429q + 0.0034(q^2) + 0.506q^*$ $-0.0023(q^{*2}) - 0.0021qq^*$
Sips	LDF	$0.22q^* - 0.22q$	$-0.275 - 0.210q + 0.214(q^*)$
Sips	improved LDF	$0.22(q^* + 0.2789q^*e^{\frac{-q}{2q^*}} - q)$	$-0.183 - 0.215q - 0.00029(q^2) + 0.272(q^*)$
Sips	Vermeulen's	$0.22\frac{q^{*2}-q^2}{2.0q}$	$-0.309 - 0.449q + 0.00037(q^2) + 0.438q^*$ $-6.326 \times 10^{-5}(q^{*3}) + 6.214 \times 10^{-5}q(q^{*2})$

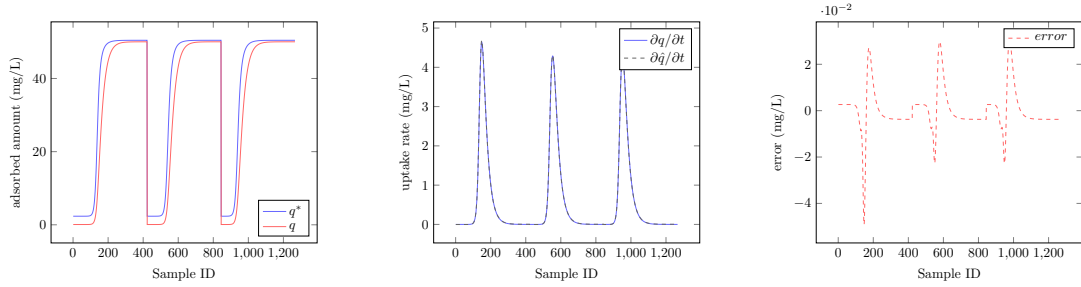


Figure 4.8: Sparse regression results for the best number of active terms (3) - Langmuir isotherm + LDF kinetics at three positions in the bed.

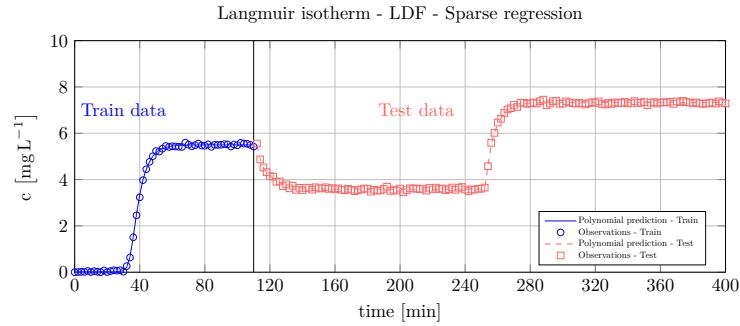


Figure 4.9: Results for Langmuir isotherm and LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.

fitting is achieved. Figure 4.11 shows that the breakthroughs are well predicted with this 4-term polynomial in the kinetics.

Finally, Figure 4.13 shows that training and test breakthroughs are well predicted with this 6-term polynomial in the Langmuir isotherm and Vermeulen's kinetics.

For Sips isotherm, similar fitting results were obtained; hence, similar conclusions can be made. The plots showing the fitting quality for this isotherm are displayed at the Appendix A.

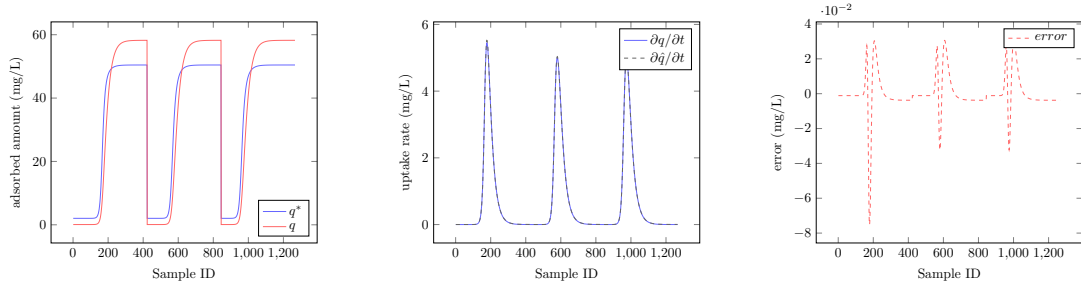


Figure 4.10: Sparse regression results for the best number of active terms (4) - Langmuir isotherm + improved LDF kinetics at three positions in the bed.

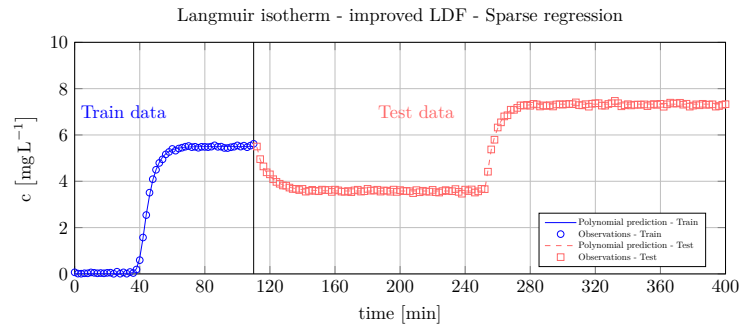


Figure 4.11: Results for Langmuir isotherm and improved LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.

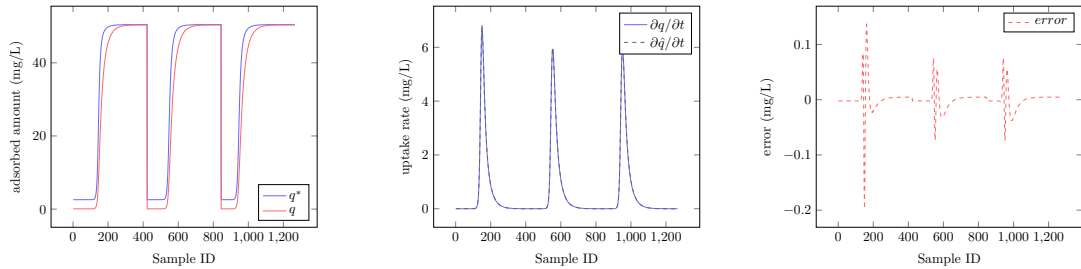


Figure 4.12: Sparse regression results for the best number of active terms (6) - Langmuir isotherm + Vermeulen's kinetics at three positions in the bed.

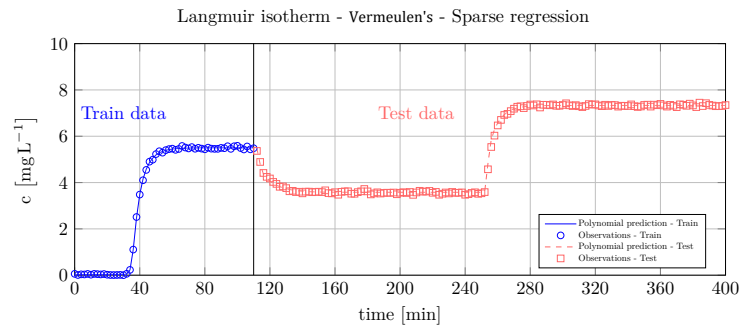


Figure 4.13: Results for Langmuir isotherm and Vermeulen's kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.

4.4.3 Symbolic Regression

The trained ANNs' outputs were also regressed by using symbolic regression. Symbolic regression is a technique that aims to uncover a model's structure and parameters in a single step. This process involves searching through a space of mathematical expressions to find the best representation of the data rather than relying on a pre-defined model structure. One of the most common approaches to solving the symbolic regression problem is genetic programming (GP). This method involves creating a system of arithmetic operators (addition, subtraction, multiplication, and division) and functions (identity, logarithmic, exponential, trigonometric, and polynomial) to generate an expression tree. The expression tree represents a mathematical formula, and the goal is to optimize the order of the operators in the tree to achieve the best fit for the data. Several techniques of GP have been used to carry out symbolic regression.

In this case, two binary operators were used: sum and multiplication. The scoring method proposed by Cranmer et al. (2020) was implemented in Julia language using the SymbolicRegression.jl library (Cranmer, 2020). The algorithm was run for over 30 iterations with a population size of 30. A Pareto front of polynomials was generated, and the polynomial with the lowest mean squared error was chosen as the optimal solution.

One of the advantages of symbolic regression compared to sparse regression is that it does not require tuning the sparsity parameter Φ – a parameter with considerable influence in sparse regression results. Table 4.4 shows the obtained polynomials using symbolic regression. Learned polynomials are similar to the ones obtained via sparse regression – differences appeared in structure, but residuals were at the same order of magnitude (10^{-2}). It is important to note that our discovery of only polynomial structures as optimal configurations is a consequential finding rather than a premeditated selection. In LDF kinetics and Sips isotherm, the obtained structure is identical to the original, with very close coefficients. Expressions in symbolic regression, in general, are simpler (with fewer terms) than those obtained with sparse regression.

Table 4.4: Symbolic regression learned polynomials.

Isotherm	Kinetic	True kinetics	Learned kinetics
Langmuir	LDF	$0.22q^* - 0.22q$	$-0.535 - 0.225q + 0.234(q^*)$
Langmuir	improved LDF	$0.22(q^* + 0.2789q^*e^{\frac{-q}{2q^*}} - q)$	$-0.554 - 0.234q + 0.281(q^*)$
Langmuir	Vermeulen's	$0.22\frac{q^{*2}-q^2}{2.0q}$	$-0.6098 + 0.0122q + 0.263q^*$ $-0.00526qq^*$
Sips	LDF	$0.22q^* - 0.22q$	$0.198q^* - 0.200q$
Sips	improved LDF	$0.22(q^* + 0.2789q^*e^{\frac{-q}{2q^*}} - q)$	$0.277q^* - 0.241q$
Sips	Vermeulen's	$0.22\frac{q^{*2}-q^2}{2.0q}$	$-0.003557q^{*2} - 0.216q + 0.395q^*$

These results demonstrate that retrieving the correct interaction term from noisy observations in such problems is not trivial despite the high-accuracy fitting. Most of the obtained polynomials resemble the true kinetic models that produced the data. The obtained

polynomials, specifically those from the improved LDF and Vermeulen's cases, cannot be directly compared with the original terms. However, a Taylor expansion analysis of the mechanistic model results, as shown in the equations in Appendix A, can demonstrate why a polynomial with positive integer coefficients (i.e., without $1/q$ and $\exp(-0.5q^*/q)$ basis) can approximate the true kinetics for such cases. For example, a second-order Taylor expansion was conducted for Vermeulen's model and Langmuir isotherm at sample ID = 50, around $q^* = 50.13$ and $q = 48.11$. This resulted in the following equation:

$$0.22 \frac{q^{*2} - q^2}{2.0q} (50.13, 48.11) \approx 0.454 + 0.229q^* - 0.229q + 0.002286q^{*2} - 0.004764q^*q + 0.002482q^2 + \mathcal{O}(\|x^3\|) \quad (4.19)$$

Figure 4.14 shows how well a simple second-order Taylor expansion around sample ID = 50 can capture the entire behavior of the uptake rate for points far away from the expansion point. It can explain why polynomials in sparse and symbolic regression explain the observations satisfactorily.

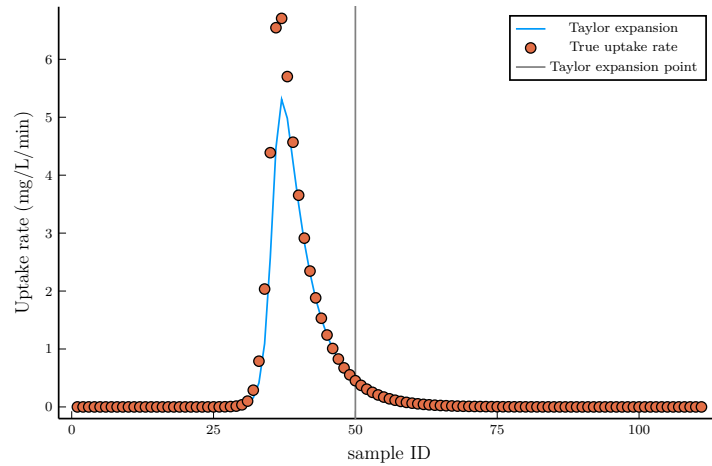


Figure 4.14: Taylor expansion of uptake rate for Vermeulen's model and Langmuir isotherm - Mechanistic model

A similar analysis can be done for improved LDF kinetics. A first-order Taylor expansion at sample ID = 50 ($q^* = 49.23$ and $q = 49.22$) of the mechanistic model results gives:

$$0.22(q^* + 0.2789q^*e^{\frac{-q}{2q^*}} - q)(49.23, 49.22) \approx 1.834 + 0.275q^* - 0.238q + \mathcal{O}(\|x^2\|) \quad (4.20)$$

Once again, by plotting the Taylor polynomial against the true uptake rate (Figure 4.15), it is possible to see how well the points match and explain why both symbolic and sparse regressions succeed in explaining observations with such structure (polynomials).

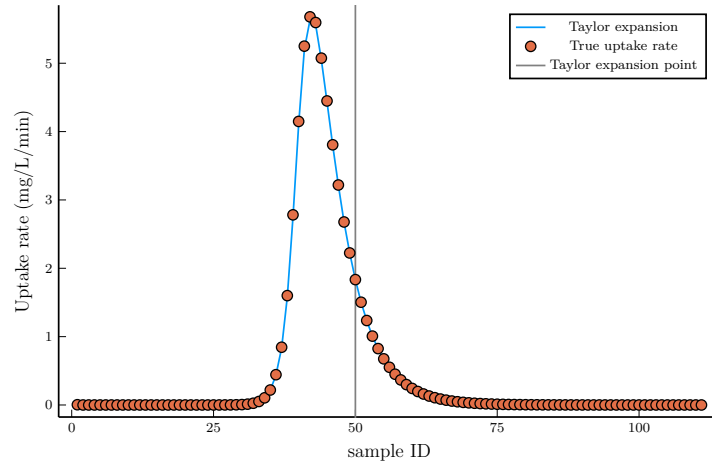


Figure 4.15: Taylor expansion of uptake rate for improved LDF model and Langmuir isotherm - Mechanistic model

We remark that the outstanding performance of Artificial Neural Networks (ANN) in the hybrid advection-diffusion-sorption PDE can be primarily attributed to its capabilities in precisely recovering the sorption uptake expressions. The comparison of original and recovered interaction terms, as illustrated in Tables 4.3 and 4.4, where both sparse and symbolic regression were used to extract interpretable/straightforward mathematical expressions that encapsulate the high-dimensional mappings learned by the network, combined with Taylor expansion analysis, reveals that the trained ANNs mirror the characteristics of the original sorption uptake functions. These findings suggest a strong alignment between our trained ANNs' performance and the original function's properties, demonstrating the network's successful approximation capabilities. Nevertheless, it is crucial to acknowledge that the ANNs' central role in our research was to serve as a universal approximator, with their detailed interpretation being executed via sparse/symbolic regression. Once we have derived simplified expressions through the regression of the trained ANN, these can essentially replace the ANN. This substitution is viable given these simplified expressions' impressive predictive power.

In conclusion, combining the uptake rate plots with sparse/symbolic regression produces an efficient tool for investigating sorption mass transfer kinetics in non-linear advection-diffusion-sorption systems. Compared to the original functions, the difference in the functional representation learned by Lasso/symbolic regression could be attributed (i) to noisy data and (ii) the fact that the two functional representations are numerically equivalent in this range of values and response shape.

4.5 Conclusions

The present work investigated the numerical aspects of hybrid non-linear advection-diffusion-sorption PDE problems. In this sense, we proposed a framework for efficiently training ANN-based hybrid non-linear advection-diffusion-sorption PDE problems using gradient-based optimizers via (i) continuous adjoint sensitivity analysis, (ii) orthogonal collocation on finite element PDE discretization scheme and an (iii) fixed-leading coefficient adaptive-order adaptive-time BDF method ODE integrator. Another contribution of the present work was the investigation of the interpretability of ANNs' prediction through sparse and symbolic regression. To demonstrate the contributions above, an in-silico data set was used. The data set was built by simulating a hypothetical single-component non-linear advection-diffusion-sorption system with two different adsorption equilibrium (Langmuir and Sips) and three kinetic models – Linear Driving Force (LDF), Vermeulen's model and improved LDF. The results show that the ANN can learn adsorption uptake from single noisy breakthroughs at column outlet and extrapolate in adsorption and desorption well.

Moreover, by carrying sparse and symbolic regression on ANNs' outputs, polynomials resembling the true kinetic models that produced the data were obtained - especially for the LDF kinetics. In the cases of non-trivial resemblance, i.e., improved LDF and Vermeulen's, a polynomial with high predictive power was obtained, and its appearance was elucidated with Taylor expansion analysis. It was demonstrated that the proposed framework could be used for kinetic model discovery from single-position noisy breakthrough data in advection-diffusion-sorption PDEs.

Part II

Model Acceleration and Scale Integration with Data-Driven Modeling in Chemical Engineering

Chapter 5

Optimal fragrances formulation using a deep learning neural network architecture: a novel systematic approach.

5.1 Abstract

This chapter within Part II delves into exploring modeling strategies and applications in chemical engineering, particularly in contexts characterized by abundant, informative data and limited domain knowledge. This chapter focuses on the optimal formulation of perfumes, a complex yet understudied challenge in the Flavor and fragrance industry. Leveraging the data-rich environment typically associated with reduced-order or surrogate models, the chapter introduces a novel framework incorporating a phenomenological model to produce an extensive training set for an AI model. This AI model employs an objective function designed to encapsulate the desired olfactory properties of a perfume. The chapter underscores the utility of this approach in solving intricate problems in fragrance formulation, thereby making significant strides in merging data-driven methods with domain-specific challenges. It exemplifies the overarching aim of Part II, which is to demonstrate effective strategies for leveraging abundant data in chemical engineering.

5.2 Introduction

The human olfaction is a complex sensorial system recently discovered to discriminate over 1 trillion different smells (Bushdid et al., 2014). The reasons it detects and responds pleasantly to certain substances are not fully understood. Nonetheless, humans have been economically exploring the enjoyable smell of substances for centuries, giving rise to the Flavor and Fragrance (FF) multi-billion dollar business (Associates, 2018).

Fragrances are part of two chemical product classes – functional and formulated (Zhang et al., 2020b). First, fragranced molecules are used to improve consumer acceptance or mask unpleasant odors on daily used products such as soap, shampoos, cosmetics, and toothpaste. In the second, they are active ingredients with the main objective to be noticeable, such as perfumes.

The Flavor and Fragrance (FF) industry sales have experienced sustained growth in the last decades (Associates, 2018), leading manufacturers to seek new tools for driving innovation while meeting the growing demand. Even so, the scientific literature is surprisingly scarce on works that develop tools to foster innovation in the field. This is particularly pronounced for fine fragrances (perfumes). This might result from a collective trend to believe that well-formulated perfumes result from artistic alchemy. However, it is indeed a complex task to formulate a perfume. First, a set of molecules is chosen to satisfy a compositional rule of three notes: top, middle, and base (Carles, 1962).

Each note is used to cause distinct olfactory impressions and improves perfume evaporation features: Top notes are intended to be the protagonist’s smell in the first few minutes but should not last too long. The middle notes must dominate the scent right after the top notes and can last for hours – it is considered the “body” and the “heart” of the perfume. The base notes must last for hours and even days and act as a fixative by reducing the evaporation rate of top and middle note chemicals.

The notes are the building blocks of the perfumes but are not stand-alone chemicals in a real perfume formulation. They are incorporated in a solvent matrix – generally a blend of solvents comprised mostly of ethanol. Unfortunately, the human olfactory system is highly sensitive to ethanol and responds unpleasantly to it, which poses a challenge in the next stage of the formulation: choosing how much of each chemical to use to achieve an odor characteristic, avoiding ethanol smell dominance.

From that perspective, perfumes are liquid solutions of fragrances in a proper solvent. To perceive the fragrances, their vapors must diffuse through the surrounding space and be detected by a human nose. From this point of view, a perfume allows incorporating several consolidated methods from chemical engineering disciplines; the process of transferring chemicals from liquid to vapor phase is very well described by models for vapor–liquid phase equilibrium and gas molecular diffusion. Human detection can be analyzed using models from psychophysics. Several works have addressed the experience of scenting perfume from this perspective (Almeida et al., 2019; Teixeira et al., 2009b, 2010b, 2013a). These works have addressed the perfume release and propagation by considering the evaporation of odorants from human skin and modeling it with a skin absorption rate and air evaporation, proposing a phenomenological model to represent the phenomena, opening space to the application of engineering tools to solve F&F industry problems.

Though a non-conventional technique, the use of high-fidelity simulation to optimize a perfume was introduced by Mata et al. (2005) through the perfume ternary diagram (PTD) method. It offers a quick guide to choosing a ternary perfume and a solvent composition

to optimize its odor features. The composition must be chosen according to the works: (1) The top note should be noticed strongly near the perfume or shortly after the application. (2) The base note should be noticed long after the application, and (3) ethanol should not be noticed since it is unpleasant and unappreciated. Despite its simplicity and capacity to provide an efficient guide for simple perfume mixtures (three components), the method can easily lose effectiveness in more complex scenarios. Real perfume mixtures can contain up to 100 components (Teixeira et al., 2013a). On the other hand, high-fidelity simulation relies on mathematical models traditionally grounded in the natural sciences, where phenomenology is important and may demand significant computational power.

From that perspective, some works in the literature have proposed using Artificial Intelligence (AI) tools to help design perfumes. Zhang et al. (2018) trained an artificial neural network model that learned how to classify a fragrance molecule among 20 possible descriptors with great accuracy. It encoded the molecular structure into numerical predictors using a fragment (group) based representation. Researchers from Google introduced a graph neural network for Quantitative Structure-Odor Relationship (QSOR) modeling (Sanchez-Lengeling et al., 2019a). The work proposes to encode molecular structures of odorants as graphs in which nodes are atoms and edges are chemical bonds. The application of AI in F&F-related problems is still limited to the classification point of view. No reports were found applying AI in the field to use its computationally cheaper models for expensive simulations, as suggested in order fields (von Rueden et al., 2020). Furthermore, up to the present date, a systematic approach has not been found to perform an optimal perfume formulation considering its whole spectrum. Addressing a similar issue, Zhang et al. (2020b) proposed a systematic methodology to maximize the sensorial quality of cosmetics.

Thus, the present work aims to develop a systematic strategy for optimal perfume design considering all four pillars that compose the fragrance, the solvent, and the top, middle, and base notes. The strategy comprises a phenomenological model to generate an extensive database to train an AI model, an objective function that precisely reflects how desirable the construction of the smell of the perfume is, and a robust optimization method to maximize it. The Deep Neural Network (DNN) architecture was chosen among the AI techniques because of its notorious capacity to model dynamic systems. A Deep Learning (DL) based model can efficiently replace the high-fidelity simulation and save considerable CPU time. A meta-heuristic technique was used to solve the optimization problem, Particle Swarm Optimization (PSO), highlighted as an effective method to find minima in complex optimization tasks (Bohorquez et al., 2020; Schwaab et al., 2008). In this way, coupling a surrogate model with a powerful optimization method seems more appropriate for designing and optimizing real fragrance structures following specific requirements, as it will be further demonstrated.

The proposed methodology was applied in a case study where the goal was to find the optimal perfume composition that would provide an equivalent odor of pine forest and

floral through space and time while minimizing as much alcohol odor as possible.

5.3 Methodology

This section is divided into two parts. The first part details the stages to identify the deep learning models representing the fragrance release and propagation through time and space. The second part deals with the underlying optimization problem formulation for fragrance performance.

5.3.1 The perfume diffusion process and model

Understanding the release of chemicals from a liquid mixture and their propagation through the air is important in designing perfume products with the desired performance. A real perfume experience would require the odorants to evaporate from the liquid phase, travel through the surrounding air space, and reach a human nose to be perceived. A way to mimic this process chain is by framing the release and propagation of perfumes as the well-known Stefan tube device (Heinzelmann et al., 1965; Lee and Wilke, 1954). Figure 5.1 illustrates it. The core idea is that odorant molecules in liquid i (perfume) can escape to the gas phase (air) and diffuse through it in the z -direction. On top of the tube, a passing gas stream quickly removes any chemical at that point, preventing its accumulation in detectable amounts.

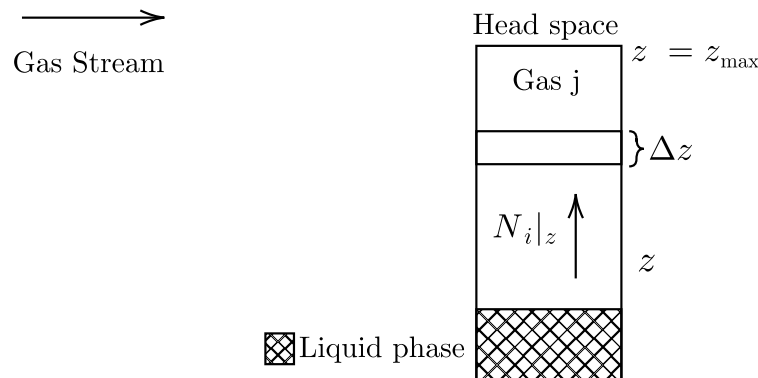


Figure 5.1: Illustration of the device for perfume release and propagation study.

In a real perfume experience, transport mechanisms other than molecular diffusion may play an important role in traveling odorants through the surrounding air. Even so, the Stefan tube offers a simple and effective tool for evaluating perfume behavior in a controllable way.

The model presented below was proposed by Mata et al. (2005). In the referred work, a first principle model using mass conservation, Fick's law of diffusion, and phase equilibria equations with good agreement between experimental observations and predictions was proposed. The model was verified experimentally in Teixeira et al. (2013a). Thus, this

model was implemented here to accurately represent the perfume release and propagation and generate the database for the empirical model development. It assumes:

- The radial flux in the gas phase is negligible;
- The air is not soluble in the liquid phase nor react with it; it can be considered a stagnant gas;
- The gas phase is an ideal gas mixture;
- The mass transfer resistance in the liquid phase is negligible;
- The liquid/gas interface is at chemical equilibrium.

The diffusion for a given chemical component relative to a stationary coordinate z can be expressed as (Bird et al., 1961):

$$N_{i,z} = -C_T D_{i,j} \left(\frac{\partial y_i}{\partial z} \right) + y_i (N_{i,z} + N_{j,z}) \quad (5.1)$$

Considering species j as the air, which is stagnant, Equation 5.1 can be rearranged to:

$$N_{i,z} = -\frac{C_T D_{i,j}}{1 - y_i} \frac{\partial y_i}{\partial z} \quad (5.2)$$

Thus, the differential material balance of component i can be written as:

$$C_T \frac{\partial y_i}{\partial t} = -\frac{\partial N_{i,z}}{\partial z} = -\frac{\partial}{\partial z} \left(-\frac{C_T D_{i,j}}{1 - y_i} \frac{\partial y_i}{\partial z} \right) \quad (5.3)$$

By applying the derivative operator on the right-hand side of Equation 5.3, it becomes:

$$\frac{\partial y_i}{\partial t} = \frac{D_{i,j} \left[\left(\frac{\partial y_i}{\partial z} \right)^2 + (1 - y_i) \frac{\partial^2 y_i}{\partial z^2} \right]}{(1 - y_i)^2} \quad (5.4)$$

Equation 5.4 describes the relationship of the gas mole fraction with respect to time t and position z . Similarly, a material balance in the liquid phase can be written considering a sink term as the flux at the liquid/gas interface. It can be written as:

$$\frac{dn_i}{dt} = A_{gl} \frac{C_T D_{i,j}}{1 - y_i} \frac{\partial y_i}{\partial z} \Big|_{z=0} \quad (5.5)$$

Initial and boundary conditions are necessary to find a unique solution to Equations 5.4 and 5.5. They can be written as:

$$\text{I.C1: } t = 0, \quad 0 < z \leq z_{\max} \Rightarrow y_i = 0 \quad (5.6)$$

$$\text{I.C2: } t = 0, \quad z = 0 \Rightarrow n_i = n_i^0, \quad n_i^0 \in \mathbb{R}_{0+} \quad (5.7)$$

$$B.C : \quad t > 0, \quad z = 0 \Rightarrow y_i = y_i^{\text{eq}} = \frac{\gamma_i x_i P_i^{\text{sat}}}{P} \quad (5.8)$$

$$t > 0, \quad z = z_{\text{max}} \Rightarrow y_i = 0 \quad (5.9)$$

where z is the axial position, t is the time, $N_{i,z}$ is the molar flux of component i in the z direction, C_T is the total molar concentration, $D_{i,j}$ is the diffusion coefficient of component i in component j , y_i is the mole fraction of component i in the gas phase, $N_{j,z}$ is the molar flux of component j in the z direction, n_i is the number of moles of component i , A_{gl} is the liquid/gas interfacial area, n_i^0 is the number of moles in the liquid phase at $t = 0$, γ_i is the activity coefficient of component i , x_i is the equilibrium mole fraction of component i in the liquid phase, y_i^{eq} is the gas phase equilibrium mole fraction corresponding to x_i , P_i^{sat} is the vapor pressure of component i , P is the gas phase pressure.

The activity coefficient (γ_i) quantifies the deviation of components' properties in a mixture from an ideal solution behavior. A solution is ideal if its extensive thermodynamic properties are very close to the sum of its parts. High deviations are expected if the mixture is composed of chemically dissimilar parts, which is usually the case of fragranced mixtures. This way, the well-known UNIFAC method (Poling et al., 2001a) was used here to consider such deviations.

The presented model comprises a set of non-linear coupled partial and ordinary differential equations that must be solved numerically. The numerical solution was implemented in the gPROMs software by dividing the domain into 200 elements using a second-order finite-difference discretization scheme. The time integration was done using the default DASOLV library available in gPROMs. The parameters P_i^{sat} , $D_{i,j}$, P , A_{gl} were taken from Teixeira et al. (2013a).

5.3.2 Odor intensity model

One of the challenges in the design of perfume products is to find precise ways to map changes in physical quantities (increasing the composition of a given component, for example) to changes in the perception of the human olfactory system. In this sense, psychophysics, a scientific field of psychology that investigates how human perception systems respond to physical stimuli, provides valuable tools to help overcome such difficulties Teixeira et al. (2010b). Among the contributions of the field for olfactory perception, two models are emphasized the most in the Perfume field: Power Law (PL) and Odor Value (OV). The OV model establishes a linear relationship between the concentration of a component and the corresponding olfactory perception. The PL is inspired by Stevens's Psychological law (Stevens, 1957) in which a linear relationship between odor intensity and concentration is present on a logarithmic scale. The advantages and drawbacks of each model are detailed in (Teixeira et al., 2010b). The power-law model was used here to map odorant concentration to perceived intensity. It can be written as:

$$\Psi = \left(\frac{C_g}{C_g^{Trh}} \right)^n \quad (5.10)$$

where C_g is the concentration of an odorant in the gas phase, C_g^{Trh} is the threshold concentration in the air, representing the minimum amount required for a human nose to detect the odorant, and n is the exponent.

The concentration of each odorant at a given distance from the liquid source z , and time t can be obtained by solving the equations described which in turn can be mapped to odor intensity Ψ .

The described PL model defines the odor intensity of pure components in the air. However, a typical perfume formulation incorporates several chemicals that pose a new challenge to define the odor features of perfumes. One approach is characterizing the odor of a mixture using the stronger component model (Teixeira et al., 2013a). It states that the odor intensity of the mixture is the maximum intensity among the components. Furthermore, it provides a qualitative measurement of the odor by stating that the odorant with maximum intensity dominates the odor character and can be recognized by the human nose. It can be mathematically expressed as:

$$\Psi_{\text{mix}} = \max(\Psi_i), \quad i = 1, 2, \dots, 3, N \quad (5.11)$$

5.3.3 System Identification

As previously mentioned, this work proposes the identification of a DNN model to represent with fidelity the perfume release and propagation in space and time. The model identification is a crucial step during the identification of an empirical model. Therefore, it should be carefully done. Thus, the identification process was divided into three steps (1) data acquisition procedure, (2) definition of predictor and model structure, and (3) training and performance evaluation procedure.

The data acquisition step consists of the planning of the experiments that need to be done to generate the necessary database. The Latin Hypercube Sampling (LHS) method was employed. The details of this step are given in the data acquisition procedure. The data set containing the information about the perfume release and propagation was obtained by implementing and simulating first principles models – the diffusion model and psychophysical model for olfactory perception identified in Teixeira et al. (2013a). The previously detailed models, as well as the perfume release and propagation process under study were used as a base for the system identification.

The proper definition of the predictor structures to describe dynamic systems is a subject of great interest in the Engineering community. In this work, a non-linear representation of the system dynamics was done using a Non-Linear Output Error (NOE) predictor. The approximating function employed is a Deep Neural Network, specifically, the long short-term memory (LSTM) cell.

In the third step, the parameters of the model need to be estimated/trained. Here, the least-squares estimation is used. Finally, the final model performance has to be assessed. Figure 5.2 summarizes the whole system identification procedure.

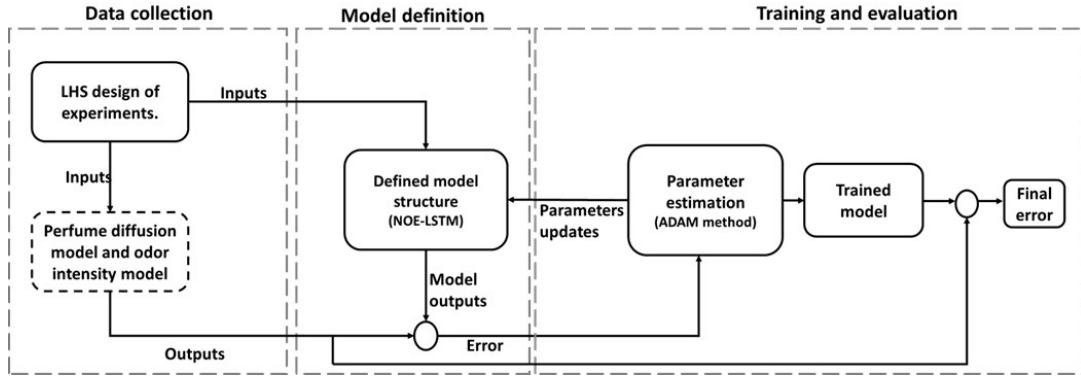


Figure 5.2: Diagram of system identification steps

5.3.3.1 Data acquisition procedure

The choice of input training data may greatly impact the final model performance. Although a vast literature exists on the design of experiments, none offers specific solutions to the many processes that can be identified by empirical means. Thus, this step requires the analyst to have a specific domain knowledge to choose the most appropriate excitation to the process.

In the present work, we use the Latin Hypercube Sampling (LHS) input design method first proposed in McKay et al. (1979). It is a computer algorithm to generate near-random samples of stratified uni/multivariate data drawn from a specified distribution. According to Helton and Davis (2003), it incorporates many of the desirable features of random and stratified sampling and produces more stable analysis outcomes than random sampling. Furthermore, its efficient stratification properties allow for covering the physics or explicable part of the process with a smaller sample size than pure random sampling. The parameters of the LHS input design are the sampling distribution, the number of random variables, the sample size, and the variables' ranges. The variables to be sampled in the present work are the number of moles of odorants per mol of solvent in the liquid. Furthermore, the sample size was chosen here by considering the computational resources needed to train the model. The choice of variables' ranges considered limits useful in the perfume design point of view – pure solvent is an example of meaningless training data in this context.

The other part of data acquisition involves collecting outputs, namely, output variables and their sampling rate. In the study of perfume release and propagation, the primary interest is to have samples of odorant intensities over a long-time horizon and at different distances from the liquid source to characterize the perfume's performance. Thus, the

chosen outputs in the present work are the dynamics of odorants' intensities at several distances from the liquid source. In this work, the sampling rate was set small enough to capture the whole dynamic trajectory of the variables.

The inputs were designed using the `lhsdesign` function from Matlab 2019b. The values were stored in an array and passed as input data to the perfume diffusion and odor intensity models implemented in gPROMs through the Go: Matlab communication software. After every 5 minutes of process simulation, gPROMs returned Matlab the odor intensity of each odorant at several distances from the perfume source.

5.3.3.2 Predictor and model structure

The variables in dynamic systems, such as the process of perfume release and propagation, are generally represented as time series. It means that the output observations at a given time depend on past observations of either input or output. However, traditional machine learning regression algorithms such as linear regression, support regression machines, kernel ridge regressors, and artificial neural networks seek to find a relation between features and the targets regardless of the order of data appearance. Most supervised learning algorithms shuffle the data at every iteration to hasten the training.

Thus, before using a machine learning model, it is necessary to consider a data structure that will introduce time dependencies in such models. It can be done considering lagged input and/or output signals as regressors and additive noise. Two classes of non-linear predictors are emphasized the most in the field of system identification: Non-Linear AutoRegressive Model with Exogeneous Inputs (NARX) and Non-Linear Output Error (NOE) (Nogueira et al., 2018).

An extensive discussion about each of them can be found in Vardanega (2011). In general, the NOE predictor is more appropriate for simulation purposes, which is the case of the present work. The simulation here means that after training, the model can output accurate values of a process dynamic behavior without relying on measured past observations of the output signal. Predictions in the unknown future solely depend on input signals and past predictions. This is also known as a recurrent structure. In mathematical terms, this can be represented as:

$$\hat{z}(t) = f[\hat{z}(k-1), \dots, \hat{z}(k-1-n_a), u(k-d), \dots, u(k-d-n_b+1), \theta], \quad (5.12)$$

$$\hat{y}(k) = \hat{z}(k) + v(k) \quad (5.13)$$

where $\hat{z}(k)$ is the model output prediction, $\hat{y}(k)$ is the output, $u(k)$ is the input signal, and $v(k)$ is the white noise at time k . The function f is a non-linear function parameterized by θ , d is the time delay, and n_b and n_a are the order of the regressors. Figure 5.3 illustrates this relationship.

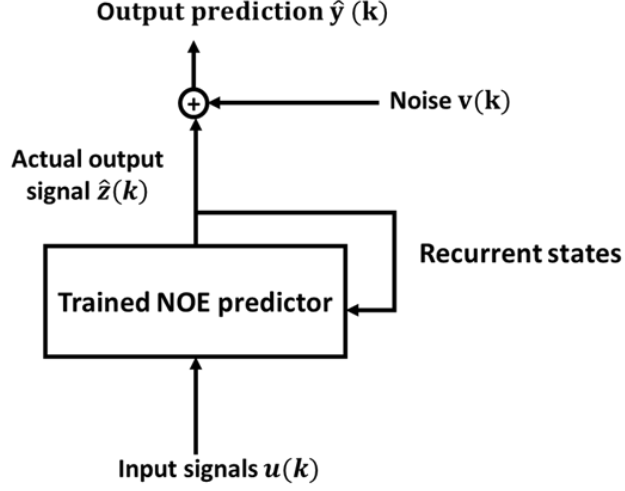


Figure 5.3: Illustration of NOE predictor

5.3.3.3 Long Short-Term Memory

The recurrent non-linear function approximator here is the Long Short-Term memory (LSTM) used in a deep neural network (DNN) model structure. This structure, first proposed by Hochreiter and Schmidhuber (1997), can learn long-time-lag sequence tasks. It was conceived to solve the vanishing gradient problems by enforcing constant error flow through multiplicative gates.

The LSTM computational cell has two variables incorporating past information in new predictions. These variables are called the hidden states. At first, the hidden states are initialized at random or with zero values, and at each time-step computation, they are updated using trainable weight matrices and non-linear functions. Three gates control the updates of these two variables: the update gate, the forget gate, and the output gate. The gates also have trainable weight matrices that define how much information is forward passed over the time steps computations. To illustrate the LSTM structure, the equations for a single forward-time computation of one example can be written as:

$$\tilde{c}(k) = f_c(W_{ca}a(k-1) + W_{cx}X(k) + b_c) \quad (5.14)$$

$$\Gamma_f = \sigma(W_{fa}a(k-1) + W_{fx}X(k) + b_f) \quad (5.15)$$

$$\Gamma_u = \sigma(W_{ua}a(k-1) + W_{ux}X(k) + b_u) \quad (5.16)$$

$$c(k) = \Gamma_u * \tilde{c}(k) + \Gamma_f * c(k-1) \quad (5.17)$$

$$\Gamma_o = \sigma(W_{oa}a(k-1) + W_{ox}X(k) + b_o) \quad (5.18)$$

$$a(k) = \Gamma_o * f_a(c(k)) \quad (5.19)$$

where $*$ is element-wise matrix multiplication, $X(k)$ is the input signal at the time step k . $c(k)$ and $a(k)$ are the hidden states at time step k . The term "hidden" means they generally do not carry the information of main interest but are intermediate values in the computational process. $\tilde{c}(k)$ can be interpreted as the variable that brings fresh information from the current time step k . The update and forget gates Γ_u, Γ_f use the sigmoid function σ to decide how much past and new information is incorporated in the new value of $c(k)$. The output gate Γ_o updates the a state by incorporating information from the updated c and past a . W_{ij} and b_i are the trainable weight matrices and biases respectively, f_i are the non-linear activation functions.

The open-source deep learning programming framework TensorFlow (Abadi et al., 2016) was used in this work. It operates under the high-level Keras API (Chollet, 2015) and counts with highly optimized implementations of recurrent structures, including LSTM.

Choosing an artificial neural network architecture is an iterative process. Even though guidelines exist, the most successful applications of neural networks generally involve trial and error. A common practice is starting with the simplest possible architecture and embodying complexity as more accuracy is needed and data and computational resources become available. Another important aspect is choosing an appropriate predictor for a problem. In this sense, using a recurrent structure for time series simulation is a reasonable choice. Thus, the present work proposes a two-layer DNN structure. The first layer is the LSTM structure with hyperbolic tangent activation. The second is a fully-connected layer with linear activation.

As the main interest is knowing the dynamics of odorants' intensities at different distances from the liquid source, several two-layer neural networks were trained in a way that each can learn the dynamics at a given distance from the perfume source. Thus, it was built as a multiple-sensor structure to assess the evaluation through time and space of the perfume diffusion. The overall sensors' structure is illustrated in Figure 5.4.

5.3.3.4 Training and evaluation

As it is possible to notice, the DNN model has several parameters to be estimated. The estimation of those parameters constitutes the artificial neural network training. Using the outputs of the system, namely the odorants' intensities, which are measured in a continuous numerical scale it is possible to apply the Euclidean distance as fitting criteria in a least-squares estimation function. Thus, the model is fitted when the optimization method finds a set of trainable parameters that minimize the average of the loss computed across the

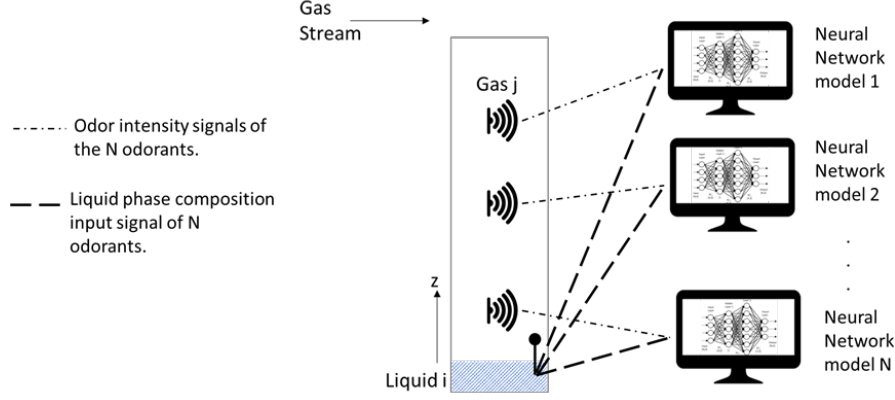


Figure 5.4: Overall predictor structure of perfume release and propagation.

training set. This average loss is also named the cost function, in this case, the mean squared error. In this case, the cost function is given by:

$$J(W_{ij}, b_i) = \frac{1}{2m} \sum_{i=1}^m \sum_{j=1}^{n_y} \sum_{k=1}^{T_y} (\hat{y}_{i,j,k}(W_{ij}, b_i, X_{i,k}) - y_{i,j,k})^2 \quad (5.20)$$

In Equation 5.20, i varies from 1 to the number of training examples m . j varies from 1 to the number of output features n_y , and k varies from 1 to the number of output time-steps T_y . W_{ij} and b_i are the trainable parameters, and X_{ik} is the i^{th} input signal at time step k .

The majority of the popular deep learning frameworks like TensorFlow train the models with gradient-based optimizers by using back-propagation. Among the optimizers, the adaptive moment estimation (Adam) (Abadi et al., 2016) has gained great acceptance in the deep learning community by speeding up several fold the learning process. Thus, this method was used here. Several parameters of the optimization algorithm control the final fitting such as the learning rate, learning rate decay, number of epochs, and mini-batch size. All the mentioned parameters are not trained during cost function optimization but are guessed before training. They are usually called hyperparameters.

Although statistical methods for hyperparameter tuning exist, the most common approach is to use a random grid search. It consists of selecting a finite number of values for each hyperparameter, finding all possible combinations, randomly choosing a subset, training the model, and finding the corresponding values with the best performance. This can be done several times up to a point where one has to manually decide among a few options. The random grid search was used in the present work to search the initial learning rate, learning rate decay, and mini-batch size that achieved the lowest mean squared error. It is important to mention that the proposed hyperparameter tuning assesses which combination has the best performance in a data set not used during the cost function optimization step – the validation set.

The number of neurons is a parameter that defines the artificial neural network architecture and has a significant influence on the model performance. It deserves a separate explanation for not being part of the optimization algorithm but of the model architecture. In a similar way, it needs to be defined before the identification method. In non-linear representations like artificial neural networks, increasing the number of neurons of a layer means increasing the number of learnable parameters. Ideally, one can increase it up to the number of training examples which would ideally lead to a perfect model fitting. However, fitting a model with too many parameters would cause it to lose predicting capacity on data out of the training set (overfitting). Thus, the tuning of the number of neurons in layers must consider the existence of a performance gap between training data and non-training data, which is an indication of overfitting.

Regularization is the most common method to tackle overfitting problems and was used here. Specifically, the early-stopping regularization. The optimal number of neurons in the hidden layer was embodied as part of the random grid search described in the previous paragraph. Finally, the final model performance was assessed by calculating the mean squared error on the same data set in which the hyperparameters were tuned.

The predictor structure, the training process and the hyperparameters tuning were implemented using TensorFlow on the Google Colaboratory environment with Python kernels. Google Colaboratory is a free serverless jupyter notebook environment for prototyping machine learning models on powerful hardware options such as graphical process units (GPUs) and tensor process units (TPUs) (Bisong, 2019).

5.3.4 Perfume performance optimization

The present work uses a model-based optimization with a surrogate model-based approach using a trained DNN to provide an accurate prediction of the system. It seeks to find the optimal composition of a perfume that will result in a specific fragrance while minimizing the solvent odor. In line with the idea of maximizing the odor effect of desired odorants, the objective function that we seek to maximize has the following form:

$$\max_n F_{\text{obj}}(n_1, n_2, n_3, \dots, n_N) \quad (5.21)$$

$$F_{\text{obj}}(n_1, n_2, n_3, \dots, n_N) = \frac{1}{\Delta x \Delta t} \sum_{\substack{i,j=1 \\ i \neq j}}^{N_o} \iint_{t_{\min}, x_{\min}}^{t_{\max}, x_{\max}} (w_{i,i} \Psi_i(t, x, n_1, n_2, \dots, n_N) + w_{i,j} |\Psi_i(x, t, n_1, n_2, \dots, n_N) - \Psi_j(x, t, n_1, n_2, \dots, n_N)|) dx dt \quad (5.22)$$

Subject to:

$$n_1^l < n_1 < n_1^u, \quad (5.23)$$

$$n_2^l < n_2 < n_2^u, \quad (5.24)$$

$$\vdots \quad (5.25)$$

$$n_N^l < n_N < n_N^u \quad (5.26)$$

The objective function is a weighted sum of odor intensity Ψ_i integrals in the spatial dimension z and time t of each odorant i which is sought to compose the perfume aura. N is the total number of components, N_o is the number of odorants, and n_k is the number of moles per mol of the solvent of component k . N_o can be lower than N since perfume formulations can contain odorless components. The integrals in the objective function are numerically approximated by using the trapezoidal rule.

It is important to note in Equation 5.22 that $w_{i,i}$ and $w_{i,j}$ are weight matrices that can assume integer values. They work as a quantitative representation of how desirable or undesirable is the odor of a given component i . They also consider how desirable it is for an odor value to be equal to another, which is represented in the absolute difference between the odorants j and i . Finally, if $w_{i,i}$ is high and positive, for instance, the optimization algorithm will avoid formulations that result in an intense smell of the corresponding component i across time and space, as is the case with alcohol for example. On the other hand, if $w_{i,i}$ is high and negative, the optimization algorithm will search for formulations that lead to an intense odor of the corresponding component i across time and space.

The constraints n_i^l and n_i^u are the lower and upper limits of the decision variables. They coincide with the ranges on which the artificial neural networks were trained. Allowing an optimization algorithm to search on regions outside these limits may cause inaccurate evaluations of the odorants' intensities and the objective function.

The algorithm employed to maximize the objective function 5.22 was the Particle Swarm Optimization (PSO). The optimization implementation using PSO was done using the research toolkit PySwarms (Miranda, 2018) on the Google Colaboratory environment, using Python language. It offers a black-box approach for solving optimization tasks with few lines of code yet allows a flexible API for non-standard swarm models. Here we used the classic global PSO.

5.4 Results and Discussion

5.4.1 Data Acquisition

A quaternary perfume formulation with pinene, linalool, tonalide, and ethanol was considered here as the study case, whose formulation is based on the one presented in Teixeira

et al. (2013a). They formed a simple set of components that play the role of top, middle, base notes and solvent, respectively. This formulation allows to evaluate the methodology proposed in this work while comparing the optimized formulation with Teixeira et al. (2013a) study.

Table 5.1 shows the mole fraction ranges of pinene, linalool and tonalide in the liquid phase used in the input design. The ranges were chosen respecting a heuristic for perfume formulation in which top note should comprise 15% — 25%, middle 30% — 40% and base 45% — 55% (Mata et al., 2005).

Table 5.1: Mole fraction ranges.

Component	Class	Minimum	Maximum
Pinene	Top note	0.01	0.44
Linalool	Middle note	0.06	0.78
Tonalide	Base note	0.03	0.76
Ethanol	Solvent	0.07	0.40

An experiment consists of a set of initial mole fractions of the liquid phase. For the present case, 500 sets of moles fractions were generated using the Matlab lhsdesign function with ranges displayed in Table 5.1. Therefore, the LHS defined 500 experiments to be done.

The univariate and pairwise distributions of the mole fractions are displayed in Figure 5.5. The diagonal contains the histograms. The upper corner contains pair-wise scatter plots, and the lower contains contours pair-wise kernel density plots. It shows that some of the mole fractions distributions are skewed, especially the ethanol. Furthermore, the mole fractions are correlated at a certain extent, which is expected since the sum of mole fractions in each sample must result in one.

After doing the design of experiments, the samples were taken using the phenomenological model as a virtual plant, implemented in gPROMs, which was accessed through Matlab in a software-in-the-loop (SIL) approach. Here, our analysis considered the evolution of odor intensities over time in five distinct positions along the tube: (13, 38, 63, 113, and 163) centimeters. These values coincide with the positions that Teixeira et al. (2013a) used to collect samples from their experimental setup. Furthermore, 600 samples of odorants' intensities were collected at a 5-minute measurement frequency for each of the 500 experiments done. It resulted in a total of 300,000 data points generated for system identification.

Figure 5.6 illustrates the time evolution of odorants' odor intensities at five distinct positions in the tube for one of the 500 experimental conditions generated by Latin Hypercube Sampling (LHS). In this experiment, the initial compositions of the liquid phase were: 0.02 - pinene, 0.70 - linalool, 0.17 - tonalide and 0.11 - ethanol. It is possible that tonalide has the lowest odor intensity among odorants. This is verified in all 500 experiments since tonalide has a considerably smaller vapor pressure than the other components. However, it plays an important role in the formulation by acting as a fixative.

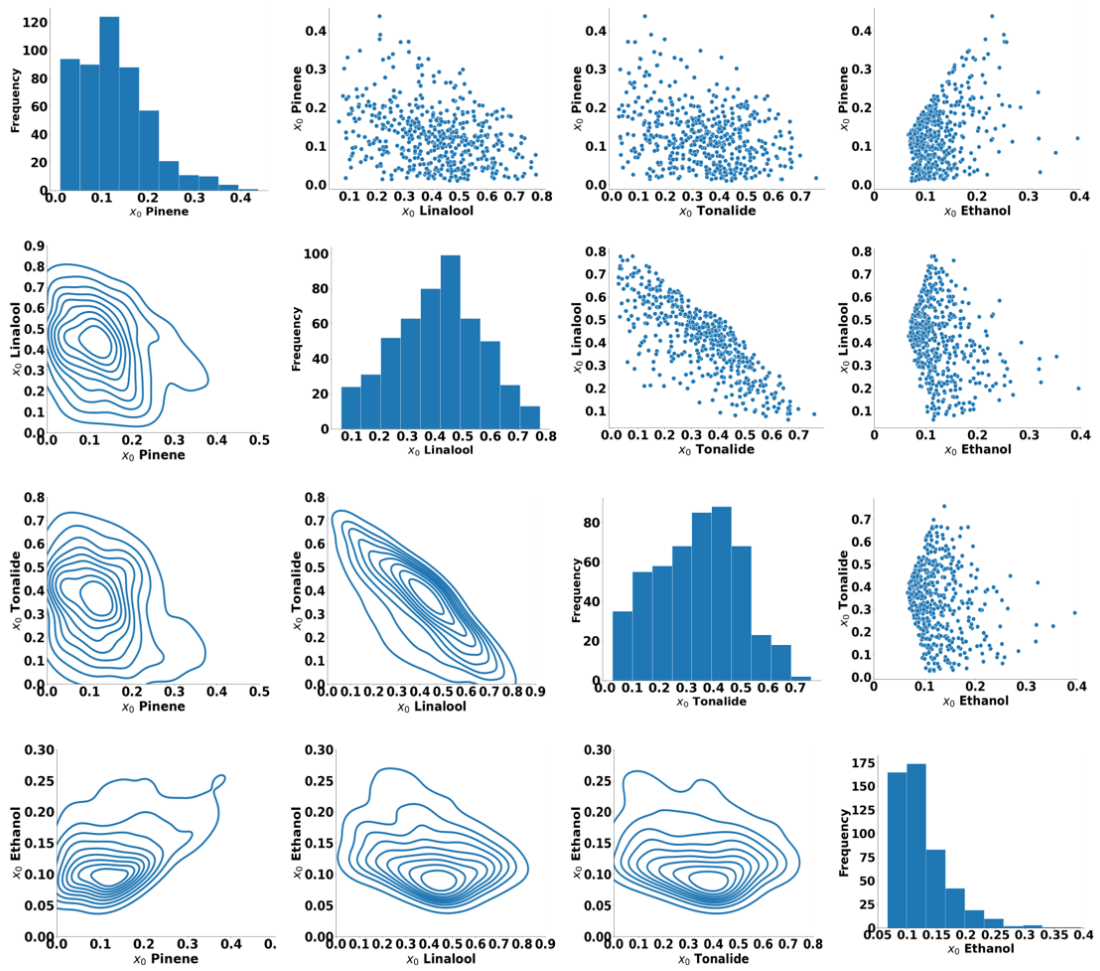


Figure 5.5: Distributions of mole fraction inputs.

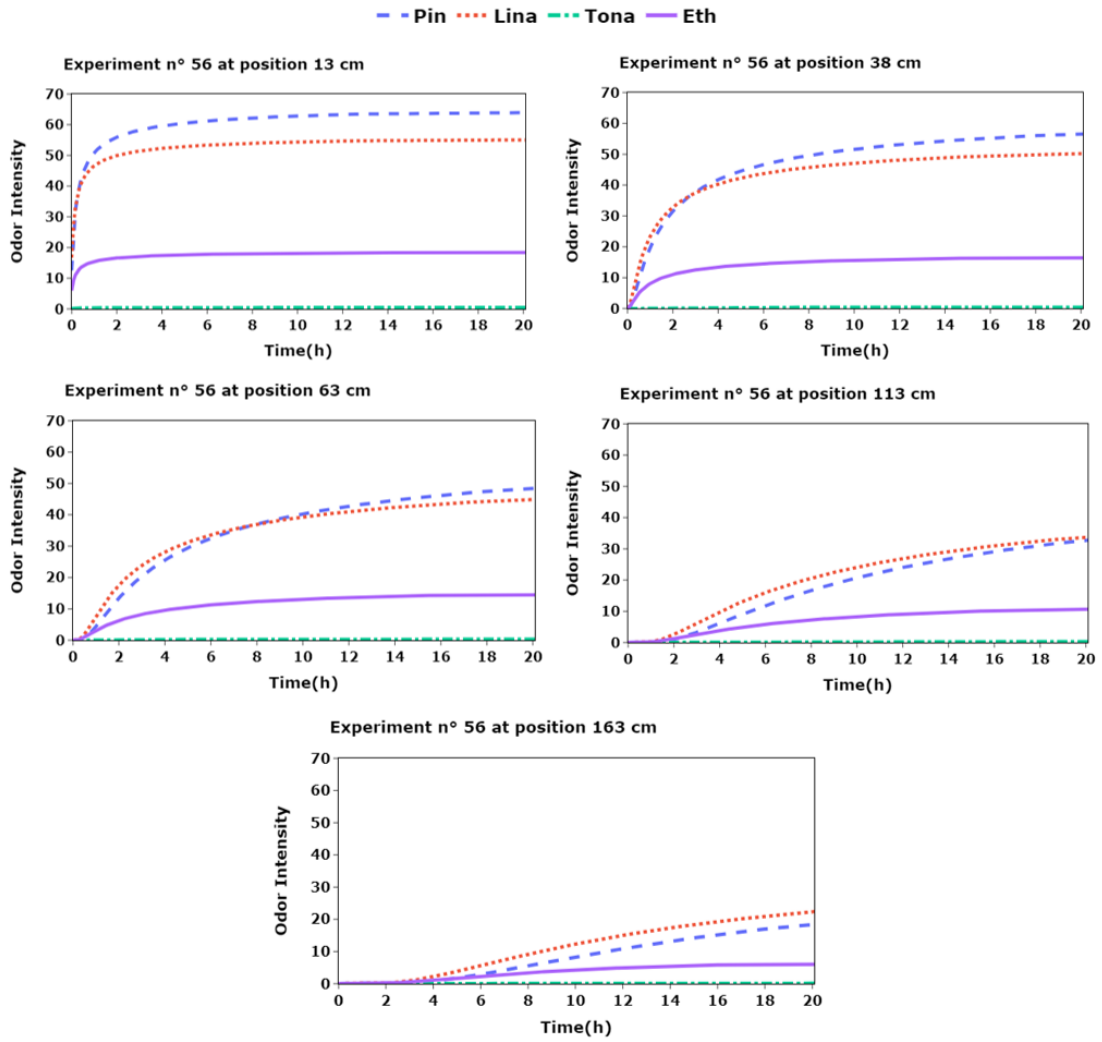


Figure 5.6: Time evolution of odorant's intensities at five distances from the liquid perfume source for experiment number 56.

5.4.2 Training and evaluation

As mentioned, identifying a machine learning model is an iterative process. According to the described methodology, five DNNs were trained to predict the dynamic of odor intensities at each distance from the perfume source. First, the data from data acquisition were split into two sets – 400 experiments from training and 100 for hyperparameter estimation and performance evaluation.

One common and useful step before training is data preprocessing. The inputs (also named features) in typical machine learning applications are measured in different magnitude orders; this difference can affect the training performance and model quality. They were scaled by using maximum and minimum values from the training set. In this way, in the worst case, scaling does not affect the training speed. Similarly, the outputs were scaled between 0 and 1 by estimating maximum and minimum values from the training set.

As shown in Figure 5.6, the odor intensities of the components at a given distance may have very distinct ranges. In this case, it can cause vanishing or exploding gradients, making the training unfeasible. The methodology previously described was followed to find a combination of hyperparameters that resulted in the minimum mean squared error. Table 5.2 shows the overall search space of the hyperparameters: initial learning rate, learning rate decay factor, neurons, and mini-batch size.

Table 5.2: Hyperparameter search space.

Initial learning rate	Learning rate decay ratio	Neurons in the hidden layer	Mini-batch size
5.00×10^{-5}	5.00×10^{-1}	20	1
1.58×10^{-4}	3.33×10^{-1}	40	2
5.00×10^{-4}	2.50×10^{-1}	60	8
1.58×10^{-3}	1.67×10^{-1}	80	32
5.00×10^{-3}	-	-	-

This small set of values leads to a total of 750 distinct combinations. Evaluating all the possibilities is time-consuming and ineffective – one would end up training several models to discover only that just a few combinations lead to satisfactory results. In this way, only 5% of all combinations were equally likely drawn without replacement. This resulted in 40 combinations of hyperparameters for each of the five DNNs – 200 train procedures. The networks were trained for 350 epochs with early-stopping regularization and patience 30 to prevent overfitting. Furthermore, the learning rate scheduler comprised a maximum of 10 discrete drops with several decay ratios, as shown in Table 5.2.

Table 5.3 shows the results of hyperparameter tuning for combinations with the three best performances for each of the five networks. A considerable improvement was attained with a maximum mean squared error of 1.65×10^{-6} for the five DNNs. Furthermore, the results showed that 80 neurons in the LSTM layer and the mini-batch size 2 comprise the best combination for these two parameters. The best combination of the initial learning

rate and learning rate decay ratio is less clear, but 5×10^{-3} for the first and 2.50×10^{-1} for the second seem the most promising.

Table 5.3: Results of hyperparameter tuning with random grid search.

DNN for distance 13 cm					
Initial learning rate	Learning rate decay ratio	Neurons in the hidden layer	Mini-batch size	Validation mse	Rank
5.00×10^{-3}	2.50×10^{-1}	80	2	1.19×10^{-6}	1
5.00×10^{-3}	1.67×10^{-1}	80	1	1.85×10^{-6}	2
5.00×10^{-3}	2.00×10^{-1}	60	1	2.04×10^{-6}	3
DNN for distance 38 cm					
Initial learning rate	Learning rate decay ratio	Neurons in the hidden layer	Mini-batch size	Validation mse	Rank
5.00×10^{-3}	2.50×10^{-1}	80	2	1.20×10^{-6}	1
5.00×10^{-3}	2.00×10^{-1}	60	1	2.06×10^{-6}	2
5.00×10^{-3}	1.67×10^{-1}	80	1	2.10×10^{-6}	3
DNN for distance 63 cm					
Initial learning rate	Learning rate decay ratio	Neurons in the hidden layer	Mini-batch size	Validation mse	Rank
1.58×10^{-3}	2.50×10^{-1}	40	2	1.54×10^{-6}	1
1.58×10^{-3}	2.00×10^{-1}	40	2	1.63×10^{-6}	2
5.00×10^{-4}	5.00×10^{-1}	60	2	1.79×10^{-6}	3
DNN for distance 113 cm					
Initial learning rate	Learning rate decay ratio	Neurons in the hidden layer	Mini-batch size	Validation mse	Rank
5.00×10^{-4}	3.33×10^{-1}	80	2	1.65×10^{-6}	1
1.58×10^{-3}	5.00×10^{-1}	60	1	1.80×10^{-6}	2
5.00×10^{-3}	2.50×10^{-1}	80	2	1.97×10^{-6}	3
DNN for distance 163 cm					
Initial learning rate	Learning rate decay ratio	Neurons in the hidden layer	Mini-batch size	Validation mse	Rank
5.00×10^{-3}	2.50×10^{-1}	80	2	1.43×10^{-6}	1
1.58×10^{-3}	5.00×10^{-1}	60	1	1.50×10^{-6}	2
5.00×10^{-3}	1.67×10^{-1}	80	1	1.65×10^{-6}	3

Once the optimal number of neurons and hyperparameters were found, all the DNNs were retrained with the best hyperparameters to be used in the perfume performance optimization. Figure 5.7 shows a visual representation of the fitting quality achieved by the DNNs for a given experiment in the validation set. The lines represent the predictions of perfume diffusion and odor intensity models implemented in gPROMs and markers represent the output of the DNNs.

In Figure 5.7, it can be seen as a good fitting for pinene, linalool, and ethanol. However, the fitting quality is not as clear for tonalide due to the very distinct scales. Thus, a separate visual is displayed in Figure 5.8 for two distances (38 and 163 cm) to show that similar predicting performance is achieved for tonalide.

One of the advantages of using a DNN as a surrogate model in the optimization step is the required CPU time to simulate the whole dynamic of odorants' intensities. While gPROMs takes 6s to simulate one experiment, TensorFlow 2.3 takes about 0.16 ms on an Intel i7-7500U, 2.7 GHz CPU with little sacrifice on prediction accuracy (the validation mean squared error is around 1.00×10^{-6}). For instance, running 500 experiments in the phenomenological model takes 50 minutes, while using the DNN model takes 0.08 seconds. Considering an optimization scenario where tens of thousands of experiments are necessary, it becomes a noticeable difference. The data-driven approach is time-demanding at the training stage, but predictions are delivered quickly and accurately afterward.

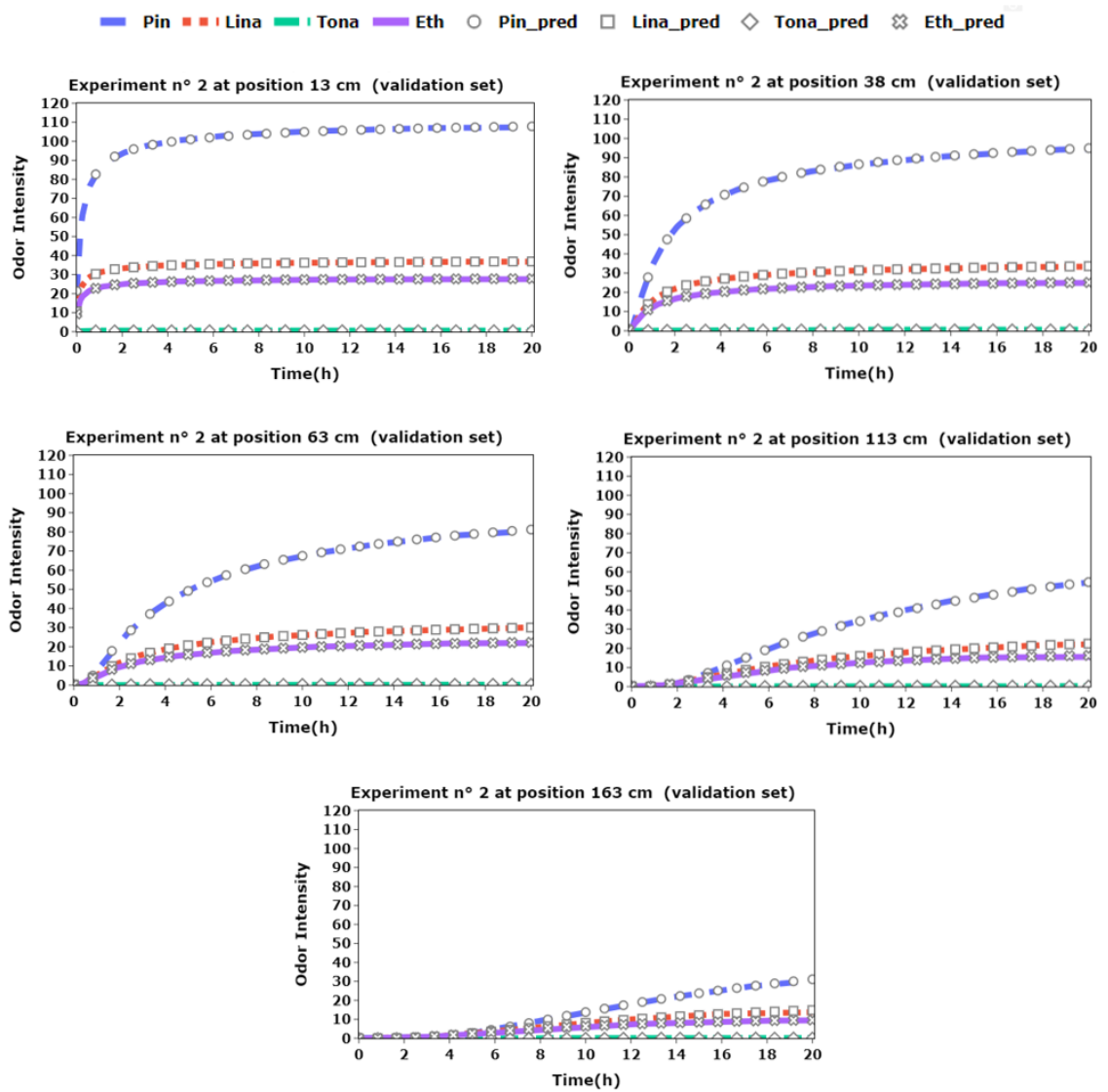


Figure 5.7: DNN models validation.

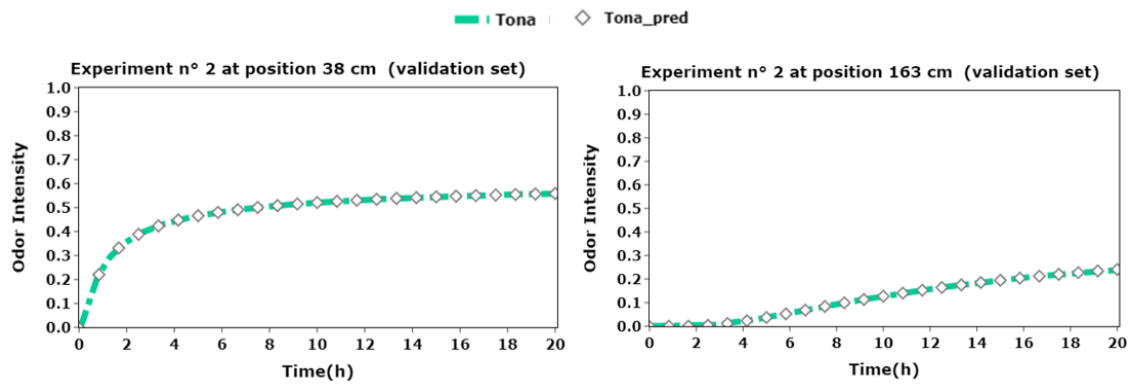


Figure 5.8: DNN models validation for tonalide.

5.4.3 Perfume performance optimization

As explained in the methodology, the proposed objective function and constraints require input weights that reflect how desirable is the odor of the fragrances and their relative discrepancy. In the present work, one of the main objectives is reducing ethanol odor intensity so it does not dominate the odor characteristic since it is irritant and not appreciated by customers (Teixeira et al., 2013a). Moreover, it is also desired to maximize the odor intensities of both top and middle notes (pinene and linalool), yet not allowing too much discrepancy between them.

Table 5.4 shows the weights that reflect our objectives, which can be expressed as: develop a perfume that will provide an equivalent odor of pinene and linalool through space and time while minimizing as much solvent odor as possible.

Table 5.4: Weights for the proposed objective function to optimize perfume performance.

	Pinene	Linalool	Tonalide	Ethanol
Pinene	-1	+10	0	0
Linalool	-	-1	0	0
Tonalide	-	-	-1	0
Ethanol	-	-	-	+1

The proposed optimization problem seeks to maximize the odor value of pinene and linalool, keeping their values close to each other and minimizing the odor value of Ethanol. Therefore, it produces a perfume with pine forest and floral scents using the proposed objective function and the weights in Table 5.4. The optimization was carried out using the PSO algorithm, with a swarm size of 150, particle velocity scaling factor of 0.5, scaling factor to search away from the particle’s best position of 0.5, scaling factor to search away from the swarm’s best position of 0.5 and 90 iterations. The values of the decision variables at the optimal point, namely, the number of moles per mole of ethanol, are 0.17, 7.00, and 0.86 for pinene, linalool, and tonalide, respectively. This corresponds to a mole fraction of 0.02, 0.77, 0.09, and 0.11 for pinene, linalool, tonalide and ethanol, respectively.

Figure 5.9 displays several frames of the perfume radar to show the time evolution of optimized perfume performance. Each frame is a snapshot of odor intensities over time. The radial axis encodes the odor intensity, while the angular axis corresponds to the component names. The polygons connect the components’ odor intensities for a given distance – the higher the perimeter of the polygon at a given time, the smaller the distance from the perfume source. At the beginning of the release (0.083 hours), the components had a detectable smell only at a distance of 13 cm with smell predominance of linalool. At about 0.5 hours, only at distances 13 and 38 cm, the odorants have a significant odor intensity – the first with a well-balanced linalool-pinene smell and the latter with a slight linalool predominance. At 1 hour, at a distance of 63 cm, a blend of linalool, pinene, and ethanol smell appears but with the dominance of linalool smell. At a distance of 38 cm, linalool smell dominance attenuates, and at 13 cm, pinene slightly dominates the

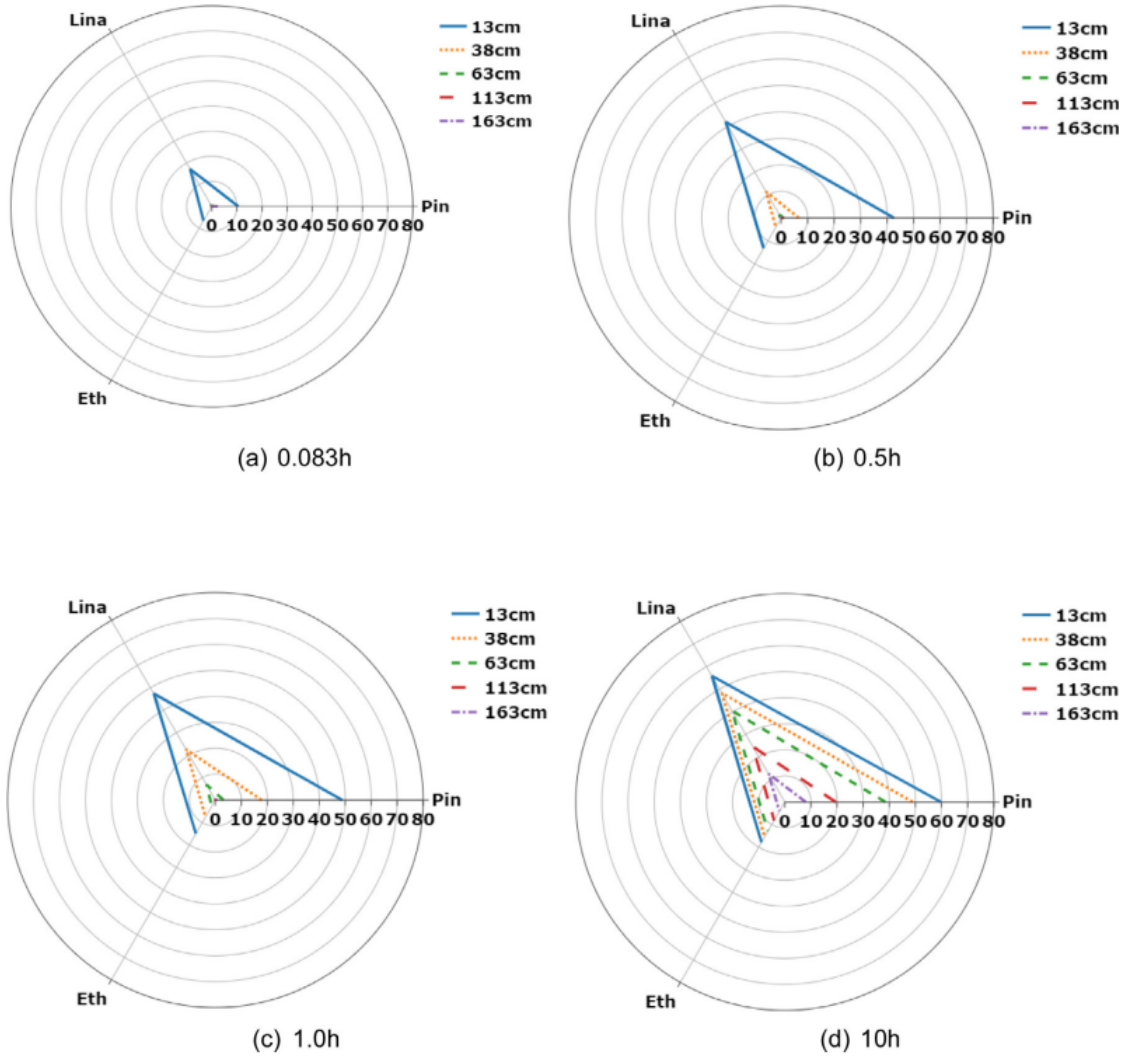


Figure 5.9: Odor intensity radar plot for optimized perfume performance at different time instants.

odor characteristics. At 10 hours, all distances have been affected by the perfume with a well-balanced blend of pinene and linalool. Figure 5.9 shows that the proposed objective function could set the perfume performance as desired – the ethanol odor intensity is kept far below pinene and linalool, which have very close intensities across time and space.

The formulations from Teixeira et al. (2013a) with the same four components were used as a reference to compare with the formulation resulting from the present optimization. The referred work aimed to assess the fixative effect of tonalide. Thus, it is important to highlight that this comparison is a mere reference. The mole fractions are 0.27, 0.38, 0.24, and 0.11 for pinene, linalool, tonalide, and ethanol. This composition was given to the DNN models to predict the perfume evolution. It is important to note that these mole fractions are within the training set input distribution that makes DNNs' predictions accurate.

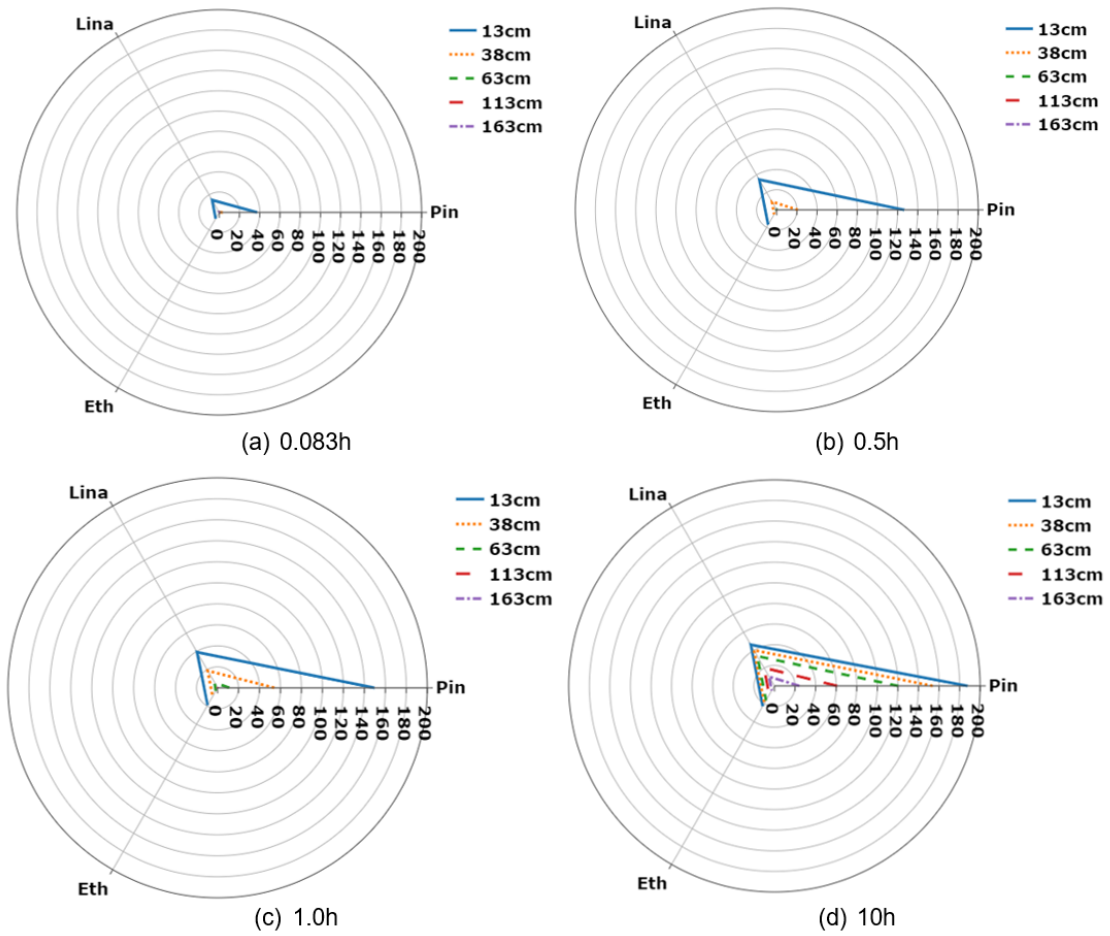


Figure 5.10: Odor intensity plot for (Teixeira et al., 2013a)’s work perfume formulation

Figure 5.10 shows the radar for Teixeira et al. (2013a)’s formulation. It can be seen that ethanol odor intensity does not predominate at any of the distances and time frames – pinene has by far the most intense but highly unbalanced compared to linalool odor intensity.

5.5 Conclusions

Defining, predicting, and designing the performance of fragranced products is still a challenge. Few tools were tested and uncovered in the scientific literature to address such difficulties. In this context, the present work proposes a Deep Learning simulation-optimization approach, identifying a surrogate model that can learn the behavior of perfume release and propagation while correlating it with the perception. The identified model was then used to optimize a perfume formulation composed of a quaternary mixture. The developed DNN models demand considerably low CPU time to predict the full dynamic behavior of perfume release – 0.16 ms against 6s for the first principle model.

This way, a novel systematic methodology for optimal perfume formulation was presented. A single objective function was constructed to encode the peculiarities of a fragrance construction that comprises the question: Which is the perfume composition to attain the desirable odor spectrum of the perfume across time and space? The particle swarm optimization method was used to find the optimal compositions. It is a flexible algorithm in which calculations are fast and are only limited to objective function evaluations. Thus, our work proposes using a trained recurrent neural network surrogate model since it demands considerably lower CPU time than the first principles model. The PSO method maximized the proposed objective function and found the perfume formulation that precisely reflected our desire: minimize ethanol odor intensity and allow the top and middle notes (pinene and linalool) to be the predominant odors with little discrepancy across time and space.

Chapter 6

PUFFIN: A Path-Unifying Feed-Forward Interfaced Network for Vapor Pressure Prediction

6.1 Abstract

The previous chapter outlined a deterministic model for perfume diffusion, assuming known parameters for simulation. However, obtaining accurate thermodynamic data like vapor pressure is a bottleneck, especially for complex perfumes with over 40 components. The issue is compounded when temperature-dependent data is needed. To tackle this, we present PUFFIN (Path-Unifying Feed-Forward Interfaced Network), a machine learning model integrating transfer learning with a specialized inductive bias node based on the Antoine equation for enhanced vapor pressure prediction. PUFFIN surpasses alternatives by exploiting inductive bias and using transfer learning via graph embeddings. This approach has applications beyond fragranced products, serving as a versatile tool for accurate thermodynamic predictions.

6.2 Introduction

Predicting vapor pressure, particularly its temperature-dependent behavior, plays a central role in various industrial and environmental contexts. From refining operations in the petrochemical industry to projecting pollutant spread in environmental sciences, vapor pressure quantification is a foundational underpinning. However, due to chemical compounds' vast diversity and unique properties, its accurate measurement remains an arduous and resource-intensive task that becomes more intractable when temperature dependence is considered.

Traditional techniques necessitate specialized and often costly equipment and time-consuming procedures to ensure the accuracy of vapor pressure measurements (Goll and

Jurs, 1999). Furthermore, obtaining reliable experimental vapor pressure data for many compounds can be challenging, particularly for compounds with normal boiling points above 200 °C, due to a scarcity of data (Yaffe and Cohen, 2001). This reality presents a barrier to systematically exploring new applications for chemical molecules, particularly in high-throughput screening, a technique widely used in drug discovery and materials science. Added intricacies stem from compound availability, stability, and potential safety hazards. Therefore, there is a burgeoning interest in developing alternative methodologies, specifically computational methods based on machine learning, to predict vapor pressure with a high degree of accuracy (Khosravi et al., 2018; Lansford et al., 2023).

Historically, a few methods have attempted to address these challenges. For instance, Tu (1994) leveraged a group contribution method for predicting vapor pressure at a fixed temperature. Later, Zang et al. (2017b) introduced an open-source quantitative structure-property relationship (QSPR) workflow to predict various physicochemical properties, including vapor pressure. Their approach aimed to facilitate high-throughput screening by proposing a predictive model for thermodynamic properties that surpassed the accuracy of the original EPI Suite models for five out of the six predicted properties. This marked a significant step forward in demonstrating the potential of artificial intelligence (AI) tools in vapor pressure prediction. Furthermore, Zang et al. (2017b) performed a statistical analysis, suggesting a strong correlation between the vapor pressure and the boiling point, evidence that will be further explored in this work in a transfer learning context. In Ohe (2019), a method was proposed to predict vapor pressure by directly determining the constants of the vapor pressure equation from boiling point data. The approach is rooted in the observed linear relationship between the logarithm of measured vapor pressure and the reciprocal value of absolute temperature across homologous series. Thus, by utilizing the slope of this relationship for a homologous substance, the intercept of the vapor pressure equation for a substance whose vapor pressure has not been measured can be estimated from its boiling point data.

Developing predictive models for vapor pressure estimation based on molecular structure is a critical task, previously undertaken in vapor-liquid equilibrium studies to ascertain activity coefficients using group contribution methods. Specifically, Nannoolal et al. (2008) introduced a methodology for predicting vapor pressure of non-electrolyte organic compounds, leveraging group contributions and inter-group interactions derived from manually curated molecular groupings

Accurately predicting sought-after properties, particularly when data is limited, remains a significant hurdle across diverse scientific and engineering fields. In these contexts, transfer learning has emerged as an invaluable technique, given its capacity to repurpose a model engineered for one task as the foundation for another task (Sanchez-Lengeling et al., 2019a). This approach has gained increasing interest where pre-trained models are usually used as the starting point for computer vision and natural language processing tasks (Zhuang et al., 2021; Vermeire and Green, 2021). Applying transfer learning can be

particularly beneficial in our specific context when we aim to predict a property with limited data availability. Thus, the premise of transfer learning is grounded in the assumption that the features learned from the task with abundant data are relevant and useful for the task with limited data. This assumption holds especially well when the properties are physically related. For instance, compounds’ boiling point and vapor pressure are linked to the strength of the intermolecular forces. Consequently, they could be expected to have related feature representations. We can leverage the abundant data for one related property to train an initial model and transfer the learned features and patterns to predict the property of interest.

Recent advancements in artificial intelligence and machine learning have further corroborated the capabilities of machine learning models for learning the relationship between molecular structures and properties at high dimensional data (Lansford et al., 2023; Fang et al., 2022), with graph neural networks (GNNs) emerging as a potent approach to tackle these challenges in QSPR modeling (Aouichaoui et al., 2023; Schweidtmann et al., 2020). Using graph theory to represent a compound as a graph, where each node is an atom, and each edge is a bond between atoms, GNNs capture the complex 3D structure of molecules, which is crucial for predicting their properties (Zhang et al., 2020a; Ohe, 2019; Zhang et al., 2022).

In this sense, some works have proposed innovative approaches using machine learning targeting thermodynamic and temperature-dependent molecular properties prediction. Winter et al. (2023) applied a physics-guided machine learning model for thermodynamically consistent activity coefficients. Vermeire et al. (2022) have introduced a model to calculate the solid solubility in organic solvents at different temperatures. Chung et al. (2022) proposed a group contribution method and a machine learning model to predict the Abraham solute parameters, as well as a machine learning model to predict solvation free energy and enthalpy at 298 K. Di Caprio et al. (2023) introduced hybrid modeling to formulate a thermodynamically consistent machine-learning based prediction of activity coefficients.

To further exploit these advancements, this chapter introduces PUFFIN (Path-Unifying Feed-Forward Interfaced Network), a novel machine learning framework. PUFFIN integrates transfer learning and inductive bias to improve prediction accuracy and reliability. Specifically, transfer learning leverages the graph embeddings from a pre-trained GNN model. These graph embeddings, which encapsulate meaningful representations of molecules and their structural relationships (Sanchez-Lengeling et al., 2019a), serve as inputs to a modified feed-forward neural network (FFNN) within the PUFFIN framework. This approach allows the FFNN to benefit from the GNN’s understanding of complex relationships and patterns within the molecular graph data, thereby enhancing the accuracy of vapor pressure predictions. Additionally, PUFFIN incorporates inductive bias, ensuring the network output conforms to the functional form of known physical laws relevant to vapor pressure. This integration reinforces the model’s capability to capture the underly-

ing physical phenomena governing vapor pressure. To properly assess the advantage of the transfer learning approach, a Graph Neural Network (GNN) model capable of predicting vapor pressure across various temperatures was implemented. This model is designed to take in dual inputs: the molecular structures represented as feature and adjacency matrices and the corresponding temperature values.

The introduction of PUFFIN marks a significant stride towards an effective methodology for predicting temperature-dependent vapor pressure rooted in a strong theoretical foundation. PUFFIN presents a comprehensive framework for accurate and interpretable predictions by combining the strengths of GNNs, transfer learning, domain knowledge, and inductive bias. By ensuring alignment with known physical laws through an internal hybridization of the neural network structure, PUFFIN enriches the existing array of computational models and fortifies the bridge between machine learning and domain-specific knowledge of physical phenomena. This holds promising implications for the future of scientific research and practical applications in industries reliant on accurate vapor pressure prediction.

6.3 Methodology

The methodology section in this chapter focuses on implementing three distinct methodological paths, each of which plays a crucial role in the overall analysis. Figure 6.1 serves as a visual representation of these paths, with each path being highlighted in the figure using its respective color. The first three paths are introduced as benchmark tests, serving as reference points for comparison with the novel approach proposed in the chapter. The subsequent subsections of the methodology will delve into detailed descriptions of each path, providing a comprehensive understanding of their individual contributions and how they relate to the central focus of this research.

First, we present a methodology that focuses on creating molecular descriptors based on a specific target property - the boiling point at atmospheric pressure. To achieve this, we leverage the power of graph neural networks, which allow us to find a mathematical representation of the molecules guided by the desired property. This approach enables us to capture the intricate relationships and structural patterns within the molecules. The hypothesis underlying this approach is that it can create more accurate and informative descriptors. This hypothesis is tested against a generic descriptor. It is important to highlight that when predicting vapor pressure using the estimated molecular embeddings, it is crucial to incorporate temperature data as an input to the model. For each molecule, we scale the temperature value with train data mean and standard deviation and concatenate it with the embedding. This combined data then serves as the input for the feed-forward neural network.

Second, we introduce an alternative approach that uses a generic descriptor of the

molecules. This descriptor utilizes logical rules to obtain a numerical molecular representation as a binary vector. This contrasts with the embedding, a vectorial representation from a neural network trained for another related task. By comparing these two descriptor types, we can assess the potential benefits of utilizing embeddings obtained from a trained model that predicts a property similar to the target one (such as boiling point).

Third, a Graph Neural Network (GNN) model capable of predicting vapor pressure across various temperatures was implemented to properly assess the advantage of the transfer learning approach. This model is designed to take in two inputs: the molecular structures encoded as graphs and the corresponding temperature values. Figure 6.1 depicts the architecture of this approach.

Lastly, we present our main proposal, which centers around integrating graph descriptors and knowledge reinforcement in a single framework. This novel methodology aims to enhance the predictive capabilities of neural networks to a specific property. By leveraging the unique strengths of task-specific descriptors and incorporating knowledge reinforcement techniques, we strive to develop a robust and accurate neural network model for property prediction. As the molecules' estimated embeddings from boiling point enter the feed-forward neural network, they remain constant. These embeddings do not undergo additional fine-tuning for vapor pressure prediction following the estimation of boiling points. The parameters that can be adjusted are situated within the feed-forward neural network's parameters.

The benchmark methodologies discussed—including Morgan Fingerprints for vapor pressure prediction, Graph Convolutional Networks (GCN) with Antoine equation reinforcement, and boiling point embeddings for vapor pressure estimation—are based on established quantitative structure-activity relationship (QSAR) protocols within cheminformatics.

6.3.1 Graph Neural Networks and Transfer Learning

In this study, we have employed a two-tier approach to model and predict the vapor pressure of chemical compounds. The first stage utilizes a graph neural network to predict the boiling point of the compounds, a property typically more accessible in data terms than vapor pressure. This GNN then serves as a benchmark model whose generated embeddings are transferred to a feed-forward neural network (FFNN) to predict the vapor pressure. The principal idea is to evaluate the accuracy of our proposed machine learning framework—the Path-Unifying Feed-Forward Interfaced Network—compared to the two-stage GNN and FFNN model. The transfer learning principle presented here will also be used within the PUFFIN framework. Figure 6.1 provides a concise visual depiction of the comprehensive framework underlying this approach, highlighted in blue.

A GNN is well-suited for the task of representing molecules as it excels at handling graph-structured data. This hypothesis will be verified in this chapter by comparing the GNN results with the outcome obtained by employing a generic descriptor (the Morgan

- Path – unifying feed – forward interfaced network for vapor pressure prediction (PUFFIN without transfer learning)
- Path – unifying feed – forward interfaced network for vapor pressure prediction (PUFFIN with transfer learning from boiling point)
- GNN transfer learning for boiling point for vapor pressure prediction
- Morgan fingerprint to neural network for vapor pressure prediction

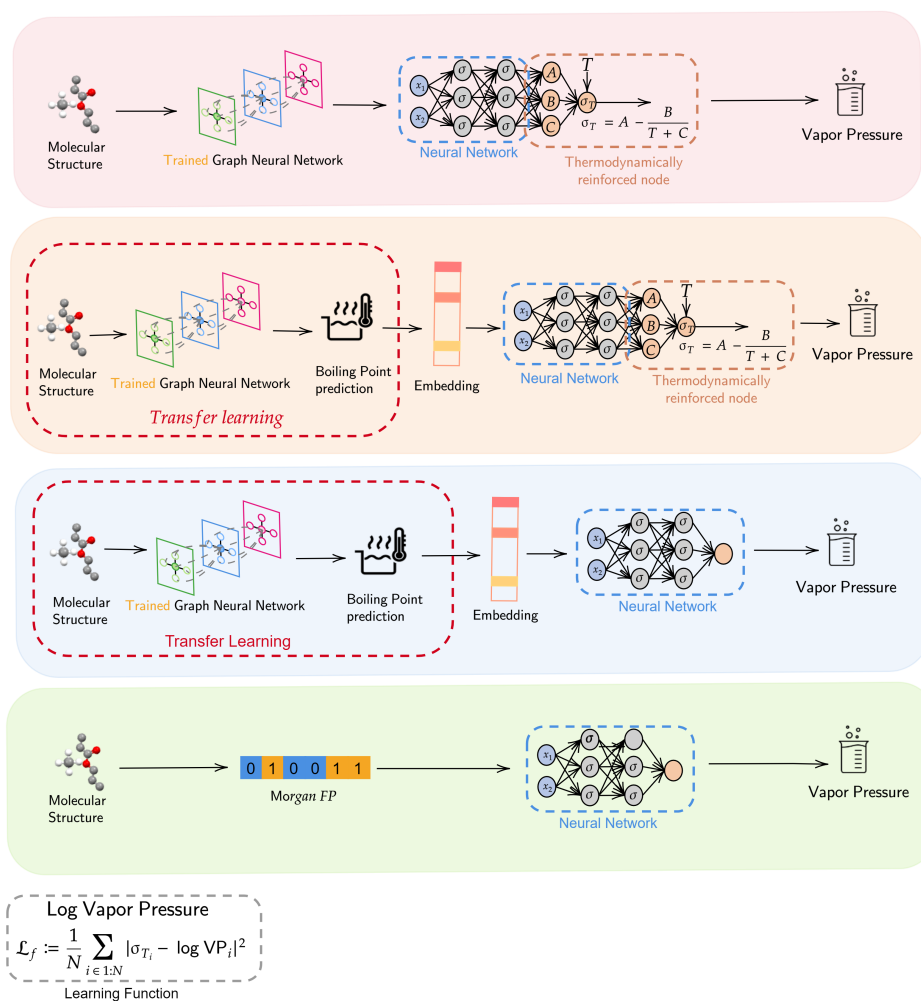


Figure 6.1: Schematic diagram of the methodology presented in this work

fingerprint). In this context, the molecules of the compounds are modeled as graphs where atoms are nodes and chemical bonds are edges. These GNNs are capable of capturing the local and global relationships between atoms in a molecule, which is essential for the accurate prediction of physicochemical properties. Mathematically, a GNN works by iteratively updating the representations (or embeddings) of each node based on the representations of its neighbors, following a procedure often referred to as message passing or neighborhood aggregation:

$$h_i^{(l+1)} = f_h^{(l)} \left(h_i^{(l)}, \text{Aggregate}^{(l)} \left(\{h_j^{(l)} : j \in N(i)\} \right) \right) \quad (6.1)$$

Here, $h_i^{(l)}$ denotes the feature vector of node i at layer l , $N(i)$ is the set of its neighboring nodes, and $f_h^{(l)}$ and $\text{Aggregate}^{(l)}$ are parameterized functions implemented as neural networks. This process is repeated for a fixed number of iterations (or layers), after which the final node representations are used to make the prediction. The aggregation function employed in this work was the one proposed by Kipf and Welling (2017) in their semi-supervised classification with graph convolutional networks. It involves summing the features of neighboring atoms along with self-loops, with normalization carried out by the degree matrix of each node.

Within the transfer learning framework, the generated embeddings from the GNN model serve as inputs to the FFNN model, allowing it to leverage the GNN’s understanding of complex relationships and patterns in the molecular graph data. FFNNs are simple, yet powerful, neural networks that map inputs to outputs through a series of transformations across multiple layers, each followed by a non-linear activation function. The mathematical formulation of a layer in an FFNN is as follows:

$$h^{(l+1)} = f_h^{(l)}(W^{(l)}h^{(l)} + b^{(l)}) \quad (6.2)$$

where $h^{(l)}$ denotes the activations of layer l , $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of layer l , respectively, and $f_h^{(l)}$ is the activation function of layer l . As indicated by the preceding equation, the inputs of the FFNN undergo internal processing to orient toward a new target. In the current context, the GNN-derived embeddings, originally intended for boiling point prediction, are processed by the FFNN to predict vapor pressure. This methodology is backed by evidence in scientific literature pointing to the boiling point as a critical and insightful variable in transfer learning applications aiming at vapor pressure prediction (Zang et al., 2017b; Ohe, 2019). This established correlation between boiling point and vapor pressure reinforces the potential of transfer learning.

6.3.2 Generic Descriptor: Morgan Fingerprints

In our methodology, we also consider another benchmark strategy for evaluating the performance of the PUFFIN framework. This alternative approach bypasses the use of inductive

bias and transfer learning, instead directly predicting the vapor pressure of compounds. At the heart of this strategy is using Morgan’s fingerprints as generic molecular descriptors (Morgan, 1965). These fingerprints provide a numerical representation of molecules and offer an alternative to the GNN used in our benchmark model and within the PUFFIN framework. The Morgan algorithm starts by assigning each atom in the molecule an initial identifier (usually its atomic number). Then, for each iteration up to a specified radius, it updates the identifier for each atom to a hash of its own identifier and the identifiers of its immediate neighbors. Finally, the identifiers are hashed into the fingerprint bit vector. Figure 6.1 provides a concise visual depiction of the comprehensive framework underlying this approach, highlighted in green.

Mathematically, a molecule’s Morgan fingerprint, $\mathbf{F}$, is a binary vector where each element, f_i , represents the presence (1) or absence (0) of a specific substructure in the molecule. For a molecule with n atoms, the fingerprint is computed as follows:

$$\mathbf{F} = [f_1, f_2, \dots, f_n] \quad (6.3)$$

where f_i is computed based on the connectivity and types of atoms in a circular neighborhood with a given radius around the i^{th} atom.

In more detail, the f_i is determined as:

$$f_i = \begin{cases} 1 & \text{if specific substructure exists around the } i^{th} \text{ atom,} \\ 0 & \text{otherwise.} \end{cases} \quad (6.4)$$

In this strategy, each molecule is represented by its Morgan fingerprint, which serves as input to a feed-forward neural network (FFNN). The FFNN, as a universal function approximator, aims to learn a mapping from these fingerprints to the vapor pressure, similar to our two other approaches.

This benchmark strategy not only serves as a comparative measure of performance but also highlights the efficacy of the PUFFIN framework and illuminates the enhancements gained through the transfer learning over using a direct numerical molecular representation like Morgan fingerprints.

6.3.3 PUFFIN: Path-Unifying Feed-Forward Interfaced Network

Our proposed PUFFIN framework enhances this two-stage model by integrating domain-knowledge reinforcement via inductive bias, ensuring alignment with known physical laws. This methodology aims to develop a robust, efficient, and accurate tool for predicting vapor pressure that extends beyond the capabilities of conventional GNNs and FFNNs. Figure 6.1 concisely depicts the comprehensive framework underlying this approach, highlighted in orange.

The Path-Unifying Feed-Forward Interfaced Network incorporates an inductive bias node—an Antoine equation layer—into the traditional FFNN model, thereby reinforcing

the phenomena within the model and offering a more accurate prediction of the vapor pressure. The Antoine equation is a well-established empirical model for describing the relationship between vapor pressure and temperature for a given substance, often expressed as $\log P = A - \frac{B}{(C+T)}$ (Thomson, 1946), where P represents vapor pressure, T is the temperature, and A , B , and C are substance-specific parameters. The PUFFIN method uses a graph neural network (GNN) embeddings from boiling point data. These embeddings are rich in essential structural and chemical information of the molecules. The resulting fixed-size vector representations, designed specifically to contain features related to the boiling point, are then input into a feedforward neural network (FFNN).

Figure 6.1 demonstrates that the FFNN is configured to accept these pre-determined boiling point embeddings. The network architecture is crafted to yield the variable σ_T , indicative of the vapor pressure calculated by the Antoine equation. In this setup, A , B , and C are coefficients emerging from an intermediate layer of the network, embodying the coefficients of the Antoine equation, with T denoting the input temperature for vapor pressure determination.

In this hybridized FFNN, the Antoine equation serves as the activation function of the output layer. This is an innovative approach that exploits both the neural networks and the domain-specific knowledge encapsulated in the Antoine equation. The bias of this output neuron is considered an exogenous variable, the temperature (T), while its inputs are the outputs of the previous layer. These inputs, therefore, are effectively equivalent to the Antoine coefficients. This design essentially bridges the gap between the high-dimensional latent space represented by the previous layer’s outputs and the real-world physical property to be predicted, the vapor pressure. The output of the Antoine neuron is the predicted vapor pressure, directly incorporating the temperature dependency encapsulated in the Antoine equation. This predicted vapor pressure is then compared with the true value to compute the loss function, which guides the training of the neural network.

By assessing PUFFIN’s performance against the two benchmark models described in the subsections above, we seek to substantiate the advantages of incorporating transfer learning and inductive bias into machine learning models for predicting physical properties. This comparison demonstrates the effectiveness of PUFFIN’s unique design, thereby showcasing its potential in advancing the field of computational prediction of physicochemical properties. The loss propagation through this custom layer is performed using standard trace-based automatic differentiation, suitable for straightforward operations present in the Antoine layer, such as addition, division, and multiplication.

6.3.4 PUFFIN: Assessment of transfer learning efficacy

The original hypothesis of PUFFIN posits that transfer learning plays a vital role in predictive performance. To evaluate this, a complete Graph Convolutional Network (GCN) with an Antoine inductive bias was trained. Adding specific details about the inductive bias

and the training process of the GCN provides a clearer understanding of the methodology and its significance in validating the hypothesis.

This model incorporates three layers of graph convolution, each succeeded by a fully connected layer containing 256 neurons. Subsequent to these layers is a global sum pooling layer. Following this, there are two additional fully connected layers, each comprising 128 neurons. The training of the model employed an early stopping technique, set with a patience parameter of 50. The learning rate during training was established at 1×10^{-4} and a dropout regularization with 25% dropout probability.

6.3.5 Vapor Pressure and Boiling Point Data Sets

The database used in this study consists of 1851 molecules of organic chemicals, encompassing a wide range of classes such as industrial chemicals, antimicrobials, dyes, fertilizers, flame retardants, fragrances, pharmaceutical products, herbicides, pesticides, inert ingredients, petrochemicals, food additives. The Simplified Molecular Identification and Line Entry System (SMILES) system represents these molecules in the database (Weininger, 1988). SMILES employs an alphanumeric character sequence to describe the connectivity of atoms in a molecule. Atoms are represented by their chemical symbols, and their bonds are indicated by special characters such as hyphens and numbers. Additionally, SMILES can incorporate information about functional groups, isomerism, stereochemistry, and electric charge.

To characterize the thermodynamic properties of each molecule, vapor pressures at five distinct temperatures were used: 298 K, which is approximately room temperature; 273 K; an intermediate temperature of 325 K; and two elevated temperatures of 350 K and 375 K. This range was chosen to ensure a thorough coverage under varying thermal conditions. To populate the database with reliable data, initial vapor pressure values at the standard reference temperature of 298 K were obtained from the experimental work published by Zang et al. (2017a). These experimental findings serve as a benchmark for the other values. The vapor pressures at the other specified temperatures were acquired through the *Thermo: Thermodynamics and Phase Equilibrium* component of the Chemical Engineering Design Library (ChEDL), accessed via its Application Programming Interface (API) (Bell and Contributors, 2023).

The dataset, which includes vapor pressure (VP) values, covers a comprehensive list of 1,852 different molecules. For each molecule, the dataset provides vapor pressure measurements at five specific temperatures, enhancing its utility for various scientific analyses. This vapor pressure dataset has been released to the public for wider use. Furthermore, the boiling point dataset, as reported by Zang et al. (2017a), has also been made available to the scientific community via a dedicated GitHub repository.

Creating the training, validation, and test sets from the vapor pressure (VP) dataset involved random shuffling followed by structured partitioning. Initially, 80% of the shuffled data was allocated for training and validation, while the remaining 20% was set aside for

testing. This initial split ensured a distinct test set. The training and validation subset was then further divided, with 64% of the original dataset becoming the final training set, and 16% serving as the validation set. To prepare the target property, the VP values, for more effective machine learning model training, we transformed them by taking the base-10 logarithm.

It is crucial to note that because the dataset includes measurements at different temperatures for the same molecule, a single molecule may appear in multiple sets but at varying temperatures. This approach ensures that the model learns to predict vapor pressures at temperatures it has not explicitly been trained on, enhancing its generalizability.

6.3.6 Hyperparameter Identification

Hyperparameter optimization is an essential part of any machine learning project, and in the development of the PUFFIN model, as well as the benchmark models, we have employed a state-of-the-art hyperparameter optimization method called HYPERBAND (Li et al., 2018a) to identify the optimal hyperparameters. Ensuring that the same method of optimization is used across all models allows for a fair and unbiased comparison of their performance.

The HYPERBAND algorithm, introduced by Li et al. (2018a), is an efficient random search algorithm designed for hyperparameter optimization. It tackles the fundamental challenge of allocating limited computational resources to evaluate numerous hyperparameter configurations. Unlike conventional hyperparameter tuning methods that follow a sequential process, HYPERBAND employs a more dynamic approach, enabling it to explore the hyperparameter space more efficiently.

HYPERBAND operates on the principle of adaptive resource allocation and early-stopping. It begins by sampling many hyperparameter configurations and allocating a few resources to each. After an initial evaluation, only a fraction of these configurations, specifically those showing the most promise, are retained and given additional resources for further evaluation. Mathematically, HYPERBAND employs a multi-armed bandit strategy known as "successive halving" to allocate resources. If n is the number of configurations and r is the initial amount of resources allocated to each configuration, the algorithm iteratively reduces n by a factor of μ (usually set to 3) while increasing r by the same factor until only the most promising configuration is left. This is done over a series of s brackets, each representing a round of resource allocation. By doing this, HYPERBAND eliminates sub-optimal configurations early in the process, allowing it to dedicate more resources to evaluating promising configurations. As a result, it can effectively identify the optimal hyperparameters within a relatively short time and with limited computational resources.

6.4 Results

6.4.1 Molecular Embedding—Estimation From Boiling Point

Morgan fingerprints are a form of generic molecular descriptors that serve to encode molecular structures into a compact vector form. While these descriptors can provide a useful overview of molecular properties, they may not be the most suitable for predicting specific properties like vapor pressure due to their generic nature. An alternative approach could involve designing a more specialized descriptor for the task at hand. For instance, following the methodology presented in the previous section, we trained a separate neural network to predict the boiling point of a compound. The intermediate representation or "latent space" (also known as the embedding) from this boiling point prediction model is then used as a more specialized descriptor for predicting vapor pressure. Unlike traditional neural networks, a GNN is specifically designed to work with graph-like data structures, such as molecular structures, by considering their topological information. The input to the GNN was the molecular graph. During the training process, the GNN learned to transform these molecular graphs into a continuous vector representation, known as an embedding, which effectively captured the essential structural features relevant to the boiling point.

The architecture of our model consisted of three GNN layers. These were followed by a global addition pooling layer, two dense layers with rectified linear unit (ReLU) activation functions, and dropout layers interspersed for regularization. A final readout layer completed the architecture. Before proceeding with model identification, data preprocessing steps were undertaken. DeepChem library (Ramsundar et al., 2019) was used to convert the molecular SMILES strings into graph data. Subsequently, the target variable (boiling point) was standardized; this was achieved by subtracting the mean and dividing by the standard deviation of the boiling points in the training set.

The boiling point (BP) dataset was then partitioned into training and validation subsets, which constituted 80% of the total dataset. These subsets were used for hyperparameter tuning and for scaling operations. The remaining 20% of the dataset was designated as the test set and was reserved exclusively for the final evaluation of model performance. Hyperparameters such as the number of neurons per layer, batch size, and dropout probability were tuned using the ASHAScheduler (AsyncHYPERBANDScheduler). Specifically, the search space for the number of neurons per layer was defined as 2^i , where i could be any value from the set 6, 7, 8, 9. Dropout probability was allowed to vary among 0.2, 0.3, 0.4, providing a balance between model complexity and risk of overfitting. The learning rate was searched in a log uniform space, ranging from 5×10^{-4} to 1×10^{-2} . Finally, the batch size was varied among the set 32, 64, 128.

The optimal model architecture determined from the hyperparameter search is detailed in Table 6.1. Upon establishing the optimal architecture, the model underwent further training for 300 epochs, with an early stopping protocol to mitigate the risk of overfitting.

The final mean squared error for the test set with respect to the rescaled boiling point was 0.0471.

Table 6.1: Architecture summary of identified graph neural network (GNN) for predicting boiling point.

Layer Type	Units	Activation
GNN	128	reLU
GNN	64	reLU
GNN	64	reLU
Global Add	-	-
Dense	128	reLU
Dropout	0.2 probability	-
Dense	128	reLU
Dropout	0.2 probability	-
Dense	1	linear

Figure 6.2 displays a parity plot showing the performance of the identified model on BP test set data. It can be seen that the data points are randomly spread over the identity line, which indicates satisfactory predictive performance.

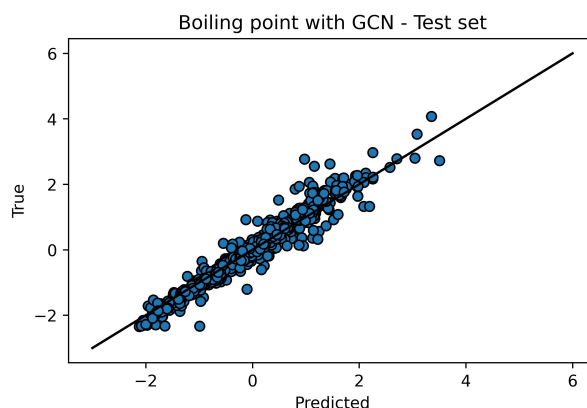


Figure 6.2: Parity plot of graph neural network (GNN) prediction of boiling point compared to the ground truth in the test set

The learned embeddings serve as a more specialized form of the molecular descriptor and learn a representation of the molecular structures in relation to their boiling point, which is intended to be used as input descriptors instead of the more generic Morgan fingerprints for predicting vapor pressure.

6.4.2 Molecular Embedding Without Domain Knowledge Node—Vapor Pressure Prediction

As mentioned, the 128-dimensional learned embeddings are intended to serve as a more specialized form of the molecular descriptor and learn a representation of the molecular structures in relation to their boiling point. As in the baseline model, to determine the

optimal neural network architecture, we used HYPERBAND for hyperparameter search. We allowed for a range in the number of layers (from 1 to 6) and varied the number of neurons per layer (from 20 to 440 with increments of 40 neurons). Additionally, the learning rate was set to vary between 1×10^{-5} and 1×10^{-2} .

Table 6.2 presents the architecture of the identified feed-forward neural network used to predict vapor pressure at different temperatures from the learned embeddings.

Table 6.2: Architecture summary of the identified feed forward neural network for predicting vapor pressure from embeddings.

Layer Type	Units	Activation
Dense	260	reLU
Dense	220	reLU
Dense	180	reLU
Dense	1	linear

Upon establishing the optimal architecture, the model underwent further training for 600 epochs, with an early stopping protocol to mitigate the risk of overfitting. The resultant mean absolute errors for the training and validation sets were measured at 0.2016 and 0.3646, respectively.

Figure 6.3.b shows a visual illustration of the model’s predictive performance in the test set.

As can be seen from test set errors and plots, using embeddings from boiling point GNN predictive model did not produce a better predictive performance for the vapor pressure prediction task than the Morgan fingerprint generic descriptor.

6.4.3 Molecular Embedding With Domain Knowledge Node (PUFFIN)—Vapor Pressure Prediction

Finally, we evaluated the performance of the PUFFIN. To determine the optimal neural network architecture, we used HYPERBAND for hyperparameter search and allowed the same search space as for the other approaches. Table 6.3 presents the architecture of the identified feed-forward neural network used to predict vapor pressure at different temperatures from the learned embeddings.

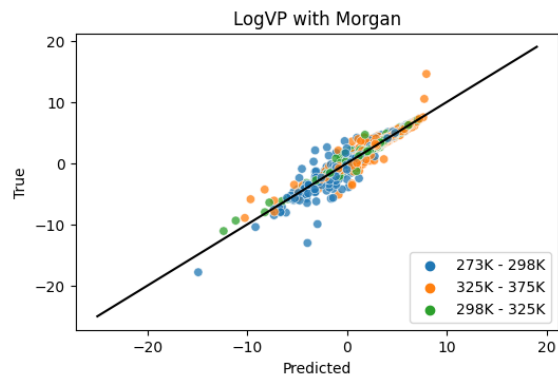
Upon establishing the optimal architecture, the model underwent further training for 600 epochs, with an early stopping protocol in place to mitigate the risk of overfitting. The resultant mean absolute errors for the training and validation sets were measured at 0.1227 and 0.1609, respectively.

Figure 6.3.c illustrates the model’s predictive performance in the test set. As can be seen both from the test set error and from the figures, the performance of PUFFIN reduced the error by about 50%. The residuals are considerably more concentrated around zero and with smaller variances than other approaches. This improvement can be attributed to the inductive bias introduced by using the Antoine equation in an intermediary node of the

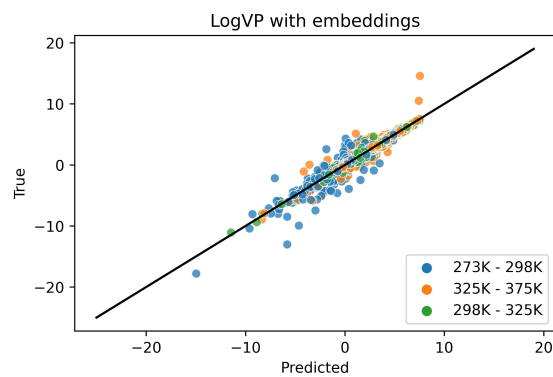
Table 6.3: Architecture summary of the identified feed-forward neural network for predicting vapor pressure from embeddings with domain knowledge reinforcement (the Antoine node).

Layer	Type	Units	Activation
	Dense	180	reLU
	Dense	380	reLU
	Dense	180	reLU
	Dense	420	reLU
	Dense	380	reLU
	Dense	380	reLU
	Dense	3	linear

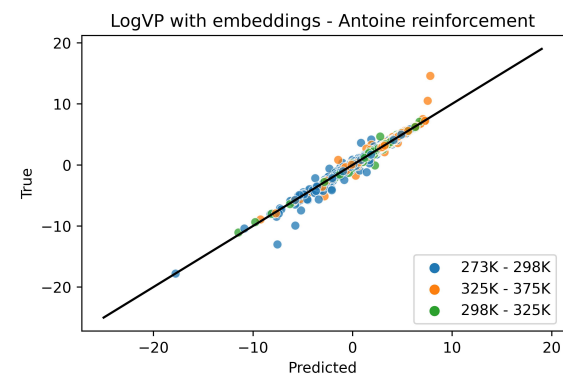
neural network. Within the proposed framework, we encode molecular information and boiling point data into the domain knowledge-reinforced PUFFIN, allowing a considerably better model for predicting vapor pressure.



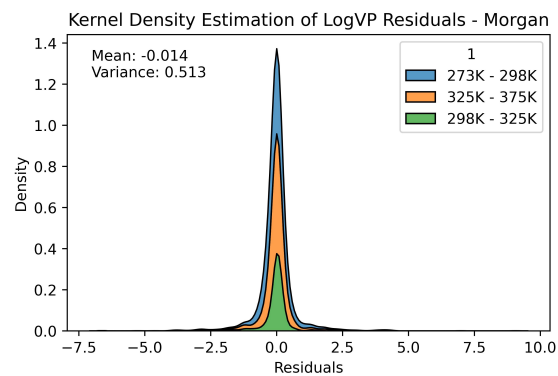
(a) Morgan



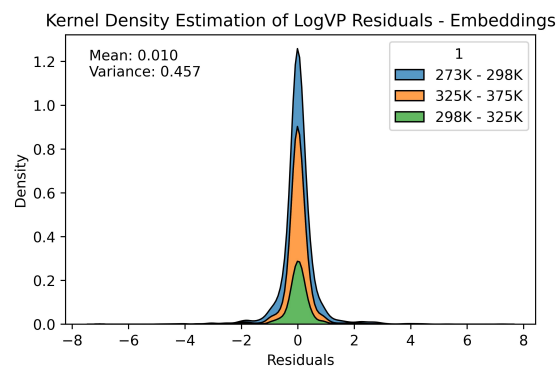
(b) Embeddings



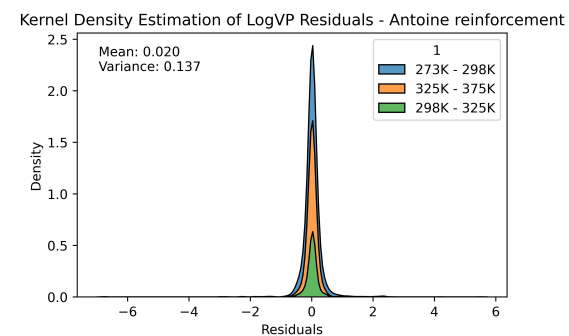
(c) Puffin



(d) Morgan - Residuals' distribution



(e) Embeddings - Residuals' distribution



(f) Puffin - Residuals' distribution

Figure 6.3: Comparative analysis: parity plots and kernel density estimation of residuals, including mean and variance for each model in the test dataset.

Table 6.4 offers a summary of the test set performance for three different approaches: using Morgan fingerprints, using boiling point specialized embeddings, and using boiling point specialized embeddings with thermodynamic reinforcement. The performance is measured in terms of mean squared error (MSE) on the test set.

The analysis of the results indicates that the method employing Morgan fingerprints is the least effective, evidenced by a mean squared error (MSE) of 0.3690. A marginal enhancement in performance is observed when the model utilizes embeddings specialized for boiling points, which yields an MSE of 0.3646. Notably, using a Graph Convolutional Network (GCN) without transfer learning, but supplemented with Antoine Reinforcement, leads to an improvement of nearly 20% in the MSE on the test set.

In Figure 6.4.a, the parity plot for vapor pressure with GCN and Antoine reinforcement shows that most data points cluster closely around the line of perfect agreement, indicating that the model’s predictions correlate well with the true values for the test set. The distribution of points suggests a strong predictive performance, particularly for values closer to zero. However, there is a visible dispersion for higher true values, implying that the model may be less accurate in that range.

Figure 6.4.b displays the kernel density estimation of the logarithm of vapor pressure (LogVP) residuals for the GCN model with Antoine reinforcement. The plot reveals a sharp peak at zero, which implies that most residuals are small and close to zero, signifying accurate predictions. The distribution appears to be symmetrical and tightly centered around the zero line, which suggests that there is no significant bias in the predictions. The tails of the distribution are thin, indicating that there are few large residuals, which is a desirable characteristic in a predictive model.

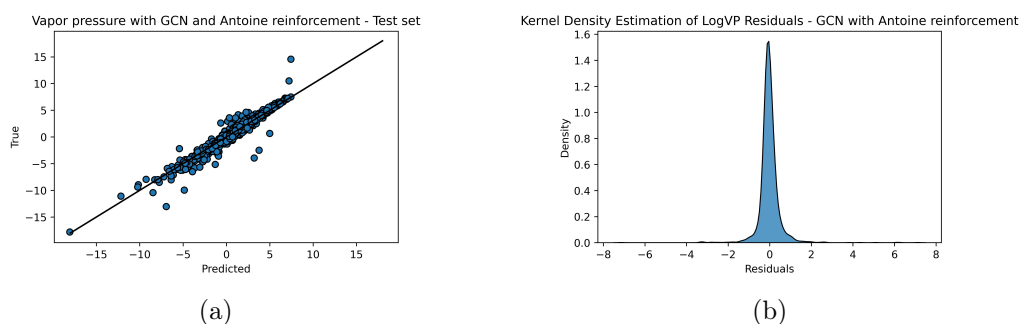


Figure 6.4: (a) Parity plot for GCN with Antoine reinforcement on vapor pressure prediction. (b) Residuals’ Kernel Density Estimation for LogVP predictions by the same GCN model

These findings support the hypothesis that guided graphs can generate more precise and informative descriptors. However, this hypothesis warrants further investigation in future studies to understand its advantages and limitations comprehensively. The most significant improvement is observed when boiling point specialized embeddings are combined with domain knowledge reinforcement—in this case, the Antoine equation. This

approach results in an MSE of 0.1609, more than halving the error obtained using the other methods.

This result illustrates the utility of integrating domain-specific knowledge (in this case, boiling point) and physical laws (thermodynamic reinforcement) into the learning process, leading to enhanced prediction accuracy. These results underscore the potential for specialized embeddings and physics-informed machine learning to be used as predictive models for complex properties like vapor pressure.

Table 6.4: Summary of test set performance for the three approaches evaluated based on Mean Squared Error (MSE) metric.

	Morgan fingerprint	Embeddings without Antoine	GCN with Antoine	Embeddings with Antoine (PUFFIN)
Test set mse	0.3690	0.3646	0.3159	0.1609

6.4.4 Morgan Fingerprints—Vapor Pressure Prediction

Our baseline model was constructed as a feed-forward neural network taking inputs from Morgan fingerprints combined with temperature data. The output from this model was the base-10 logarithm of vapor pressure. To determine the optimal neural network architecture, we used HYPERBAND for hyperparameter search. We allowed for a range in the number of layers (from 1 to 6) and varied the number of neurons per layer (from 20 to 440 with increments of 40 neurons). Additionally, the learning rate was set to vary between 1×10^{-5} and 1×10^{-2} .

To measure the effectiveness of these variations, we used a validation set to compute the mean absolute error, which served as our optimization metric. The training process was conducted with the Adaptive Moment Estimation (ADAM), Kingma and Ba (2015) 's algorithm, batching data in fixed sizes of 64. HYPERBAND was configured to a maximum of 100 epochs and a factor of 3 for the training process. The best architecture, identified by achieving the smallest validation set error, is presented in Table 6.5. Upon establishing the optimal architecture, the model underwent further training for 600 epochs, with an early stopping protocol in place to mitigate the risk of overfitting. The resultant mean absolute errors for the training and validation sets were 0.1663 and 0.3529, respectively.

To evaluate the baseline model's efficacy, the error against a test set was computed. This test dataset had been excluded from both the initial training and the hyperparameter tuning stages. The mean absolute error determined for this test set was 0.3690. To demonstrate the model's predictive proficiency with the test set, Figure 6.3.a illustrates a parity plot and a kernel density estimation of the residuals' distribution, showing that the mean of the residuals is centered around zero with low variance.

Table 6.5: Architecture summary of identified feed-forward neural network for predicting vapor pressure from Morgan fingerprints.

Layer	Type	Units	Activation
Input		1024	-
Dense		340	reLU
Dense		60	reLU
Dense		100	reLU
Dense		1	linear

6.4.5 Interpretability remarks

The intermediary layer of the PUFFIN model, which feeds into the Antoine neuron, produces output parameters A , B , and C . To test the partial interpretability of this structure, these parameters are evaluated against the known Antoine coefficients available at Dortmund Data Bank (www.ddbst.com) for three different components from the test set. Table 6.6 presents the values of the coefficients for three components and the embedding values provided by the PUFFIN.

Table 6.6: Comparison of coefficients predicted from PUFFIN and from literature.

Substance	A_Puffin	A	B_Puffin	B	C_Puffin	C
Acetone	9.019	12.312	1082.27	1315.67	-46.31	-32.671
Butanol	10.064	12.3013	1680.87	1285.0227	-59.25	-99.903
Octyl-benzene	9.402	12.46418	2173.842	2121.3576	-63.591	-71.162

The close agreement between these sets of coefficients underscores the model’s ability to extract meaningful, domain-specific features from high-dimensional molecular graph data. For all three components, the model’s A , B , and C parameters closely align with the empirical Antoine coefficients. The accuracy of these outputs reinforces the model’s capability to integrate the inductive bias of the Antoine equation into the predictive framework, ensuring not only high predictive accuracy but also thermodynamic consistency.

It is worth highlighting that the model was not fed any information about the Antoine coefficients during its training. Despite this, the PUFFIN model demonstrated a remarkable ability to generate embeddings consistent with the Antoine equation. The findings underscore the robustness of the PUFFIN model, given that it has been trained purely on high-dimensional molecular data and temperature–vapor pressure pairs, with no explicit information regarding the Antoine coefficients being provided.

These findings suggest that the hybridized structure of PUFFIN provides a methodology for incorporating domain-specific knowledge into the machine learning architecture, which can be extended to other fields. This mechanism promotes a richer interpretation of the model’s outputs and contributes to its performance in accurately predicting vapor pressure. In particular, the successful retrieval of Antoine-like coefficients from the intermediary layer validates the design choice to use the Antoine equation as an activation

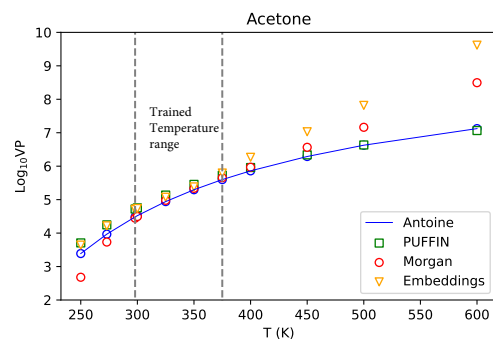
function for the output layer. This lends credence to the concept that incorporating domain knowledge into machine learning models can improve the interpretability and the predictive performance of these models.

A final assessment is provided in Figure 6.5. The figure presented in this analysis illustrates the extrapolation of three strategies in predicting vapor pressure, specifically focusing on the proposed strategy, PUFFIN. Figure 6.5 demonstrates the performance of each strategy relative to Antoine predictions, particularly in regions extending beyond the confines of the training data.

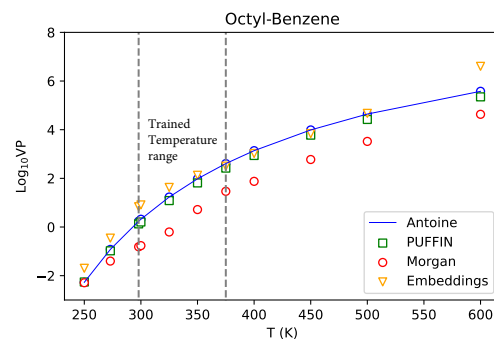
Notably, the PUFFIN strategy aligns accurately with Antoine predictions, even in regions significantly distant from the original training data where the other models drift significantly, a behavior clearly seen in Figure 6.5.a. This observation is important as it highlights the potential of our modified AI model to predict vapor pressure as a continuous function of temperature with remarkable accuracy.

In contrast, the other two strategies start to diverge noticeably from Antoine predictions when venturing into regions distant from the training area. This stark difference further emphasizes the exceptional precision and robustness of the PUFFIN strategy. Furthermore, it is essential to underline that Antoine coefficients are typically derived from experimental data, making obtaining these values for all the molecules incorporated in our database is challenging. Despite this, our figure evaluates three molecules from the unseen test dataset.

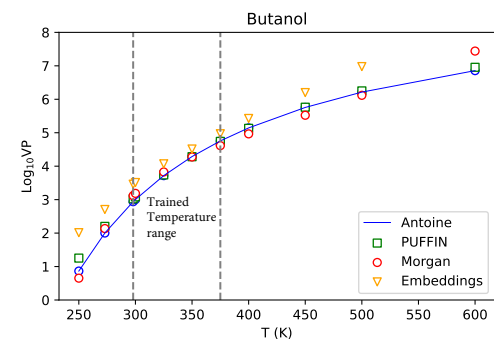
The results from these evaluations show that PUFFIN can accurately predict vapor pressure for molecules unseen during the training phase and for temperatures well beyond the range of the training set. This capability highlights the potential of PUFFIN in predicting vapor pressure for new molecules and for temperatures outside the known training data, making it a compelling area for future research.



(a)



(b)



(c)

Figure 6.5: Extrapolation assessment of all models for three test set components. (a) Acetone, (b) Octyl-Benzene and (c) Butanol.

6.5 Conclusions

In conclusion, this study presents a novel approach for predicting vapor pressure using a hybrid neural network model, the PUFFIN framework, demonstrating good accuracy and partial interpretability. We adopted a transfer learning strategy, leveraging widely available boiling point data. Furthermore, we incorporated domain knowledge (thermodynamic) concepts into the neural network architecture with an inductive bias. Our results indicate that this approach significantly enhances the model’s prediction capabilities, reducing the mean squared error by about 50% compared to the other two neural network approaches (the Morgan fingerprint network and the non-reinforced molecular embedding network).

Moreover, the PUFFIN framework’s ability to extract Antoine-like coefficients directly from high-dimensional molecular graph data offers a compelling illustration of the benefits of incorporating domain knowledge into machine learning models. This promising outcome offers a degree of model interpretability typically lacking in standard neural network approaches, contributing significantly to the model’s utility. Remarkably, these advances were achieved without providing the model with explicit information about the Antoine coefficients during training. This demonstrates the power of machine learning models to learn and utilize domain-specific relationships effectively, even without explicit domain knowledge during training.

Furthermore, the introduction of an Antoine node in the PUFFIN permitted a physically accurate extrapolation of vapor pressure data to temperatures beyond those to which training was completed. This is in contrast to the two other methods (the Morgan fingerprint and the non-reinforced molecular embedding), which were both less accurate in predicting vapor pressure at temperatures outside of the range used for training.

The improvements realized by the PUFFIN framework affirm the considerable potential of incorporating domain-specific knowledge into machine learning architectures. The demonstrated success of this approach suggests potential applicability beyond vapor pressure prediction, extending to other domains that could benefit from a similar integration of domain knowledge into machine learning models. In the face of ever-increasing complexity and data availability in various scientific and engineering domains, our study underscores the importance and potential of physics-informed and interpretable machine learning models for accurately predicting complex properties. This further underscores the potential of the PUFFIN model as a reliable instrument for predicting complex thermophysical properties.

Chapter 7

Advancing Odor Classification Models Enhanced by Scientific Machine Learning and Mechanistic Model: Probabilistic Weight Assignment for Odor Intensity Prediction and Uncertainty Analysis for Robust Fragrance Classification

7.1 Abstract

The section delineated in chapter 5 described a deterministic model for the diffusion of perfume, predicated on the assumption of well-defined simulation parameters. Despite the advances presented, critical challenges in the application of these models persist, chiefly the deficiency of rigorous uncertainty quantification. Such uncertainties arise from various sources including the fundamental mechanistic models that govern perfume diffusion—characterized by parameters like vapor pressure, diffusion coefficients, and activity coefficients—and the complex taxonomy involved in the classification of individual odorants across olfactory families. Advancements in the field involve identifying, quantifying, and systematically incorporating these uncertainties into the predictive modeling tools utilized in fragrance mixture design. Furthermore, there is an observable lacuna in the literature and the preceding chapter with respect to a fully developed probabilistic approach for scent classification that integrates the classification of singular odorants to the uncertainty present in mechanistic models. This chapter endeavors to construct a holistic methodology

that not only rectifies these noted limitations but also strives to establish a more robust and empirically reliable paradigm for fragrance design.

7.2 Introduction

Fragrances are deeply embedded in consumer products across the globe, serving as a pivotal element in captivating consumers and fueling purchases. Their core objective is to bestow products with a pleasant and harmonious aroma, evoking positive sensations upon interaction. Both a signature of distinction and a source of pleasure, fragrances can elevate a product’s desirability and functional appeal to its audience Calkin and Jellinek (1994); Cussler et al. (2010); Teixeira et al. (2011); Butler (2013).

However, crafting a harmonious and pleasant fragrance is complex considering the number of available fragrances (on the order of several thousand) (Teixeira et al., 2011). The possible combinations of fragrance ingredients for a perfumer are nearly endless – achieving a desired olfactory target involves a long and costly trial-and-error process for testing the perceived odor of hundreds of different samples (Santana et al., 2021). Historically, this intricate process has leaned heavily on the expertise of seasoned perfumers who meticulously adjust ingredient proportions to attain the desired olfactory experience in intensity and quality (Mata et al., 2005; Rodrigues et al., 2021). In the past two decades, the use of Product Engineering in fragrance design has emerged as a valuable tool to integrate scientific knowledge into a traditionally empirical and subjective domain (Mata et al., 2005; Rodrigues et al., 2021; Teixeira et al., 2013a). A notable benefit of this integration is the ability to decrease the number of chemicals in formulations, all while achieving the desired perceptual outcomes.

From the viewpoint of Product Engineering, understanding the perception of a fragrance in a perfume involves a sequence of events, some of which are grounded in well-defined physics and mathematical frameworks: (i) The process commences with a solution of fragrance ingredients and solvents, characterized by specific molar compositions (x_i), which users apply to their skin or clothing. (ii) Upon application, this liquid blend starts evaporating into the ambient atmosphere, resulting in alterations in both nearby gas-phase (y_i) and liquid-phase (x_i) compositions. (iii) As the evaporation proceeds, the fragrant components spread into the surrounding air. (iv) Ultimately, some fragrance molecules reach the olfactory senses of nearby individuals, conveying distinct intensities and scent characteristics (Teixeira et al., 2011).

In this sense, it is of great interest to establish reliable predictive models that can accurately describe this chain of processes from the release of a scent on a substrate to its reception by the human olfactory system. Once developed and validated, such models offer extensive utility – they can deepen our comprehension of fragrance perception or aid perfumers in the design and formulation processes, for instance.

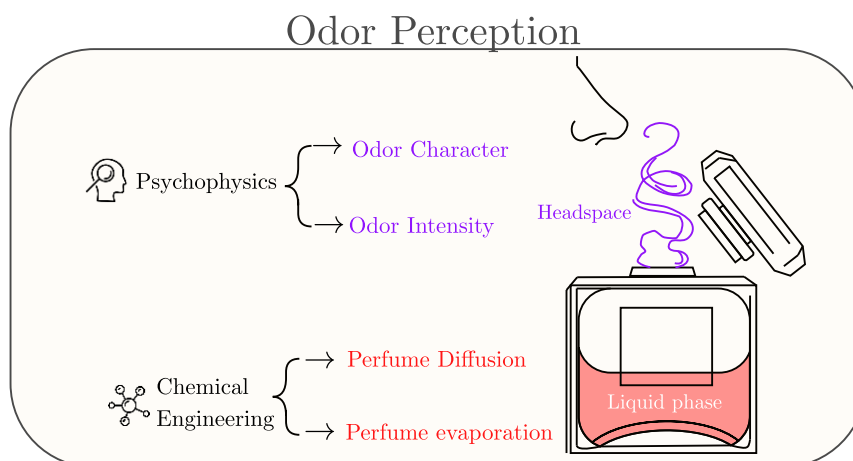


Figure 7.1: Schematic representation of the steps involved in the odor perception model for fragrance mixtures. Adapted from Teixeira (2011).

Over the past decades, advancements in Perfume Engineering have made remarkable strides in integrating well-established physics with the description of scent perception. This integration leans heavily on Chemical Engineering domains, encompassing solution's thermodynamics, transport phenomena, and the psychophysics of human sensory perception. Figure 7.1 visually represents this intricate chain of processes and highlights the specific scientific disciplines engaged at each juncture. Within this framework, several developments have emerged, such as the Perfumery Ternary Diagrams (PTD) methodology (Mata et al., 2005), the perfume radar (Teixeira et al., 2014), and research on the diffusion and performance of perfume mixtures (Teixeira et al., 2009a, 2013b).

From the perspective of the mentioned works, perfumes can be defined as liquid mixtures of odorants dissolved in appropriate solvents, such as a combination of ethanol and water. Chemical Engineering offers models for understanding the transition from liquid to vapor through thermodynamics and the subsequent diffusion of these scents in the air. Furthermore, the concentration of these fragrant molecules can be related to the sensory experience through models of odor intensity and perception, drawing insights from psychophysics. This provides an appropriate framework for describing and evaluating perfume behavior as demonstrated in Teixeira et al. (2009a, 2013b); Zhang et al. (2020b).

Although this approach has shown strong predictive capabilities, uncertainties in the model's parameters can significantly influence their predictions and subsequent tasks where the model is applied. The evaporation of ingredients used in the formulation is a pivotal concern for product design. A component's vapor pressure is essential in evaporation rate prediction within this framework. Yet, vapor pressure is often based on limited experiments, introducing a notable degree of uncertainty, particularly for rarer perfume ingredients. Additionally, for more complex mixtures — whether they are non-ideal, azeotropic, or close-boiling — static uncertainties are inherently present in comprehensive thermody-

namic models like UNIFAC, UNIQUAC, or NRTL, which are utilized to obtain deviations from ideality through activity coefficients (Ulas and Diwekar, 2004). Therefore, it is important to identify, quantify, and incorporate these uncertainties into the models.

When considering diffusion in model odor perception, the diffusion coefficient can significantly influence predictions about the concentration of fragrance ingredients in the surrounding air. While finding vapor pressure for rare ingredients is challenging, determining the diffusion coefficient in air is even more difficult. Researchers typically rely on predictive correlations derived from noisy experimental data to estimate it. However, as in any estimation process from noisy observations, an irreducible aleatoric uncertainty is associated with the obtained predictive correlation.

From the perspective of perception, that is, how an individual consciously experiences and describes a smell in terms of intensity and its olfactory family for a given concentration of an odorant, a lot of variation is expected among individuals, i.e., different people experience the same smell differently. This variability (uncertainty) poses a significant challenge when using psychophysics to correlate concentrations with scent intensity and associating individual component intensities and olfactory families with the overall perception of a fragrant mixture. In this way, in olfactory perception studies, factors like inherent variability and response bias are inevitable and should be considered (Teixeira et al., 2013a; Chambers and Wolf, 1996).

Traditionally, the categorization of individual odors or molecules has relied heavily on the expert sensory perception of human evaluators. However, a paradigm shift is occurring, as evidenced by recent works like those of Sanchez-Lengeling et al. (2019b) and Keller and Vosshall (2016); Keller et al. (2017). These researchers have adopted machine learning techniques to approach odor classification from a Quantitative-Structure-Odor-Relationship standpoint. They aim to establish correlations between chemical descriptors of individual odorants and their associated olfactory families or nuances.

When it comes to fragrances, which typically comprise 50 to 100 different aromatic components (odorants), the classification challenge grows even more complex. Conventional methods often categorize perfumes into specific classes or families, sometimes further delineating subfamilies or nuances, largely based on empirical assessments. To inject a layer of scientific rigor into fragrance classification, a novel methodology called Perfumery Radar (PR) has been developed, as cited in the work by Teixeira et al. (2014). This approach employs physicochemical models and qualitative descriptors to categorize commercial perfumes systematically.

In Teixeira et al. (2014) methodology, a significant limitation exists in using traditional techniques for weight assignment. This approach risks introducing bias, as it does not consider the probabilistic nature of olfactory classifications. The consequence is a potential failure to capture the full spectrum of odor intensity variations across different molecular components. This rigidity in weight allocation can limit the model's adaptability to new or complex olfactory scenarios, thereby affecting the overall accuracy and reliability of the

odor classification model. A further limitation of Teixeira et al. (2014) methodology is the absence of a mechanism for handling uncertainty. These uncertainties can manifest from the mechanistic models describing perfume diffusion and the variable classifications within olfactory families for individual odorants. Additionally, it is essential to identify, quantify (when feasible), and integrate these uncertainties into the predictive modeling tools used for designing fragrance mixtures.

Building upon Teixeira et al. (2014), our methodology introduces two key advancements: weight assignment using scent classifier output probabilities and an integrated uncertainty analysis framework. The former enhances the model's adaptability in diverse olfactory landscapes, while the latter fortifies its robustness and generalizability, especially in real-world settings with data imperfections. Uncertainties are quantified via probability statements and density functions, which are derived from prediction residuals for least squares estimated parameters.

These quantified uncertainties are then propagated through the perfume diffusion model using sampling methods, allowing us to assign probabilities to dependent variables, such as the likelihood of evoking a specific olfactory response. This holistic approach not only addresses the limitations of the original work but also paves the way for a more robust and reliable fragrance design.

7.3 Methodology

This section is divided into three parts. The first and second parts detail the perfume diffusion and odor perception models. The third part identifies the source of uncertainties and proposes ways to quantify them.

7.3.1 Perfume diffusion model

This study employs as a base the perfume diffusion model previously proposed and validated by Teixeira et al. (2009a, 2013b). The proposed model focuses on the amount of liquid perfume typically released with a single spray. The model aims to simulate the evaporation process observed when a perfume bottle's diffuser is pressed. To validate the methodology proposed here, a mixture of limonene, geraniol, vanillin, and ethanol, also investigated in Teixeira et al. (2010a), was used as a case study. The initial composition of the components is 0.12, 0.12, 0.06, 0.7, respectively.

The model aims to describe the evaporation of a small amount of liquid perfume, which rises through the air above it (headspace) in the axial direction z dominantly by diffusion where interactions between the scent in the gas phase and surrounding air molecules are negligible, implying an ideal gas phase. It is assumed that the ambient air does not mix with the liquid and that mass transfer resistance in the liquid can be neglected due to the small quantity of liquid. Figure 7.2 illustrates the modeled process.

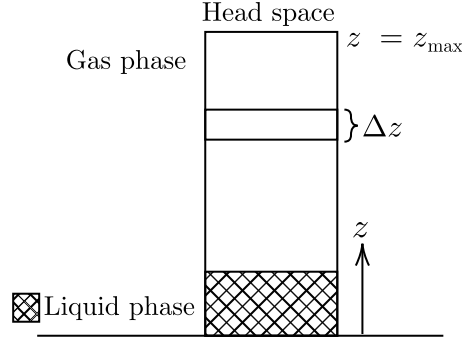


Figure 7.2: Scheme of the simulated liquid and gas phases of the perfume system. Adapted from Teixeira et al. (2009a)

Over time, as evaporation proceeds, the volume and composition of the fragranced mixture change. The liquid phase number of moles is denoted by n_i , while the concentration of each component in the air is represented by c_i . Additionally, it is assumed that the evaporation happens at constant pressure (P) and temperature (T).

Considering the above conditions, the time-dependent mass balance in the gas phase can be expressed by the following partial differential equation (PDE):

$$\frac{\partial y_i}{\partial t} = D_{i,air} \frac{\partial^2 y_i}{\partial z^2} \quad (7.1)$$

Here, y_i represents the mole fraction of component i in the gas phase, linked to the gas-phase concentration as $C_{i,g} = y_i C_T$. In this relation, C_T is the total concentration in the gas phase, defined by both pressure and temperature. The variable t stands for time, z the direction perpendicular to the liquid-gas interface, while $D_{i,air}$ indicates the diffusivity of component i in the air. $D_{i,air}$ was estimated from Fuller et al. (1966) correlation and given as:

$$D_{i,air} = \frac{1 \times 10^{-8} T^{1.75} \left(\frac{1}{M_i} + \frac{1}{M_{air}} \right)^{1/2}}{P[(\sum_A v_i^{1/3} + 20.1)^{1/3}]} \quad (7.2)$$

v_i is the Diffusion Volume based on the group contribution method and tabulated in the work, T is the temperature in Kelvin, P is the pressure in atm, M_i is the molecular weight in g/mol.

The PDE requires two boundary conditions at the extrema of the z domain. They can be written as:

$$t > 0, z = 0 : y_i(t, z) = \frac{\gamma_i(x_i(t), T) x_i(t) P_i^{\text{Sat}}(T)}{P}; x_i(t) = \frac{n_i(t)}{\sum_{i=1}^N n_i(t)} \quad (7.3)$$

$$t > 0, z = z_{\text{max}} : y_i(t, z) = 0 \quad (7.4)$$

As seen from Equation 7.3, the vapor-liquid equilibrium is predicted using Raoult's law, and the liquid-phase non-ideality accounted by the activity coefficients (γ_i) calculated using the UNIFAC Dortmund group-contribution method (Poling et al., 2001b; Gmehling, 2003; Gmehling et al., 2002). y_i and x_i are the mole fractions of the i^{th} component in the gas and liquid phase, respectively. N is the number of components, $P_i^{\text{Sat}}(T)$ is the saturation pressure at temperature T .

The liquid-phase mass balance can be written considering the mass transfer of the components across the interface. This is represented by the subsequent ordinary differential equation (ODE):

$$\frac{dn_i}{dt} = D_{i,\text{air}} A_{l,g} C_T \frac{\partial y_i(z=0, t)}{\partial z} \quad (7.5)$$

in which n_i is the number of moles of component i in the liquid phase, $A_{l,g}$ is the area of the liquid-gas interface.

7.3.1.1 Simulation - Numerical solution and software implementation

The perfume diffusion simulator was implemented in the open-source Julia Programming Language (Bezanson et al., 2012). The code is publicly available in the GitHub repository described in the Code Availability section.

To numerically solve the PDE, we employed the method of lines. The axial coordinate z was discretized into an evenly-spaced grid with 40 nodes. The second-order derivative in Equation 7.1 was approximated with a centered finite difference scheme, i.e.:

$$\frac{\partial^2 y_i}{\partial z^2} \sim \frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} \quad (7.6)$$

where n is the grid node index, i.e., $n \in \{1, 2, \dots, 40\}$, and h is the distance between two nodes in the grid.

The derivative at $z = 0$ in Equation 7.5 was approximated using a 3-point forward difference for improved accuracy of flux calculation at the boundary:

$$\frac{\partial y_i(z=0, t)}{\partial z} \sim \frac{-y_3 + 4y_2 - 3y_1}{2h} \quad (7.7)$$

The assignment of UNIFAC Dortmund groups for each component in the chosen mixture (limonene, geraniol, vanillin, ethanol) was automated using the GCIdentifier.jl library using the SMILES representations as inputs. Subsequently, activity coefficients were computed using the Clapeyron.jl (Walker et al., 2022) library in the Julia programming environment.

The resulting ordinary differential equation (ODE) was integrated using a fixed-leading coefficient adaptive-order adaptive-time BDF method (FBDF) for stiff ODEs available in DifferentialEquations.jl (Rackauckas and Nie, 2017).

In Fuller et al. (1966) correlation, the groups required for calculating diffusion volumes were also automated using RDKit python API.

7.3.2 Odor intensity perception model

Of the various models, the Odor Value (OV) is a simple yet intuitive, dimensionless variable that estimates the scent intensity of fragrant molecules. The OV is determined by dividing the concentration of an odorant species i in the gas phase, denoted as $C_{i,g}$, by its odor threshold concentration, Thr_i .

In odor science literature, four principal types of odor thresholds are commonly discussed:

- Detection Threshold (Absolute Odor Threshold): The minimum concentration at which an odor becomes detectable.
- Recognition Threshold: The concentration at which the odor is detectable and identifiable.
- Terminal Threshold: The point at which the odor becomes pungent or irritating.
- Difference Threshold: A comparative or relative odor evaluation, quantifying the minimum change in stimulus—such as gas concentration—required to produce a noticeable difference in perception.

Experimentally, the Difference Threshold is often determined by comparing a standard stimulus to a variable one, measured on a metric scale (Teixeira et al., 2013a). Despite these numbers' differences, detection thresholds are often reported and used in perfume engineering. In this way, it was also employed in this work.

The OV considering the detection threshold of a component i can be written as:

$$OV_i = \frac{C_{i,g}}{Thr_i} \quad (7.8)$$

When several components are present in the air, the perception of that mixture will be affected by all components. The perception of a mixture of odorants is a complex process not understood thoroughly (Teixeira et al., 2011; Gilbert, 2008). Several models have been proposed to explain the perception of odorants in mixtures, although no consensus was reached about the best one (Laffort and Dravnieks, 1982; Cain et al., 1995; Olsson, 1998). The stronger component model is the simplest and most often used in Product Engineering of fragrant mixtures. It states that the one with the highest odor intensity is the most strongly perceived in a mixture of odorants. It means that the odor with the highest intensity determines the odor intensity of the mixture:

$$OV_{mix} = \max\{OV_1, OV_2, \dots, OV_N\} \quad (7.9)$$

The odor intensity model can be linked with the perfume diffusion model by substituting its numerical solution in the odor value expression. In this way, if $C_{i,g}$ is a variable distributed in (t, z) , one can know the corresponding odor value in time and space.

7.3.3 Odor classification

Building on the work in Teixeira et al. (2014), we also classify scents by combining the olfactory families and predicted intensities of individual odorants. However, we improve upon this by using a probabilistic approach to construct the radar chart initially proposed in Teixeira et al. (2014). Unlike the original study, which used arbitrary weights for olfactory families, our method uses a classifier to calculate the probability for each family of odorants.

The classifier comes from the methodology, the Principal Odor Map (POM), outlined in the study by Lee et al. (2023). Barsainyan (2023) reproduced the methodology and made the model implementation publicly available. A comprehensive dataset was curated by extracting the CAS number of odorants and their corresponding smell semantic descriptors. This curated data set has been organized into a well-structured data frame with records and made publicly accessible via a GitHub repository.

The POM approach employed the message-passing neural network classifier. The molecules were encoded into adjacency and feature matrices. The feature matrix includes binary features regarding each atom’s atomic type, bonding properties (orbitals hybridization), and the presence of a chirality center. The target variable was a 138-dimensional binary vector, indicating an odorant’s membership in various olfactory families.

7.3.4 Uncertainty quantification and propagation

The predictive perfume diffusion and odor intensity models contain parameters/variables $(T, P, D_{i,air}(T, P), P^{Sat}(T), \gamma(x, T), Thr_i, A_{l,g})$ that must be known to be used for predictions. These quantities contain uncertainty as they are estimated from direct or indirect noisy observations. This section describes the methodology used to quantify uncertainties using probability theory and the method to propagate them through the simulator.

While this study addresses the uncertainties in $D_{i,air}(T, P)$, $P^{Sat}(T)$, and Thr_i , it should be noted that temperature and pressure (T, P) can also fluctuate during experimentation, affecting the results. However, these factors are not considered the most critical variables for the scope of this work. Another important parameter is the area at the gas-liquid interface $A_{l,g}$, commonly set at 0.071 m^2 (Teixeira et al., 2013a). However, no rational basis exists for this specific value in the literature. Consequently, this study does not account for the uncertainty associated with this particular parameter. Finally, the uncertainty in the activity coefficient $\gamma(x, T)$ significantly influences predictions, as quantified in distillation processes in Ulas and Diwekar (2004). Unfortunately, an accurate assessment of the uncertainty of this variable requires multiple measurements of liquid-vapor

equilibrium, which are unavailable for the case study mixture explored here, consisting of limonene, geraniol, vanillin, and ethanol.

7.3.4.1 Quantifying uncertainties from Least Squares Estimation (LSE) of parametrized models: Vapor Pressure and diffusion coefficient

In the literature, the Antoine equation is frequently used to derive vapor pressure, expressed as $\log P_{sat} = A - \frac{B}{T_{sat} + C}$. This equation correlates the saturation temperature T_{sat} with the saturation pressure P_{sat} and is considered valid within a specific temperature and pressure range. To estimate the parameters A , B , and C , a finite dataset $\mathcal{D}_{P,T} = \{(P_{sat,i}, T_{sat,i})\}_{i=1}^N$ containing pairs of pressure and temperature values is collected, where N is the number of data points. These pairs are often noisy, and are utilized in an estimation problem where typically the target variable is either the temperature T_{sat} or the logarithm of the pressure $\log P_{sat}$, which transforms the corresponding variable into a predictor.

For diffusion coefficients, the correlation proposed by Fuller et al. (1966) has been utilized in perfume engineering, specifically within the perfume diffusion model as cited in studies by Teixeira et al. (2013b, 2009a). The authors assert that this model offers satisfactory accuracy when applied to volatile compounds diffusing through air. The correlation has grounds in the hard-sphere model from the Chapman-Enskog equation (Chapman and Cowling, 1990). However, Fuller et al. (1966) proposed an improved form of the equation that contains parameters to be identified from experimental data, i.e., from a data set $\mathcal{D}_{D_{i,j},P,T} = \{(D_{1,j}, P_1, T_1), (D_{2,j}, P_2, T_2), \dots, (D_{N,j}, P_N, T_N)\}$, where $D_{i,j}$ is given by:

$$D_{i,air} = \frac{CT^b \left(\frac{1}{M_i} + \frac{1}{M_j} \right)^{1/2}}{P[(\sum_A v_i^{\alpha_1} + \sum_A v_j^{\alpha_2})]} \quad (7.10)$$

v_i is the diffusion volume based on the group contribution method, C is an arbitrary constant, b is the exponent dependency of temperature, and α_1, α_2 are the diffusion volumes' exponents. All parameters were identified by solving an estimation problem. In such work, C, b, α_1, α_2 were found to be $1 \times 10^{-8}, 1.75, 1/3, 1/3$, respectively for T in K, P in atm, M_i in g/mol, $D_{i,j}$ in cm^2/s . The Diffusion Volumes can be found in Fuller et al. (1966) work.

There are two predominant methods to formulate and solve the estimation problem: the frequentist and Bayesian approaches. The frequentist method, owing to its simplicity, is more widely used. This approach can be derived using the maximum likelihood principle, resulting in the Least Squares Estimation (LSE).

To associate the LSE with a probability density function post-estimation (point predictive distribution), it is essential to derive it from the maximum likelihood principle. The initial step involves assigning a probability density function (pdf) or likelihood function to the data generation process. In LSE, the Normal distribution ($\mathcal{N}$) with a mean equal to the predictive model ($\phi(x, W)$) with predictors x , parameters W and unknown variance

σ^2 is used:

$$y_i \sim \mathcal{N}(y_i | \phi(x_i, W), \sigma^2) \quad (7.11)$$

In this work, $\phi(x, W)$ can be either the Antoine equation if the prediction is for vapor pressure or Fuller et al. (1966) correlation (Equation 7.10) if the prediction is for diffusion coefficient in air.

From the pdf of a single observation and assuming that all observations are independent and identically distributed (iid), we can write:

$$y_{1:N} \sim \prod_{i=1}^N \mathcal{N}(y_i | \phi(x_i, W), \sigma^2) \quad (7.12)$$

$$p(y_{1:N} | \phi(x_{1:N}, W), \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^N (y_i - \phi(x_i, W))^2 \right\} \quad (7.13)$$

To determine all the parameters, we employ the maximum likelihood function. This encompasses parameters from the Antoine or diffusivity coefficient equations and the measurement variance, denoted as σ^2 . Given that the values of the likelihood function can be close to zero, it is often more practical to minimize the negative logarithm of the likelihood. By doing so, we derive:

$$\begin{aligned} \min_{W, \sigma^2} \mathcal{J} &= -\log \{p(y_{1:N} | \phi(x_{1:N}, W), \sigma^2)\} = \\ &\left\{ \frac{N}{2} \log(2\pi) + \frac{N}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \left\| (y_{1:N} - \phi(x_{1:N}, W)) \right\|^2 \right\} \end{aligned} \quad (7.14)$$

Using deterministic optimization methods, taking the derivative of $\mathcal{J}$ with respect to W and setting it to zero, W optimal can be estimated. Once W has been estimated, σ^2 can be estimated by setting $\frac{d\mathcal{J}}{d\sigma^2} = 0$. This yields an analytical representation for σ^2 as:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \phi(x_i, W))^2 \quad (7.15)$$

Though this formula is commonly termed the Mean Squared Error (MSE), from a statistical lens, it corresponds to the estimation of the variance of the likelihood function.

Finally, we can incorporate this measurement uncertainty when making predictions. This is done through the point predictive distribution, which is Normal:

$$p(y | \phi(x, W), \sigma^2) = \mathcal{N}(\phi(x, W), \sigma^2) \quad (7.16)$$

After estimation, however, verifying if the residuals follow the Normal distribution, as assumed in likelihood, and if the variance is constant is important. The normality of residuals can be verified by the goodness-of-fit Kolmogorov–Smirnov test (Massey, 1951).

The homogeneity of the variance can be verified by visual inspection with parity plots or using adequate testing statistics, e.g., Levene’s test.

7.3.4.2 Quantifying uncertainties of odor threshold

Various techniques and equipment may be employed for the calculation of odor thresholds. An average odor threshold value can be ascertained by examining the distribution of relative sensitivity to a specific chemical odorant across a large population. It is important to recognize that although this final result may appear as a single threshold value or an estimated range of values with accompanying confidence limits, it is based on a substantial amount of underlying experimental tests. Inherent variability and response bias are unavoidable factors in olfactory perception studies. Even within the same panel, individual responses can vary significantly—ranging from a two-fold to a four-fold difference in the same experiment (Chambers and Wolf, 1996). Interpersonal variability can be even more pronounced, with differences sometimes escalating to as much as three orders of magnitude (Teixeira et al., 2013a).

Individual and interpersonal variability must be considered to quantify the uncertainty linked with odor detection thresholds comprehensively. A major obstacle in achieving this comprehensive understanding is the absence of standardized methods for reporting such metrics. While some studies report confidence intervals, others only offer a single threshold value, as is often seen in compilations of odor thresholds (van Gemert, 2011). We suggest employing a weighted sum of probability density functions to address these varying conditions.

For cases where an interval is reported, a uniform probability density function is assigned to that range. When only a single value is reported, a Dirac delta function is the probability density function for that specific number. If a combination of these two scenarios exists for a given component, we construct a weighted sum of probability density functions, assigning equal weights to each. This is based on the assumption that opting for a conservative distribution ensures robustness in the model. This approach is particularly relevant for input variables whose underlying distribution is unknown or insufficiently characterized. This assumption maximizes our findings’ reliability in cases where the data is limited or ambiguous.

In this study, diffusion, vapor pressure, and threshold concentration uncertainties are explicitly characterized using conservative statistical methods. For other significant variables — namely surface area and activity coefficient—either data are insufficient, or no established methodology exists for uncertainty quantification. While the present work does not encompass all sources of uncertainty, it is an initial effort to integrate uncertainty related to threshold values within the perfume engineering framework, thereby providing a basis for future advancements in robustness and accuracy.

7.3.4.3 Propagating uncertainty through the perfume diffusion simulator

In uncertainty quantification using probability theory, sampling methods are often utilized to propagate uncertainty through a simulator, yielding samples of dependent variables. In this study, our main focus is to obtain samples in the odor intensity in time and space for each component in the mixture, namely limonene, geraniol, vanillin, and ethanol. To accomplish this, we sample from the probability density functions of all uncertain parameters and run simulations of the perfume diffusion model for each set of samples. While it is possible to calculate statistical metrics such as expectation and variance related to odor intensity without propagating uncertainty, as suggested by the Koopman operator discussion in `SciMLExpectations.jl`, the present work opted to estimate these metrics using output samples.

Various sampling methods are available for this type of analysis, ranging from the straightforward “crude” Monte Carlo method, which utilizes a pseudo-random generator to approximate a uniform distribution, to more advanced techniques specifically designed for multivariate distribution sampling. Such specialized methods include Latin Hypercube Sampling (Hardin and Sloane, 1993) and Sobol sequences (Sobol’, 1967). It is worth noting that, theoretically, quasi-Monte Carlo methods such as Sobol sequences are low-discrepancy and therefore converge as integral approximations at a rate of $O(n)$, in contrast to the $O(\sqrt{n})$ convergence rate of a naive Monte Carlo method. This theoretical framework highlights the potential differences in convergence speed. In the present work, the convergence rates for calculating expectation, variance, and percentiles are compared between the crude Monte Carlo and Sobol sampling methods.

The built pdf samples were obtained using `Distributions.jl` (Lin et al., 2019; Besançon et al., 2021) along with `MonteCarloMeasurements.jl` (Bagge Carlson, 2020) for crude Monte Carlo and `QuasiMonteCarlo.jl` for Sobol sampling.

7.3.5 Integration of models for robust fragrance odor classification

While we have based our approach on the work by Teixeira et al. (2014) for integrating the perfume diffusion model with odor classification, our method introduces several enhancements. As a first contribution, we integrated a unique methodology to cope with uncertain outcomes within the odor classification model, something not present in the original work. For a more visual understanding of this work’s contribution, the methodology is schematically and visually depicted in Figure 7.3.

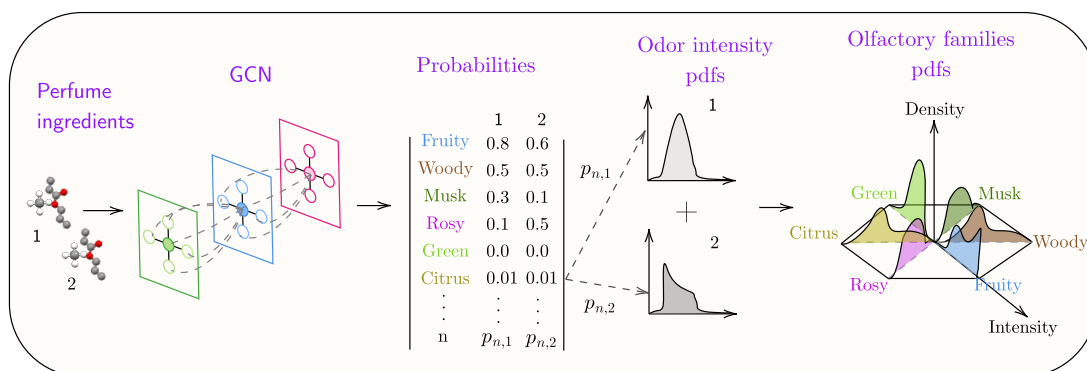


Figure 7.3: Diagram for the methodology of robust fragrance classification (olfactory families are generic).

7.4 Results

Furthermore, a significant feature of our method lies in its utilization of the classifier’s output probabilities. Instead of traditionally assigning weights, we use these probabilities as an unbiased technique to allocate weights to each molecule’s odor intensity probability distribution function.

In the process, the probabilities of different olfactory families, determined by the GCN, serve as these weights. When multiplied with the odor intensity trajectories of each component, each radar segment corresponding to an olfactory family becomes a weighted combination of these trajectories.

7.4.1 Quantifying uncertainties of diffusion coefficient and vapor pressure

Using the methodology outlined in Section 7.3.4.1, a point predictive distribution of the diffusion coefficient based on data from Fuller et al. (1966) was identified. This data comprises the diffusion of 40 compounds in the air and provides predictions and their absolute deviations. A histogram of the error distribution was plotted to evaluate its similarity to a potential Normal likelihood function (Figure 7.4). To confirm the hypothesis of a Normal likelihood, the Massey (1951) goodness-of-fit Kolmogorov-Smirnov (K-S) test was applied, where the null hypothesis postulates that the sample does not result from a Normal distribution with zero mean and standard deviation equal to 8.4297×10^{-7} calculated using Equation 7.15.

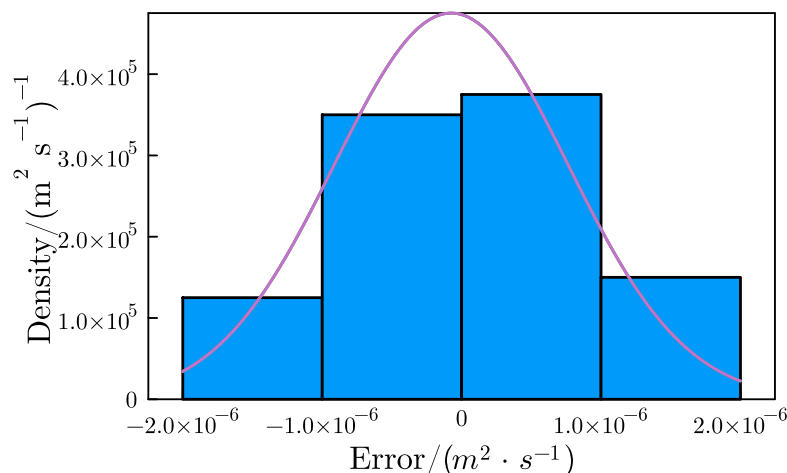


Figure 7.4: Histogram of the error of diffusion coefficients in air.

The K-S test provided a p-value of 0.6977, suggesting insufficient evidence to reject the null hypothesis.

This work assumes that for all the molecules in the fragrant mixture (limonene, geraniol, vanillin, ethanol), the point-predictive distribution of the air diffusion coefficient is given

by a Normal distribution with mean equal to correlation 7.10 and standard deviation 8.4297988×10^{-7} .

It was assumed that the residuals for the saturation temperature follow a normal distribution. To test this hypothesis, vapor pressure data for the solvent, ethanol, was sourced from the Dortmund Data Bank (GmbH, 2023). This data was then compared with the Antoine Equation, represented as $T_{\text{sat}} = \frac{B}{A - \log_{10} P + C}$, with $A = 7.239$, $B = 1582.195$, and $C = -48.855$. The data set encompasses 68 liquid-vapor equilibrium records for ethanol, ranging in temperature from 290.25 K to 351.70 K. A histogram of the saturation temperature error distribution was plotted to evaluate its similarity to a potential Normal likelihood function (Figure 7.5).

The K-S test was employed to validate the assumption of a normal distribution. The null hypothesis of this test posits that the sample is drawn from a normal distribution with a mean of zero and a standard deviation of approximately 0.1468. The K-S test provided a p-value of 0.1285, suggesting insufficient evidence to reject the null hypothesis. This result indicates that the residuals of the saturation temperature of ethanol appear to follow a normal distribution. Notably, the saturation pressure is the primary variable of interest in the perfume diffusion model. When isolating P in the Antoine formula, the resulting probability density function (pdf) for saturation pressure appears to follow a lognormal distribution. This is consistent with the lower boundary of this variable, which is greater than zero.

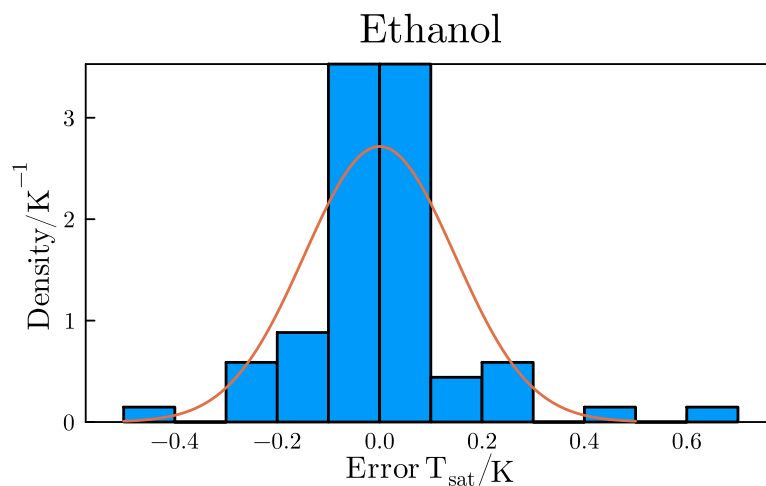


Figure 7.5: Histogram of the error of ethanol's saturation temperature

An alternative approach was implemented to address the lack of readily available experimental vapor-liquid equilibrium data for the remaining components of the fragrant mixture (limonene, geraniol, and vanillin). Given the challenges in sourcing tabulated data, we employed a different methodology. Instead of evaluating the saturation temperature residual distribution, we used an estimate of the saturation temperature standard deviation. This

estimate was obtained from the reported root mean squared errors (RMSEs) in Stephenson and Malanowski (1987).

It is worth noting that the handbook referenced also contains standard deviation values for ethanol, which exhibited values closely aligned with those we obtained through our estimation process (reported 0.02 - 0.1, this work 0.1468). This consistency underscores the reliability of our approach, affirming that it yields a robust approximation for this variable. The Antoine parameters chosen for limonene, geraniol, and vanillin were tailored for predictions at the simulated temperature of 298.15 K. Table 7.1 provides the Antoine parameters and estimated standard deviation (σ) of saturation temperature for each component.

Table 7.1: Antoine parameters and standard deviations for selected substances.

Substance	A	B	C	σ (K)
Limonene	6.81591	2075.62	-16.65	1.0
Geraniol	8.64144	3287.427	0.0	5.0
Vanillin	10.9356	4535.023	0.0	5.0

7.4.2 Quantifying uncertainties of odor threshold

Following the approach detailed in Section 7.3.4.2, we model the odor threshold uncertainty using a mix of uniform and delta distributions. The van Gemert (2011) reported odor thresholds exhibit significant variability in the mixture examined. Specifically, limonene has values of 0.21 and 1.7 mg/m³. This near eightfold increase in threshold values indicates a considerable range in the odor sensitivity for this compound. Geraniol shows an even greater contrast, with values listed as 7.1×10^{-4} and 0.6 mg/m³, suggesting an approximate 850 times difference between the lower and upper threshold. Vanillin's range, from 7×10^{-5} to 0.005 mg/m³ (modeled as Uniform distribution), highlights a 70 times difference, indicating potentially varying perceived intensities in different contexts. Notably, ethanol presents the most diverse range. Its 11 values span from 0.17 mg/m³ to 1350.0 mg/m³. This vast range, nearly 8000 times from the lowest to the highest value, underscores the potential variability in its odor perception, possibly due to multiple sources of data and different experimental conditions.

The observed variability in odor thresholds holds significant implications for the model-based design of fragranced mixtures. Different selections of thresholds can lead to vastly different mixture designs. When this variability is coupled with the uncertainties inherent in the perfume diffusion model, several orders of magnitude can diverge the resulting predictions.

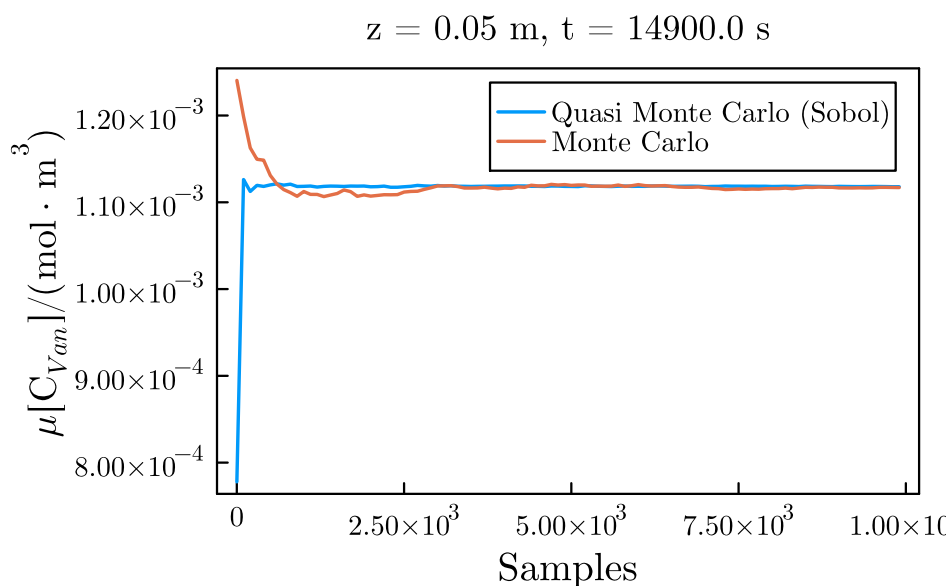


Figure 7.6: Evolution of the expected value of geraniol concentration close to the perfume source and 14900s after the release

7.4.3 Uncertainty propagation

To propagate the uncertainty of the parameters through the perfume diffusion model, 60,000 samples were drawn from each presented probability density function using both "crude" Monte Carlo and Sobol sequences as described in the methodology, except for those that are Dirac mixtures—specifically, the odor thresholds for limonene, geraniol, and ethanol. To estimate statistics of the odor values of these components, samples from the concentration are apportioned among each discrete possible value that can be assumed. For example, given that the limonene threshold can be either 0.21 or 1.7, the 60,000 trajectories are effectively doubled to 120,000 by dividing each trajectory by each threshold value.

The quantity of samples was predetermined, taking into account the runtime of the perfume diffusion model. Nevertheless, to visually assess the appropriate cutoff value for the number of samples, a convergence verification of summary statistics (mean) was executed for both sampling methods. Upon reaching this number, it is reasonable to infer that the sampling statistics exhibit negligible variation.

To observe the convergence of the mean gas-phase concentration, samples taken near the releasing source and after an extended period were employed as reference points for geraniol, given that variations under these conditions are more pronounced, as depicted in Figure B.1. Figure 7.6 illustrates that Sobol Sampling demonstrates earlier convergence than Crude Monte Carlo, in line with expectations.

To first understand how the vapor pressure and diffusivity uncertainties affect the predictions, the mean gas phase concentration and 99% credible interval (spanning the 5th to 99.5th percentiles) at a 100cm distance from the liquid-gas interface is displayed in Figure 7.7 for crude Monte Carlo samples. It can be seen that predictions can differ

by orders of magnitude with amplifying intervals as time evolves. The least pronounced amplitude of the credible interval is related to ethanol, as expected due to its lower variance in the vapor pressure point predictive distribution. Similar figures were plotted to illustrate the same conclusions at different positions from the liquid source and are displayed in the B.

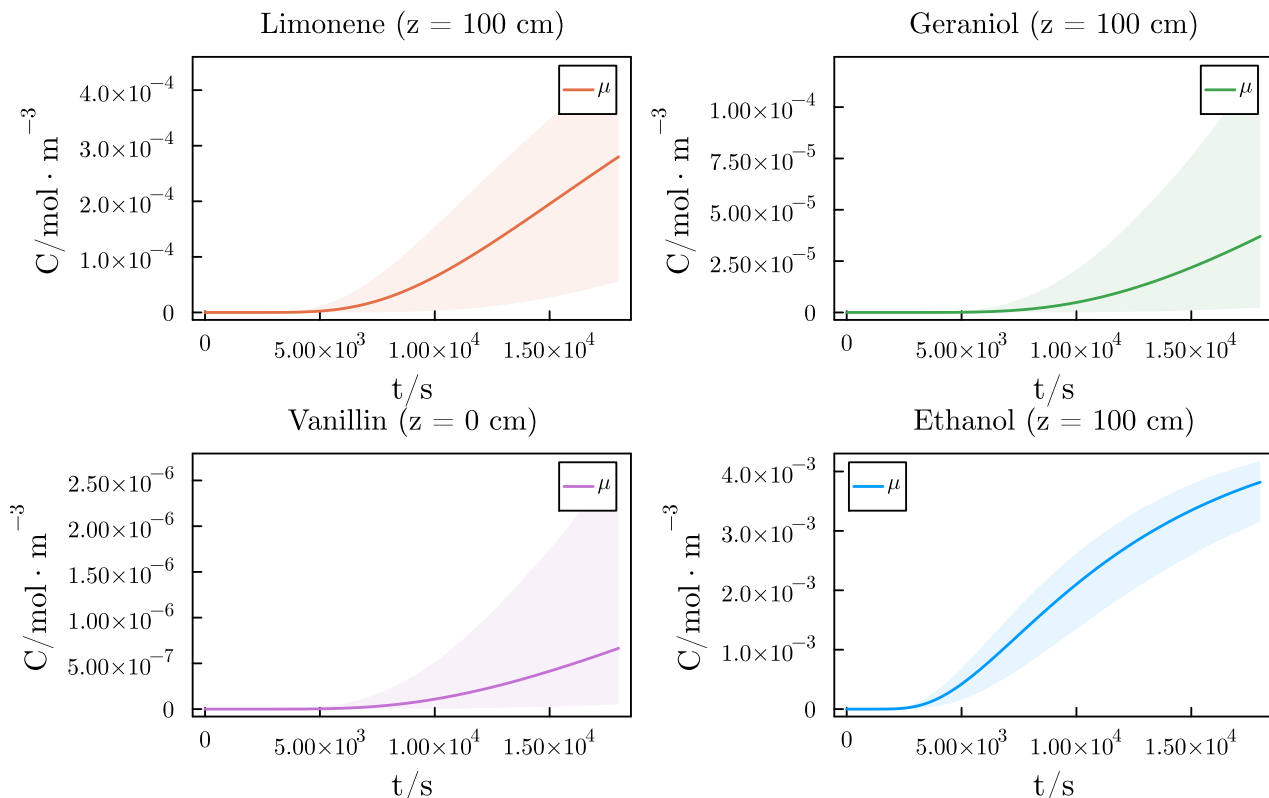


Figure 7.7: Mean gas phase concentrations and 99% credible intervals for all components at $z = 100\text{cm}$

7.4.4 Integration of models for robust fragrance classification

After analyzing the uncertainties in the mechanistic model—specifically, the effects of vapor pressure and diffusivity uncertainties on the gas phase concentration of odorants—we can apply our proposed framework (shown in Figure 7.3). This framework integrates concentration trajectories, the psychophysics model’s uncertainty, and the output probabilities from the odor classifier into a unified structure.

To accomplish the integration, the membership probabilities of 139 olfactory families were estimated for every molecule in the studied fragrance mixture (limonene, geraniol, vanillin, and ethanol). Only the families with probabilities higher than 0.55 were kept as they were highly certain to be present. Figure 7.8 displays each molecule’s corresponding values.

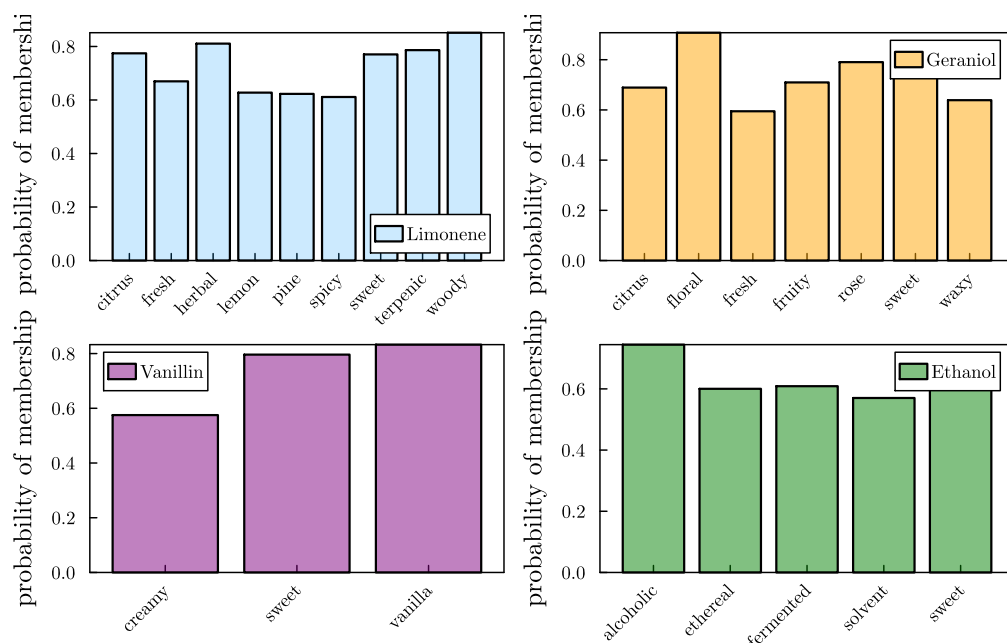


Figure 7.8: Estimated probabilities of membership for each ingredient in the fragrant mixture

These probabilities acted as weight factors to determine each family's composite odor intensity field samples $\psi_f(z, t)$. For the citrus family, as an illustration, the field is computed by taking the product of the membership probability of limonene to the citrus category with the odor intensity field samples of limonene. Subsequently, the result from the product of geraniol's citrus membership probability and its odor intensity field samples is added to this value.

The probability densities associated with each family allow for visualizing the odor intensity field samples. However, with 19 families, it becomes challenging to represent them clearly in a single graph, especially because each point in the z, t grid is a set of 19 samples with thousands of points. As a representation, Figure 7.9 displays a subset, illustrating the estimated densities for 6 odor intensities using kernel density estimation at time $t = 4$ s at a distance of 0.05 m from the releasing source. The intention is to characterize the odor shortly after the release and near the source. It can be seen that the sweet intensity is dominant over the remaining 5 ones displayed.

To better understand the dominant families, we averaged the intensity distributions for each family and displayed them in increasing order of intensity in Figure 7.10. It can be seen that the intensities associated with ethanol are the ones with the highest intensities, and "alcoholic" is the olfactory family with the highest intensity, followed by fermented and ethereal. Creamy and vanilla are the ones with the lowest values.

In addition to determining the intensities of each olfactory family, we can also evaluate which family is most dominant using a probabilistic approach. Specifically, we can estimate the probability of an olfactory family being dominant by counting the relative

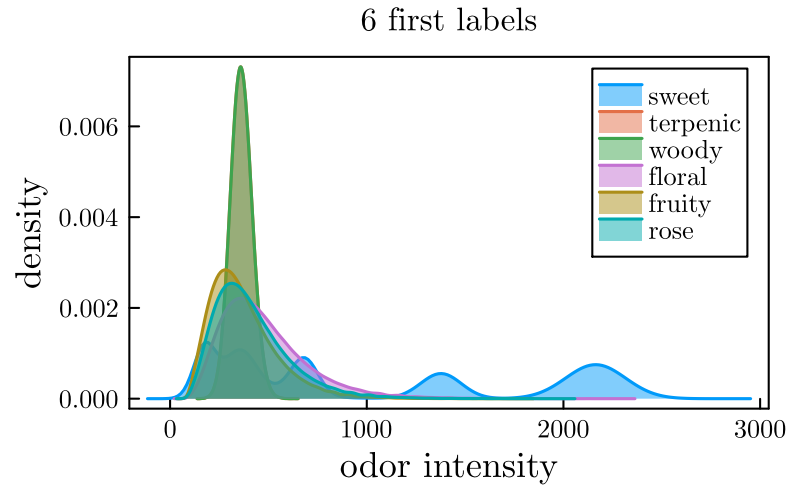


Figure 7.9: Probability density estimation of 6 olfactory families intensities at $z = 0.05$ m, $t = 4$ s

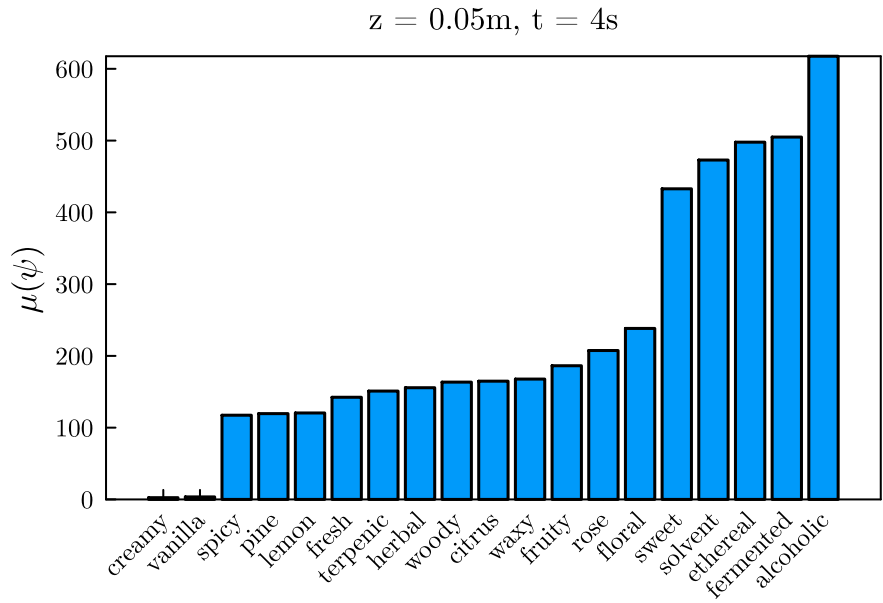


Figure 7.10: Average intensity distributions for each odor family, presented in ascending order of intensity - $z = 0.05$ m, $t = 4$ s.

frequency with which it exhibits the highest intensity. A radar chart can represent the top 7 probabilities (Figure 7.11). The probabilities estimation follows the same order as the intensity rank, with the highest being alcoholic dominating the scent character in about 25% of the samples, followed by ethereal and fermented in 16% of the samples.

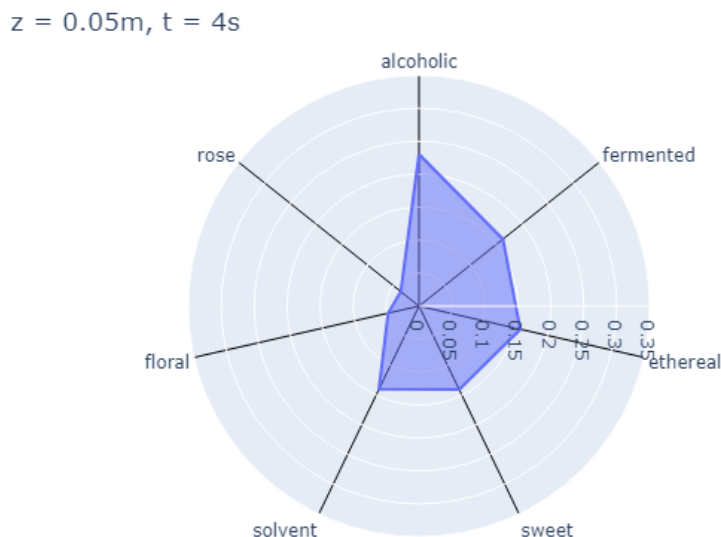


Figure 7.11: Radar chart depicting the top olfactory families based on their relative frequency of exhibiting the highest intensity at $z = 0.05\text{ m}$, $t = 4\text{ s}$.

A similar analysis can be performed to understand the perfume character after a few minutes of the release and at a higher distance from the source. In this case, the same analysis was conducted at 20 cm and 600 s after the release. Figure 7.12 shows the averaged intensities of each family. It can be seen that the intensities are considerably smaller than near the source and that the highest average intensity is the “woody” note, followed by “herbal” and “terpenic” (all similarly intense). The estimation of the dominance probability was also estimated and displayed in Figure 7.13. The probability estimates also follow the same order as the intensity, with about 30% of the samples being dominated by the “woody” note.

By employing these probabilistic statements, this work can be expanded towards robust optimization to maximize the likelihood of attaining a specific olfactory experience, given the known constraints and uncertainties in the model variables.

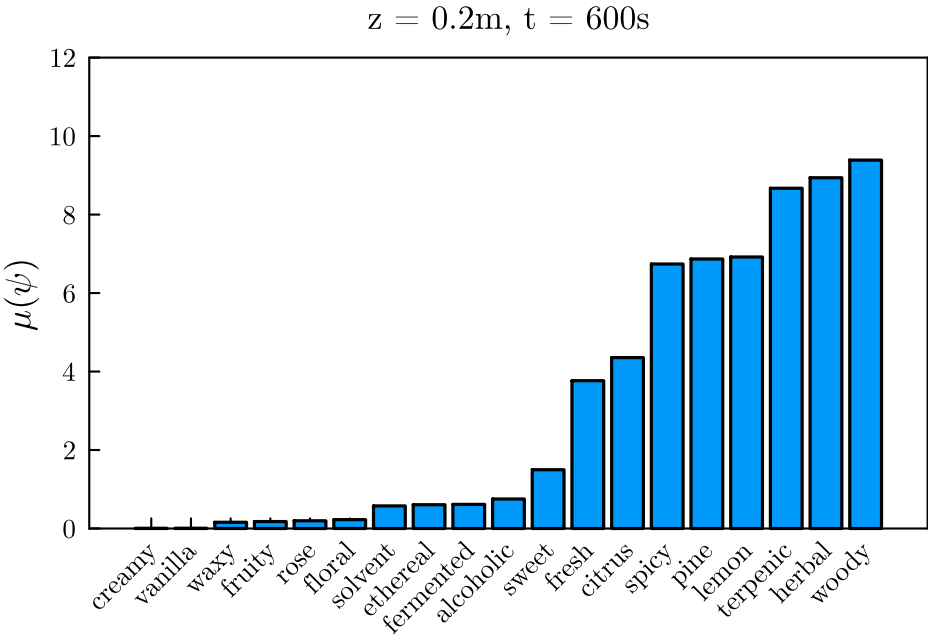


Figure 7.12: Average intensity distributions for each odor family, presented in ascending order of intensity - $z = 0.2\text{ m}$, $t = 600\text{ s}$.

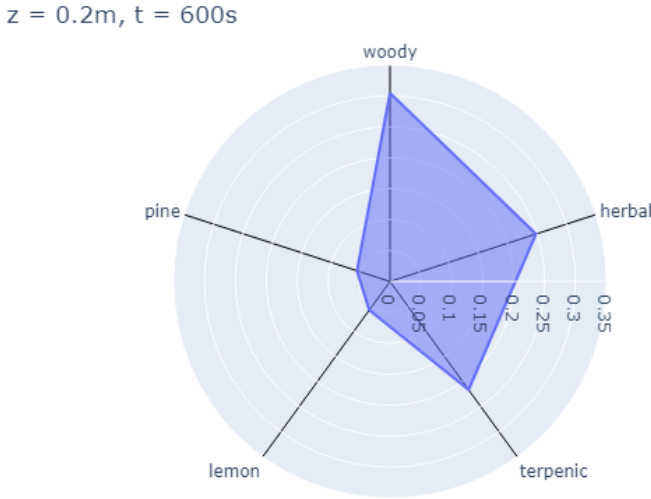


Figure 7.13: Radar chart depicting the top olfactory families based on their relative frequency of exhibiting the highest intensity at $z = 0.2\text{ m}$, $t = 600\text{ s}$.

7.5 Conclusions

In conclusion, our research has demonstrated through a quaternary fragrance case study the feasibility of the proposed probabilistic approach for predicting olfactory families of fragranced mixtures. By systematically identifying, quantifying, and incorporating uncertainties in both the perfume diffusion and odor classification models, we have laid the foundation for more robust fragrance design and formulation. It offers a valuable tool for perfumers and fragrance designers, enabling them to assign probabilities to olfactory responses and customize fragrances with a higher likelihood of achieving desired sensory experiences.

One of our unique contributions lies in integrating uncertainty quantification with the odor classification model, leveraging the output probabilities of graph convolutional neural networks to assign weights to odor intensity probability distribution functions and then transforming samples into probabilities by counting dominance. This innovative method enhances the reliability of the predictions and gives a probabilistic estimate of the fragrance’s dominant smell.

As a result, our work advances the fragrance science field and contributes to the broader domain of probabilistic modeling in sensory perception. Future research could explore refinements to the methodology, such as incorporating additional sources of uncertainty (either epistemic or aleatoric), extending its applicability to other complex odorant mixtures, and testing the predictive capability against experimental data. Ultimately, the probabilistic approach presented here empowers fragrance designers with a data-driven framework to create perfumes that closely align with desired olfactory profiles, marking a significant step forward in the science of fragrance design.

Conclusions and Future Directions

This thesis has examined the integration of mechanistic and data-driven models in chemical engineering, focusing on process systems engineering. The primary objectives were to enhance analytical tools for systems described by differential equations and explore how abundant data integration can refine mechanistic models.

Part I: Mechanistic Models with Limited Data. The first part of this thesis addressed the challenges of modeling adsorption-based separation processes under conditions of limited data availability. Implementing Physics-Informed Neural Networks (PINNs) in Chapter 1 marked a considerable progress in modeling nonlinear column chromatography, showcasing improved resilience to model stiffness. This approach, one of the first of its kind in chemical engineering, particularly in adsorption-based systems, provides a foundation for further exploration in this emerging field. Unlike traditional numerical solvers, PINNs simplify the solution process by efficiently handling spatial-temporal points and avoiding complex discretization schemes. Moreover, PINNs facilitate quick spatio-temporal predictions, promising broader applications in more complex processes like continuous chromatography. Further investigations of PINNs could focus on their application in more complex chemical processes, such as continuous chromatography.

In Chapter 2, hybrid modeling approaches for lithium recovery were developed, showing superior performance over traditional models. In the first, a hybrid modeling framework for developing a dynamic model of the flat-sheet supported liquid membrane (FSSLM) in lithium recovery was introduced. It combined a phenomenological model for system dynamics with a surrogate model for mass transport phenomena, offering an innovative alternative to traditional first-principle modeling. This method allowed for enhanced information extraction about the system, storing it within an artificial neural network. The results indicated that this hybrid approach outperformed mechanistic modeling in precision. Additionally, it enhanced understanding of the system's behavior under various mass transfer scenarios and aided in optimizing process components for improved recovery. Future works could investigate the interpretability of the obtained models. Chapter 3 pioneered using Universal Differential Equations (UDEs) to model multicomponent competitive adsorption systems. The study effectively tackled stiff ordinary differential equations representing the underlying physical processes by applying UDEs to a fixed bed adsorption unit and validating with experimental data. Integrating an Artificial Neural Network

(ANN) allowed for learning from experimental data, leading to more accurate predictions than those from a fully-mechanistic model. The ANN notably outperformed traditional models like the Langmuir isotherm in precision. This initial foray into modeling adsorption in multicomponent systems shows promise, indicating potential broader applications of UDEs in adsorption and other chemical processes. However, further research is needed to harness their capabilities fully.

Chapter 4 extended these concepts to partial differential equations, achieving new accuracy and computational efficiency levels. This chapter explored the numerical challenges in hybrid non-linear advection-diffusion-sorption Partial Differential Equation (PDE) problems, introducing a method for robustly training Artificial Neural Networks (ANNs) for these issues. Key strategies included continuous adjoint sensitivity analysis, orthogonal collocation on finite element PDE discretization, and an adaptive-order, adaptive-time BDF method ODE integrator. Another important contribution involved analyzing ANN predictions using sparse and symbolic regression. An in-silico data set, simulating a single-component non-linear advection-diffusion-sorption system with various adsorption and kinetic models demonstrated the ANN's ability to learn from noisy data and extrapolate effectively. This framework also facilitated kinetic model discovery from single-sensor noisy data. Further studies could stress the uncertainty of the other parameters of the hybrid model, including Peclet number, for instance.

Part II: Data-Abundant Scenarios with Sparse Domain Knowledge. The second part of the thesis shifted focus to scenarios rich in data but with limited domain-specific knowledge. This section demonstrated the potential of integrating machine learning with mechanistic models.

Chapters 5 through 7 exemplified this shift, with Chapter 5 focusing on the efficient formulation of perfumes using deep learning models, Chapter 6 introducing a machine learning model for vapor pressure prediction, and Chapter 7 presenting a method for classifying and predicting odor intensity in perfumery. Chapter 5 addressed the complex challenge of predicting and designing fragranced products by proposing a Deep Learning simulation-optimization method. It introduces a surrogate model trained to predict perfume diffusion and its correlation with perception. The model, significantly more efficient in computation (0.16 ms for predictions compared to 6s for first-principle models), was used to optimize a quaternary perfume mixture. The approach involved developing a single objective function focused on achieving a desired odor spectrum over time and space. Particle swarm optimization (PSO) was employed for its speed and flexibility in optimizing this function. The outcome was an optimal perfume formulation that minimized ethanol odor while highlighting pinene and linalool fragrances consistently over time and space, showcasing the effectiveness of the recurrent neural network surrogate model in fragrance formulation.

Chapter 6 introduced a novel hybrid neural network model, PUFFIN, which accurately and interpretably predicted vapor pressure. By employing transfer learning with boiling

point data and embedding thermodynamic principles into the model, PUFFIN achieved a 50% reduction in mean squared error compared to other neural network methods. A distinctive feature was its ability to extract Antoine-like coefficients from molecular data, demonstrating the benefit of integrating domain knowledge into machine learning. Additionally, PUFFIN excelled in extrapolating vapor pressure data to temperatures beyond its training range, outperforming other benchmark models. These advancements highlighted the potential of incorporating domain-specific knowledge to enhance machine learning predictions, which are applicable in vapor pressure and across various scientific and engineering disciplines. The study underscored the growing importance of physics-informed, interpretable models for accurately predicting complex properties.

Chapter 7 demonstrated, through a quaternary fragrance case study, the effectiveness of a probabilistic approach for predicting olfactory families in fragranced mixtures. By incorporating uncertainties in perfume diffusion and odor classification models, we established a more robust framework for fragrance design. This method can enable perfumers and fragrance designers to assign probabilities to olfactory responses, enhancing the likelihood of achieving desired sensory experiences. Integrating uncertainty quantification with the odor classification model using graph convolutional neural networks marked a significant advancement in fragrance science and probabilistic modeling in sensory perception. Further directions of this work could focus on the experimental validation of olfactory family probability memberships and numerical methods for robust optimization.

A central theme of this thesis is the successful integration of mechanistic and data-driven models. This combination has proven effective in increasing prediction accuracy and reducing computational demands in various applications. The methodologies developed have broad implications, demonstrating their versatility and impact in diverse areas such as perfume design and lithium recovery processes. The methods and models developed in this thesis offer significant potential for application in other areas of chemical engineering. This foundation sets the stage for further exploration into more complex systems, where integrating machine learning with mechanistic modeling could lead to innovative advancements. Despite hybrid models' high predictive power and flexibility, interpretability is still an open issue in these approaches. This could enhance the trust and usability of these models in practical applications, making them more accessible to engineers and decision-makers who are not experts in machine learning. Another further development could lie in applying these models to real-time process optimization and control in industrial settings, revolutionizing how chemical plants operate and potentially increasing efficiency and reducing environmental impact.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: A system for large-scale machine learning, 2016.
- Almeida, R. N., Costa, P., Pereira, J., Cassel, E., and Rodrigues, A. E. Evaporation and permeation of fragrance applied to the skin. *Industrial and Engineering Chemistry Research*, 58:9644–9650, 2019. ISSN 15205045. doi: 10.1021/acs.iecr.9b01004.
- Aouichaoui, A. R. N., Fan, F., Mansouri, S. S., Abildskov, J., and Sin, G. Combining group-contribution concept and graph neural networks toward interpretable molecular property models. *Journal of Chemical Information and Modeling*, 63(3):725–744, 2023. doi: 10.1021/acs.jcim.2c01091. URL <https://doi.org/10.1021/acs.jcim.2c01091>. PMID: 36716461.
- Associates, L. . Flavor and fragrance industry - top 10, 2018.
- Azevedo, S. F. D., Dahm, B., and Oliveira, F. R. Hybrid modelling of biochemical processes: A comparison with the conventional approach. *Computers and Chemical Engineering*, 21, 1997. ISSN 00981354. doi: 10.1016/s0098-1354(97)87593-x.
- Bagge Carlson, F. Montecarlomeasurements.jl: Nonlinear propagation of arbitrary multivariate distributions by means of method overloading, 2020. URL <https://github.com/baggepinnen/MonteCarloMeasurements.jl>.
- Barsainyan, M., Aryan Amit;Schmuker. Open principal odor map. <https://github.com/ARY2260/openpom>, 2023.
- Benzal, G., Kumar, A., Delshams, A., and Sastre, A. M. Mathematical modelling and simulation of cotransport phenomena through flat sheet-supported liquid membranes. *Hydrometallurgy*, 74, 2004. ISSN 0304386X. doi: 10.1016/j.hydromet.2004.01.005.
- Besançon, M., Papamarkou, T., Anthoff, D., Arslan, A., Byrne, S., Lin, D., and Pearson, J. Distributions.jl: Definition and modeling of probability distributions in the juliastats ecosystem. *Journal of Statistical Software*, 98(16):1–30, 2021. ISSN 1548-7660. doi: 10.18637/jss.v098.i16. URL <https://www.jstatsoft.org/v098/i16>.
- Bezanson, J., Karpinski, S., Shah, V. B., and Edelman, A. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.
- Bhatluri, K. K., Manna, M. S., Ghoshal, A. K., and Saha, P. Supported liquid membrane based removal of lead(ii) and cadmium(ii) from mixed feed: Conversion to solid waste

- by precipitation. *Journal of Hazardous Materials*, 299, 2015. ISSN 18733336. doi: 10.1016/j.jhazmat.2015.07.030.
- Bird, R. B., Lightfoot, E. N., and Stewart, W. E. *Transport phenomena*, John Wiley and Sons, Inc., New York. 1961. ISBN 0470115394, 9780470115398. doi: 10.1002/aic.690070245.
- Bisong, E. *Google Colaboratory*. 2019. doi: 10.1007/978-1-4842-4470-8_7.
- Bohorquez, J. F. C., Tovar, L. P., Maciel, M. R. W., Melo, D. C., and Filho, R. M. Surrogate-model-based, particle swarm optimization, and genetic algorithm techniques applied to the multiobjective operational problem of the fluid catalytic cracking process. *Chemical Engineering Communications*, 207:612–631, 2020. ISSN 15635201. doi: 10.1080/00986445.2019.1613230. URL <https://doi.org/10.1080/00986445.2019.1613230>.
- Braake, H. A. T., Can, H. J. V., and Verbruggen, H. B. Semi-mechanistic modeling of chemical processes with neural networks. *Engineering Applications of Artificial Intelligence*, 11:507–515, 1998. ISSN 09521976. doi: 10.1016/s0952-1976(98)00011-6.
- Brunton, S. L., Proctor, J. L., Kutz, J. N., and Bialek, W. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113:3932–3937, 2016. ISSN 10916490. doi: 10.1073/pnas.1517384113.
- Bushdid, C., Magnasco, M. O., Vosshall, L. B., and Keller, A. Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, 343:1370–1372, 3 2014. ISSN 0036-8075. doi: 10.1126/science.1249168. URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1249168>.
- Butler, H. *Poucher’s perfumes, cosmetics and soaps*. Springer Science & Business Media, 2013.
- Cain, W. S., Schiet, F. T., Olsson, M. J., and de Wijk, R. A. Comparison of models of odor interaction. *Chemical senses*, 20(6):625–637, 1995.
- Calkin, R. R. and Jellinek, J. S. *Perfumery: practice and principles*. John Wiley & Sons, 1994.
- Carles, J. A method of creation & perfumery. *Soap, perfumery & cosmetics*, 35:328–335, 1962.
- Chambers, E. and Wolf, M. Sensory testing methods, 1996. URL <https://books.google.no/books?id=c5I8vwEACAAJ>.
- Chapman, S. and Cowling, T. *The Mathematical Theory of Non-uniform Gases: An Account of the Kinetic Theory of Viscosity, Thermal Conduction and Diffusion in Gases*. Cambridge Mathematical Library. Cambridge University Press, 1990. ISBN 9780521408448. URL <https://books.google.no/books?id=Cbp5JP20TrwC>.
- Chollet, F. Keras: The python deep learning library. *Keras.Io*, 2015.

- Chung, Y., Vermeire, F. H., Wu, H., Walker, P. J., Abraham, M. H., and Green, W. H. Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy. *Journal of Chemical Information and Modeling*, 62(3):433–446, 2022. doi: 10.1021/acs.jcim.1c01103. URL <https://doi.org/10.1021/acs.jcim.1c01103>. PMID: 35044781.
- Constantino, D. S., Pereira, C. S., Faria, R. P., Ferreira, A. F., Loureiro, J. M., and Rodrigues, A. E. Synthesis of butyl acrylate in a fixed-bed adsorptive reactor over amberlyst 15. *AIChE Journal*, 61, 2015. ISSN 15475905. doi: 10.1002/aic.14701.
- Cozad, A. and Sahinidis, N. V. A global minlp approach to symbolic regression. *Mathematical Programming*, 170:97–119, 7 2018. ISSN 14364646. doi: 10.1007/s10107-018-1289-x.
- Cranmer, M. Pysr: Fast & parallelized symbolic regression in python/julia, September 2020. URL <http://doi.org/10.5281/zenodo.4041459>.
- Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. Discovering symbolic models from deep learning with inductive biases. *NeurIPS 2020*, 2020.
- Cussler, E., Wagner, A., and Marchal-Heussler, L. Designing chemical products requires more knowledge of perception. *AIChE Journal*, 56(2):283–288, 2010.
- Dandekar, R., Chung, K., Dixit, V., Tarek, M., Garcia-Valadez, A., Vemula, K. V., and Rackauckas, C. Bayesian neural ordinary differential equations, 2020. URL <http://arxiv.org/abs/2012.07244>.
- de Souza Gama, M., Santos, M. S., de Almeida Lima, E. R., Tavares, F. W., and Barreto, A. G. B. A modified poisson-boltzmann equation applied to protein adsorption. *Journal of Chromatography A*, 1531, 2018. ISSN 18733778. doi: 10.1016/j.chroma.2017.11.022.
- Debao, C. Degree of approximation by superpositions of a sigmoidal function. *Approximation Theory and its Applications*, 9:17–28, 1993. ISSN 10009221. doi: 10.1007/BF02836480.
- Di Caprio, U., Degève, J., Hellinckx, P., Waldherr, S., and Leblebici, M. E. Hybridgamma: A thermodynamically consistent framework for hybrid modelling of activity coefficients. *Chemical Engineering Journal*, 475:146104, 2023. ISSN 1385-8947. doi: <https://doi.org/10.1016/j.cej.2023.146104>. URL <https://www.sciencedirect.com/science/article/pii/S1385894723048350>.
- Dobbelaere, M. R., Plehiers, P. P., de Vijver, R. V., Stevens, C. V., and Geem, K. M. V. Machine learning in chemical engineering: Strengths, weaknesses, opportunities, and threats. *Engineering*, 7:1201–1211, 2021. ISSN 20958099. doi: 10.1016/j.eng.2021.03.019.
- Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, feb 2022. ISSN 2522-5839. doi: 10.1038/s42256-021-00438-4. URL <https://www.nature.com/articles/s42256-021-00438-4>.
- Finlayson, B. A. Orthogonal collocation on finite elements. *Chemical Engineering Science*, 30:587–596, 1974. ISSN 03784754. doi: 10.1016/0378-4754(80)90097-X.

- Frazier, P. I. A tutorial on bayesian optimization, 2018. URL <https://arxiv.org/abs/1807.02811>.
- Fuller, E. N., Schettler, P. D., and Giddings, J. C. New method for prediction of binary gas-phase diffusion coefficients. *Industrial & Engineering Chemistry*, 58(5):18–27, May 1966. ISSN 0019-7866. doi: 10.1021/ie50677a007. URL <https://doi.org/10.1021/ie50677a007>.
- Gama, M. S., Barreto, A. G., and Tavares, F. W. The binding interaction of protein on a charged surface using poisson – boltzmann equation : lysozyme adsorption onto sba - 15. *Adsorption*, 2021. ISSN 1572-8757. doi: 10.1007/s10450-021-00344-6. URL <https://doi.org/10.1007/s10450-021-00344-6>.
- Ganaie, I. A., Arora, S., and Kukreja, V. K. Cubic hermite collocation method for solving boundary value problems with dirichlet, neumann, and robin conditions. *International Journal of Engineering Mathematics*, 2014:1–8, 2 2014. ISSN 2356-7007. doi: 10.1155/2014/365209.
- Gilbert, A. What the nose knows: The science of scent in everyday life, 2008.
- Glueckauf, E. and Coates, J. I. 241. theory of chromatography. part iv. the influence of incomplete equilibrium on the front boundary of chromatograms and on the effectiveness of separation. *Journal of the Chemical Society (Resumed)*, page 1315, 1 1947. ISSN 0368-1769. doi: 10.1039/jr9470001315. URL <https://pubs.rsc.org/en/content/articlehtml/1947/jr/jr9470001315>.
- GmbH, D. Vapor pressure of ethanol, 2023. URL www.ddbst.com.
- Gmehling, J. Potential of group contribution methods for the prediction of phase equilibria and excess properties of complex mixtures. *Pure and applied chemistry*, 75(7):875–888, 2003.
- Gmehling, J., Wittig, R., Lohmann, J., and Joh, R. A modified unifac (dortmund) model. 4. revision and extension. *Industrial & Engineering Chemistry Research*, 41(6):1678–1688, Mar 2002. ISSN 0888-5885. doi: 10.1021/ie0108043. URL <https://doi.org/10.1021/ie0108043>.
- Goll, E. S. and Jurs, P. C. Prediction of vapor pressures of hydrocarbons and halo-hydrocarbons from molecular structure with a computational neural network model. *Journal of Chemical Information and Computer Sciences*, 39(6):1081–1089, 1999. doi: 10.1021/ci9900433. URL <https://doi.org/10.1021/ci9900433>.
- Guélat, B., Ströhlein, G., Lattuada, M., and Morbidelli, M. Electrostatic model for protein adsorption in ion-exchange chromatography and application to monoclonal antibodies, lysozyme and chymotrypsinogen a. *Journal of Chromatography A*, 1217, 2010. ISSN 00219673. doi: 10.1016/j.chroma.2010.06.064.
- Guélat, B., Ströhlein, G., Lattuada, M., Delegrange, L., Valax, P., and Morbidelli, M. Simulation model for overloaded monoclonal antibody variants separations in ion-exchange chromatography. *Journal of Chromatography A*, 1253, 2012. ISSN 00219673. doi: 10.1016/j.chroma.2012.06.081.

- Hai, W. Study on thermodynamics and kinetics of adsorption of p-toluidine from aqueous solution by hypercrosslinked polymeric adsorbents. *Environmental Chemistry*, 2004.
- Hardin, R. and Sloane, N. A new approach to the construction of optimal designs. *Journal of Statistical Planning and Inference*, 37(3):339–369, 1993. ISSN 0378-3758. doi: [https://doi.org/10.1016/0378-3758\(93\)90112-J](https://doi.org/10.1016/0378-3758(93)90112-J). URL <https://www.sciencedirect.com/science/article/pii/037837589390112J>.
- Heinzelmann, F. J., Wasan, D. T., and Wilke, C. R. Concentration profiles in stefan diffusion tube. *Industrial and Engineering Chemistry Fundamentals*, 1965. ISSN 01964313. doi: 10.1021/i160013a009.
- Helton, J. C. and Davis, F. J. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering and System Safety*, 81:23–69, 2003. ISSN 09518320. doi: 10.1016/S0951-8320(03)00058-9.
- Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 1997. ISSN 08997667. doi: 10.1162/neco.1997.9.8.1735.
- Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural Networks*, 1989. ISSN 08936080. doi: 10.1016/0893-6080(89)90020-8.
- Jagtap, A. D., Kawaguchi, K., and Karniadakis, G. E. Locally adaptive activation functions with slope recovery for deep and physics-informed neural networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2239):20200334, jul 2020. doi: 10.1098/rspa.2020.0334. URL <https://doi.org/10.1098%2Frspa.2020.0334>.
- JuliusMartensen, Rackauckas, C., et al. Datadrivendiffeq.jl, July 2021. URL <https://doi.org/10.5281/zenodo.5083412>.
- Kahrs, O. and Marquardt, W. The validity domain of hybrid models and its application in process optimization. *Chemical Engineering and Processing: Process Intensification*, 46:1054–1066, 2007. ISSN 02552701. doi: 10.1016/j.cep.2007.02.031.
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 6 2021. ISSN 25225820. doi: 10.1038/s42254-021-00314-5.
- Keller, A. and Vosshall, L. B. Olfactory perception of chemically diverse molecules. *BMC Neuroscience*, pages 1–17, 2016. ISSN 1471-2202. doi: 10.1186/s12868-016-0287-2.
- Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., Mainland, J. D., Ihara, Y., Yu, C. W., Wolfinger, R., Vens, C., leander schietgat, Grave, K. D., Norel, R., Consortium, D. O. P., Stolovitzky, G., Cecchi, G. A., Vosshall, L. B., and pablo meyer. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017. doi: 10.1126/science.aal2014. URL <https://www.science.org/doi/abs/10.1126/science.aal2014>.
- Khosravi, A., Pabon, J., Koury, R., and Machado, L. Using machine learning algorithms to predict the pressure drop during evaporation of r407c. *Applied Thermal Engineering*, 133:361–370, 2018. ISSN 1359-4311. doi: <https://doi.org/10.1016/>

- j.applthermaleng.2018.01.084. URL <https://www.sciencedirect.com/science/article/pii/S1359431117366863>.
- Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization, 2015.
- Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks, 2017.
- Laffort, P. and Dravnieks, A. Several models of suprathreshold quantitative olfactory interaction in humans applied to binary, ternary and quaternary mixtures. *Chemical Senses*, 7(2):153–174, 1982.
- Lagaris, I., Likas, A., and Fotiadis, D. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9(5):987–1000, 1998. doi: 10.1109/72.712178. URL <https://doi.org/10.1109/72.712178>.
- Lansford, J. L., Jensen, K. F., and Barnes, B. C. Physics-informed Transfer Learning for Out-of-sample Vapor Pressure Predictions. *Propellants, Explosives, Pyrotechnics*, 48(4), mar 2023. ISSN 0721-3115. doi: 10.1002/prop.202200265. URL <https://onlinelibrary.wiley.com/doi/10.1002/prop.202200265>.
- Lee, B. K., Mayhew, E. J., Sanchez-Lengeling, B., Wei, J. N., Qian, W. W., Little, K. A., Andres, M., Nguyen, B. B., Moloy, T., Yasonik, J., Parker, J. K., Gerkin, R. C., Mainland, J. D., and Wiltschko, A. B. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661):999–1006, 2023. doi: 10.1126/science.ade4401. URL <https://www.science.org/doi/abs/10.1126/science.ade4401>.
- Lee, C. Y. and Wilke, C. R. Measurements of vapor diffusion coefficient. *Industrial & Engineering Chemistry*, 1954. ISSN 0019-7866. doi: 10.1021/ie50539a046.
- Lee, J., Liu, Z., Chen, W.-H., Wei, J., Kundu, R., Adhikari, R., Rivera, J. A., Gillespie, P. M., Strych, U., Zhan, B., Hotez, P. J., and Bottazzi, M. E. Process development and scale-up optimization of the sars-cov-2 receptor binding domain-based vaccine candidate, rbd219-n1c1. *Applied Microbiology and Biotechnology*, 105:4153–4165, 5 2021. ISSN 0175-7598. doi: 10.1007/s00253-021-11281-3. URL <https://link.springer.com/10.1007/s00253-021-11281-3>.
- Leon, G., Martinez, G., Leon, L., and Guzman, M. A. Separation of cobalt from nickel using novel ultrasound-prepared supported liquid membranes containing cyanex 272 as carrier. *Physicochemical Problems of Mineral Processing*, 52, 2016. ISSN 16431049. doi: 10.5277/ppmp160107.
- Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization, 2018a.
- Li, Z., Liu, Y., Wang, H., Tsai, C. J., Yang, X., Xing, Y., Zhang, C., Xiao, P., and Webley, P. A. A numerical modelling study of so2 adsorption on activated carbons with new rate equations. *Chemical Engineering Journal*, 353:858–866, 2018b. ISSN 13858947. doi: 10.1016/j.cej.2018.07.119. URL <https://doi.org/10.1016/j.cej.2018.07.119>.
- Lin, D., White, J. M., Byrne, S., Bates, D., Noack, A., Pearson, J., Arslan, A., Squire, K., Anthoff, D., Papamarkou, T., Besançon, M., Drugowitsch, J., Schauer, M., and

- other contributors. JuliaStats/Distributions.jl: a Julia package for probability distributions and associated functions, July 2019. URL <https://doi.org/10.5281/zenodo.2647458>.
- Lu, L., Meng, X., Mao, Z., and Karniadakis, G. E. DeepXDE: A deep learning library for solving differential equations. *SIAM Review*, 63(1):208–228, 2021. doi: 10.1137/19M1274067.
- Ma, Y., Dixit, V., Innes, M. J., Guo, X., and Rackauckas, C. A comparison of automatic differentiation and continuous sensitivity analysis for derivatives of differential equation solutions. *2021 IEEE High Performance Extreme Computing Conference, HPEC 2021*, 2021. doi: 10.1109/HPEC49654.2021.9622796.
- Ma, Z. and Guiochon, G. Application of orthogonal collocation on finite elements in the simulation of non-linear chromatography. *Computers & Chemical Engineering*, 15(6):415–426, 1991. ISSN 0098-1354. doi: [https://doi.org/10.1016/0098-1354\(91\)87019-6](https://doi.org/10.1016/0098-1354(91)87019-6). URL <https://www.sciencedirect.com/science/article/pii/0098135491870196>.
- Massey, F. J. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. ISSN 01621459. URL <http://www.jstor.org/stable/2280095>.
- Mata, V. G., Gomes, P. B., and Rodrigues, A. E. Engineering perfumes. *AIChE Journal*, 51:2834–2852, 2005. ISSN 00011541. doi: 10.1002/aic.10530.
- McKay, B., Willis, M., and Barton, G. Steady-state modelling of chemical process systems using genetic programming. *Computers & Chemical Engineering*, 21(9):981–996, 1997. ISSN 0098-1354. doi: [https://doi.org/10.1016/S0098-1354\(96\)00329-8](https://doi.org/10.1016/S0098-1354(96)00329-8). URL <https://www.sciencedirect.com/science/article/pii/S0098135496003298>.
- McKay, M. D., Beckman, R. J., and Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979. ISSN 15372723. doi: 10.1080/00401706.2000.10485979.
- Minceva, M., Ribeiro, A., Silva, M., Ribeiro, A., Rodrigues, A., Graça, N., Santos, J., Pereira, C., and Pais, L. *Principles of Simulated Moving Bed*. 2015. ISBN 9780128020241. doi: 10.1016/b978-0-12-802024-1.00001-x.
- Miranda, L. J. V. Pyswarms: a research toolkit for particle swarm optimization in python. *The Journal of Open Source Software*, 2018. ISSN 2475-9066. doi: 10.21105/joss.00433.
- Mogk, G., Mrziglod, T., and Schuppert, A. Application of hybrid models in chemical industry. In Grievink, J. and van Schijndel, J., editors, *European Symposium on Computer Aided Process Engineering-12*, volume 10 of *Computer Aided Chemical Engineering*, pages 931–936. Elsevier, 2002. doi: [https://doi.org/10.1016/S1570-7946\(02\)80183-3](https://doi.org/10.1016/S1570-7946(02)80183-3). URL <https://www.sciencedirect.com/science/article/pii/S1570794602801833>.
- Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, may 1965. ISSN 0021-9576. doi: 10.1021/c160017a018. URL <https://pubs.acs.org/doi/abs/10.1021/c160017a018>.

- Moriconi, R., Deisenroth, M. P., and Kumar, K. S. S. High-dimensional bayesian optimization using low-dimensional feature spaces, 2019. URL <https://arxiv.org/abs/1902.10675>.
- Nagrath, D., Messac, A., Bequette, B. W., and Cramer, S. M. A hybrid model framework for the optimization of preparative chromatographic processes. *Biotechnology Progress*, 20, 2004. ISSN 87567938. doi: 10.1021/bp034026g.
- Nannoolal, Y., Rarey, J., and Ramjugernath, D. Estimation of pure component properties: Part 3. estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions. *Fluid Phase Equilibria*, 269(1):117–133, 2008. ISSN 0378-3812. doi: <https://doi.org/10.1016/j.fluid.2008.04.020>. URL <https://www.sciencedirect.com/science/article/pii/S0378381208001611>.
- Narayanan, H., Luna, M., Sokolov, M., Arosio, P., Butté, A., and Morbidelli, M. Hybrid models based on machine learning and an increasing degree of process knowledge: Application to capture chromatographic step. *Industrial & Engineering Chemistry Research*, page acs.iecr.1c01317, 7 2021a. ISSN 0888-5885. doi: 10.1021/acs.iecr.1c01317.
- Narayanan, H., Seidler, T., Luna, M. F., Sokolov, M., Morbidelli, M., and Butté, A. Hybrid models for the simulation and prediction of chromatographic processes for protein capture. *Journal of Chromatography A*, 1650:462248, 2021b. ISSN 18733778. doi: 10.1016/j.chroma.2021.462248.
- Narayanan, H., Cruz Bournazou, M. N., Guillén Gosálbez, G., and Butté, A. Functional-hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. *Chemical Engineering Journal*, 430:133032, 2022. ISSN 1385-8947. doi: <https://doi.org/10.1016/j.cej.2021.133032>. URL <https://www.sciencedirect.com/science/article/pii/S1385894721046088>.
- Nicoud, L., Owczarz, M., Arosio, P., and Morbidelli, M. A multiscale view of therapeutic protein aggregation: A colloid science perspective. *Biotechnology Journal*, 10:367–378, 3 2015. ISSN 18607314. doi: 10.1002/biot.201400858.
- Nocedal, J. and Wright, S. J. *Numerical Optimization Second Edition*. 2006. ISBN 2006923897.
- Nogueira, I., Fontes, C., Sartori, I., Pontes, K., and Embiruçu, M. A model-based approach to quality monitoring of a polymerization process without online measurement of product specifications. *Computers and Industrial Engineering*, 106:123–136, 2017. ISSN 03608352. doi: 10.1016/j.cie.2017.01.030. URL <http://dx.doi.org/10.1016/j.cie.2017.01.030>.
- Nogueira, I. B. R., Santana, V. V., Ribeiro, A. M., and Rodrigues, A. E. Using scientific machine learning to develop universal differential equation for multicomponent adsorption separation systems. *The Canadian Journal of Chemical Engineering*, 7 2022. ISSN 1939-019X. doi: 10.1002/CJCE.24495. URL <https://onlinelibrary.wiley.com/doi/full/10.1002/cjce.24495>.
- Nogueira, I. B., Ribeiro, A. M., Requião, R., Pontes, K. V., Koivisto, H., Rodrigues, A. E., and Loureiro, J. M. A quasi-virtual online analyser based on an artificial neural networks and offline measurements to predict purities of raffinate/extract in simulated moving bed

- processes. *Applied Soft Computing Journal*, 67:29–47, 2018. ISSN 15684946. doi: 10.1016/j.asoc.2018.03.001. URL <https://doi.org/10.1016/j.asoc.2018.03.001>.
- Nogueira, I. B., Faria, R. P., Requião, R., Koivisto, H., Martins, M. A., Rodrigues, A. E., Loureiro, J. M., and Ribeiro, A. M. Chromatographic studies of n-propyl propionate: Adsorption equilibrium, modelling and uncertainties determination. *Computers and Chemical Engineering*, 119, 2019. ISSN 00981354. doi: 10.1016/j.compchemeng.2018.09.020.
- Ohe, S. A prediction method of vapor pressures by using boiling point data. *Fluid Phase Equilibria*, 501:112078, dec 2019. ISSN 03783812. doi: 10.1016/j.fluid.2019.01.018. URL <https://linkinghub.elsevier.com/retrieve/pii/S0378381219300275>.
- Olsson, M. J. An integrated model of intensity and quality of odor mixtures. *Annals of the New York Academy of Sciences*, 855(1):837–840, 1998.
- Pan, I., Mason, L. R., and Matar, O. K. Data-centric engineering: integrating simulation, machine learning and statistics. challenges and opportunities. *Chemical Engineering Science*, 249:117271, 2022. ISSN 00092509. doi: 10.1016/j.ces.2021.117271.
- Peydayesh, M., Esfandyari, G. R., Mohammadi, T., and Alamdari, E. K. Pertraction of cadmium and zinc ions using a supported liquid membrane impregnated with different carriers. *Chemical Papers*, 67, 2013. ISSN 03666352. doi: 10.2478/s11696-013-0310-3.
- Poling, B. E., Prausnitz, J. M., and O’Connell, J. P. *The properties of gases and liquids, Fifth Edition*. 2001a. ISBN 0071499997. doi: 10.1002/aic.690240634.
- Poling, B. E., Prausnitz, J. M., and O’connell, J. P. *Properties of gases and liquids*. McGraw-Hill Education, 2001b.
- Praditia, T., Karlbauer, M., Otte, S., Oladyshkin, S., Butz, M. V., and Nowak, W. Finite volume neural network: Modeling subsurface contaminant transport, 4 2021. URL <http://arxiv.org/abs/2104.06010>.
- Psichogios, D. C. and Ungar, L. H. A hybrid neural network-first principles approach to process modeling. *AIChE Journal*, 38:1499–1511, 1992. ISSN 15475905. doi: 10.1002/aic.690381003.
- Rackauckas, C. and Nie, Q. Differentialequations.jl—a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, 5(1), 2017.
- Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., and Edelman, A. Universal differential equations for scientific machine learning, 1 2020. ISSN 2331-8422. URL <http://arxiv.org/abs/2001.04385>.
- Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 10902716. doi: 10.1016/j.jcp.2018.10.045.
- Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., and Wu, Z. *Deep Learning for the Life Sciences*. O’Reilly Media, 2019. <https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837>.

- Rasmuson, A., Andersson, B., Olsson, L., and Andersson, R. *Introduction*, chapter 1, page 1–9. Cambridge University Press, 2014. doi: 10.1017/CBO9781107279124.002.
- Rebello, C. M., Martins, M. A. F., Loureiro, J. M., Rodrigues, A. E., Ribeiro, A. M., and Nogueira, I. B. R. From an optimal point to an optimal region: A novel methodology for optimization of multimodal constrained problems and a novel constrained sliding particle swarm optimization strategy. *Mathematics*, 9(15), 2021. ISSN 2227-7390. doi: 10.3390/math9151808. URL <https://www.mdpi.com/2227-7390/9/15/1808>.
- Regufe, M. J., Ribeiro, A. M., Ferreira, A., and Loureiro, J. M. Complete and simplified modelling of flat sheet supported liquid membranes for the extraction of metal ions., 2021a.
- Regufe, M. J., Santana, V. V., Ferreira, A. F. P., Ribeiro, A. M., Loureiro, J. M., and Nogueira, I. B. R. A hybrid modeling framework for membrane separation processes: Application to lithium-ion recovery from batteries. *Processes*, 9(11), 2021b. ISSN 2227-9717. doi: 10.3390/pr9111939. URL <https://www.mdpi.com/2227-9717/9/11/1939>.
- Rodrigues, A. E., Nogueira, I., and Faria, R. P. V. Perfume and flavor engineering: A chemical engineering perspective. *Molecules*, 26, 2021. ISSN 1420-3049. doi: 10.3390/molecules26113095. URL <https://www.mdpi.com/1420-3049/26/11/3095>.
- Ruthven, D. M. *Principles of Adsorption and Adsorption Processes*. JohnWiley & Sons, 1984.
- Sanchez-Lengeling, B., Wei, J. N., Lee, B. K., Gerkin, R. C., Aspuru-Guzik, A., and Wiltschko, A. B. Machine learning for scent: Learning generalizable perceptual representations of small molecules, 10 2019a. URL <http://arxiv.org/abs/1910.10685>.
- Sanchez-Lengeling, B., Wei, J. N., Lee, B. K., Gerkin, R. C., Aspuru-Guzik, A., and Wiltschko, A. B. Machine learning for scent: Learning generalizable perceptual representations of small molecules, 2019b.
- Sansana, J., Joswiak, M. N., Castillo, I., Wang, Z., Rendall, R., Chiang, L. H., and Reis, M. S. Recent trends on hybrid modeling for industry 4.0. *Computers and Chemical Engineering*, 151, 8 2021. ISSN 00981354. doi: 10.1016/j.compchemeng.2021.107365.
- Santana, V. V., Gama, M. S., Loureiro, J. M., Rodrigues, A. E., Ribeiro, A. M., Tavares, F. W., Barreto, A. G., and Nogueira, I. B. R. A first approach towards adsorption-oriented physics-informed neural networks: Monoclonal antibody adsorption performance on an ion-exchange column as a case study. *ChemEngineering*, 6, 2022. ISSN 2305-7084. doi: 10.3390/chemengineering6020021. URL <https://www.mdpi.com/2305-7084/6/2/21>.
- Santana, V. V., Costa, E., Rebello, C. M., Ribeiro, A. M., Rackauckas, C., and Nogueira, I. B. Efficient hybrid modeling and sorption model discovery for non-linear advection-diffusion-sorption systems: A systematic scientific machine learning approach. *Chemical Engineering Science*, 282:119223, 2023. ISSN 0009-2509. doi: <https://doi.org/10.1016/j.ces.2023.119223>. URL <https://www.sciencedirect.com/science/article/pii/S0009250923007790>.

- Santana, V. V., Rebello, C. M., Queiroz, L. P., Ribeiro, A. M., Shardt, N., and Nogueira, I. B. Puffin: A path-unifying feed-forward interfaced network for vapor pressure prediction. *Chemical Engineering Science*, 286:119623, 2024. ISSN 0009-2509. doi: <https://doi.org/10.1016/j.ces.2023.119623>. URL <https://www.sciencedirect.com/science/article/pii/S000925092301179X>.
- Santana, V. V., Martins, M. A., Loureiro, J. M., Ribeiro, A. M., Rodrigues, A. E., and Nogueira, I. B. Optimal fragrances formulation using a deep learning neural network architecture: A novel systematic approach. *Computers and Chemical Engineering*, 150: 107344, 4 2021. ISSN 00981354. doi: 10.1016/j.compchemeng.2021.107344.
- Schwaab, M., Biscaia, E. C., Monteiro, J. L., and Pinto, J. C. Nonlinear parameter estimation through particle swarm optimization. *Chemical Engineering Science*, 63: 1542–1552, 2008. ISSN 00092509. doi: 10.1016/j.ces.2007.11.024.
- Schweidtmann, A. M., Rittig, J. G., König, A., Grohe, M., Mitsos, A., and Dahmen, M. Graph neural networks for prediction of fuel ignition quality. *Energy & Fuels*, 34(9): 11395–11407, 2020. doi: 10.1021/acs.energyfuels.0c01533. URL <https://doi.org/10.1021/acs.energyfuels.0c01533>.
- Schweidtmann, A. M., Esche, E., Fischer, A., Kloft, M., Repke, J. U., Sager, S., and Mitsos, A. Machine learning in chemical engineering: A perspective. *Chemie-Ingenieur-Technik*, pages 2029–2039, 2021. ISSN 15222640. doi: 10.1002/cite.202100083.
- Sobol', I. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. ISSN 0041-5553. doi: [https://doi.org/10.1016/0041-5553\(67\)90144-9](https://doi.org/10.1016/0041-5553(67)90144-9). URL <https://www.sciencedirect.com/science/article/pii/0041555367901449>.
- Stephenson, R. M. and Malanowski, S. *Handbook of the Thermodynamics of Organic Compounds*. Springer Netherlands, 1987. doi: 10.1007/978-94-009-3173-2.
- Stevens, S. S. On the psychophysical law. *Psychological Review*, 1957. ISSN 0033295X. doi: 10.1037/h0046162.
- Teixeira, M. A., Rodríguez, O., Mata, V. G., and Rodrigues, A. E. The diffusion of perfume mixtures and the odor performance. *Chemical Engineering Science*, 64(11):2570–2589, 2009a. ISSN 0009-2509. doi: <https://doi.org/10.1016/j.ces.2009.01.064>. URL <https://www.sciencedirect.com/science/article/pii/S0009250909000700>.
- Teixeira, M. A., Rodríguez, O., Mata, V. G., and Rodrigues, A. E. The diffusion of perfume mixtures and the odor performance. *Chemical Engineering Science*, 64:2570–2589, 2009b. ISSN 00092509. doi: 10.1016/j.ces.2009.01.064.
- Teixeira, M. A., Rodríguez, O., and Rodrigues, A. E. The perception of fragrance mixtures: A comparison of odor intensity models. *Aiche Journal*, 56:1090–1106, 2010a. URL <https://api.semanticscholar.org/CorpusID:97192105>.
- Teixeira, M. A., Rodríguez, O., and Rodrigues, A. E. Perfumery radar: A predictive tool for perfume family classification. *Industrial and Engineering Chemistry Research*, 49: 11764–11777, 2010b. ISSN 08885885. doi: 10.1021/ie101161v.

- Teixeira, M. A., Rodríguez, O., Mota, F. L., Macedo, E. A., and Rodrigues, A. E. Evaluation of group-contribution methods to predict vle and odor intensity of fragrances. *Industrial & Engineering Chemistry Research*, 50(15):9390–9402, 2011. doi: 10.1021/ie200290r. URL <https://doi.org/10.1021/ie200290r>.
- Teixeira, M. A., Rodríguez, O., and Rodrigues, A. E. Diffusion and performance of fragranced products: Prediction and validation. *AIChE Journal*, 59:3943–3957, 10 2013a. ISSN 00011541. doi: 10.1002/aic.14106.
- Teixeira, M. A., Rodríguez, O., and Rodrigues, A. E. Diffusion and performance of fragranced products: Prediction and validation. *AIChE Journal*, 2013b. ISSN 00011541. doi: 10.1002/aic.14106.
- Teixeira, M. A., Barrault, L., Rodríguez, O., Carvalho, C. C., and Rodrigues, A. E. Perfumery radar 2.0: A step toward fragrance design and classification. *Industrial & Engineering Chemistry Research*, 53(21):8890–8912, 2014. doi: 10.1021/ie403968w. URL <https://doi.org/10.1021/ie403968w>.
- Teixeira, M. A. A. Perfume performance and classification: Perfumery quaternary-quinary diagram (pq2d®) and perfumery radar. *Futures*, pages 725–752, 2011. doi: 10.1002/fut.
- Thomson, G. W. The antoine equation for vapor-pressure data. *Chemical Reviews*, 38(1):1–39, 1946. doi: 10.1021/cr60119a001. URL <https://doi.org/10.1021/cr60119a001>. PMID: 21016992.
- Tu, C.-H. Group-contribution method for the estimation of vapor pressures. *Fluid Phase Equilibria*, 99:105–120, aug 1994. ISSN 03783812. doi: 10.1016/0378-3812(94)80025-1. URL <https://linkinghub.elsevier.com/retrieve/pii/0378381294800251>.
- Ulas, S. and Diwekar, U. M. Thermodynamic uncertainties in batch processing and optimal control. *Computers & Chemical Engineering*, 28(11):2245–2258, 2004. ISSN 0098-1354. doi: <https://doi.org/10.1016/j.compchemeng.2004.04.001>. URL <https://www.sciencedirect.com/science/article/pii/S0098135404001061>.
- van Gemert, L. Odour thresholds: compilations of odour threshold values in air, water and other media, 2011. URL <https://books.google.no/books?id=ZfC6twAACAAJ>.
- Vardanega, T. Lecture notes in computer science: Preface, 2011. ISSN 03029743.
- Vermeire, F. H. and Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, 2021. ISSN 1385-8947. doi: <https://doi.org/10.1016/j.cej.2021.129307>. URL <https://www.sciencedirect.com/science/article/pii/S1385894721008925>.
- Vermeire, F. H., Chung, Y., and Green, W. H. Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. *Journal of the American Chemical Society*, 144(24):10785–10797, 2022. doi: 10.1021/jacs.2c01768. URL <https://doi.org/10.1021/jacs.2c01768>. PMID: 35687887.
- von Rueden, L., Mayer, S., Sifa, R., Bauckhage, C., and Garcke, J. Combining machine learning and simulation to a hybrid modelling approach: Current and future directions. *Lecture Notes in Computer Science*, 12080 LNCS:548–560, 2020. ISSN 16113349. doi: 10.1007/978-3-030-44584-3_43.

- von Stosch, M., Oliveira, R., Peres, J., and de Azevedo, S. F. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers and Chemical Engineering*, 60:86–101, 2014. ISSN 00981354. doi: 10.1016/j.compchemeng.2013.08.008.
- Walker, P. J., Yew, H.-W., and Riedemann, A. Clapeyron.jl: An extensible, open-source fluid thermodynamics toolkit. *Industrial & Engineering Chemistry Research*, 61(20): 7130–7153, May 2022. ISSN 0888-5885. doi: 10.1021/acs.iecr.2c00326. URL <https://doi.org/10.1021/acs.iecr.2c00326>.
- Wang, D., Chen, Q., Hu, J., Fu, M., and Luo, Y. High flux recovery of copper(ii) from ammoniacal solution with stable sandwich supported liquid membrane. *Industrial and Engineering Chemistry Research*, 54, 2015. ISSN 15205045. doi: 10.1021/acs.iecr.5b00297.
- Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL <https://doi.org/10.1021/ci00057a005>.
- Winter, B., Winter, C., Esper, T., Schilling, J., and Bardow, A. Spt-nrtl: A physics-guided machine learning model to predict thermodynamically consistent activity coefficients. *Fluid Phase Equilibria*, 568:113731, 2023. ISSN 0378-3812. doi: <https://doi.org/10.1016/j.fluid.2023.113731>. URL <https://www.sciencedirect.com/science/article/pii/S0378381223000134>.
- Yaffe, D. and Cohen, Y. Neural network based temperature-dependent quantitative structure property relations (qsprs) for predicting vapor pressure of hydrocarbons. *Journal of Chemical Information and Computer Sciences*, 41(2):463–477, 2001. doi: 10.1021/ci000462w. URL <https://doi.org/10.1021/ci000462w>. PMID: 11277738.
- Zaheri, P., Ghassabzadeh, H., Abolghasemi, H., Maraghe, M. G., and Mohammadi, T. Facilitated transport of europium through supported liquid membrane using cyanex272 as carrier and mass transfer modelling. *Canadian Journal of Chemical Engineering*, 95, 2017. ISSN 1939019X. doi: 10.1002/cjce.22686.
- Zander, H. J., Dittmeyer, R., and Wagenliuber, J. Dynamic modeling of chemical reaction systems with neural networks and hybrid models. *Chemical Engineering and Technology*, 22, 1999. ISSN 09307516. doi: 10.1002/(SICI)1521-4125(199907)22:7<571::AID-CEAT571>3.0.CO;2-5.
- Zang, Q., Mansouri, K., Williams, A. J., Judson, R. S., Allen, D. G., Casey, W. M., and Kleinstreuer, N. C. In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *Journal of Chemical Information and Modeling*, 57(1):36–49, 2017a. doi: 10.1021/acs.jcim.6b00625. URL <https://doi.org/10.1021/acs.jcim.6b00625>. PMID: 28006899.
- Zang, Q., Mansouri, K., Williams, A. J., Judson, R. S., Allen, D. G., Casey, W. M., and Kleinstreuer, N. C. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *Journal of Chemical Information and Modeling*, 57(1):36–49, jan 2017b. ISSN 1549-9596. doi: 10.1021/acs.jcim.6b00625. URL <https://pubs.acs.org/doi/10.1021/acs.jcim.6b00625>.

- Zante, G., Boltoeva, M., Masmoudi, A., Barillon, R., and Trébouet, D. Lithium extraction from complex aqueous solutions using supported ionic liquid membranes. *Journal of Membrane Science*, 580, 2019. ISSN 18733123. doi: 10.1016/j.memsci.2019.03.013.
- Zante, G., Boltoeva, M., Masmoudi, A., Barillon, R., and Trébouet, D. Selective separation of cobalt and nickel using a stable supported ionic liquid membrane. *Separation and Purification Technology*, 252, 2020. ISSN 18733794. doi: 10.1016/j.seppur.2020.117477.
- Zhang, L., Mao, H., Liu, L., Du, J., and Gani, R. A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. *Computers and Chemical Engineering*, 115, 2018. ISSN 00981354. doi: 10.1016/j.compchemeng.2018.04.018.
- Zhang, S., Liu, Y., and Xie, L. Molecular mechanics-driven graph neural network with multiplex graph for molecular structures, 2020a.
- Zhang, S., Liu, Y., and Xie, L. Efficient and accurate physics-aware multiplex graph neural networks for 3d small molecules and macromolecule complexes, 2022.
- Zhang, X., Zhou, T., and Ng, K. M. Optimization-based cosmetic formulation: Integration of mechanistic model, surrogate model, and heuristics. *AIChE Journal*, 2020b. ISSN 0001-1541. doi: 10.1002/aic.17064.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. doi: 10.1109/JPROC.2020.3004555.

Appendix A

Chapter 4

A.1 Fully mechanistic model

A fully-mechanistic model for column chromatography model was implemented to create the data set, also using the method of lines with cubic Hermite polynomial in OCFEM. 42 evenly-spaced finite elements were used in the discretization, and the ODE integrated with the FBDF method with 1×10^{-6} absolute and relative tolerances. The mechanistic model assumes a plug flow, negligible pressure drop, isothermal operation, and uniform distribution radial with no variation in the superficial flow velocity throughout the bed.

The adsorption uptake rate, however, can take three forms: LDF, quadratic, and improved LDF. The LDF assumes that the uptake rate of the adsorbate is proportional to the difference between the species at the outer surface of the particle (equilibrium adsorption capacity) and its average concentration with the particle. The quadratic assumes that the uptake rate is proportional to the difference of concentrations squared. The improved LDF is a proposed modification where the uptake rate is corrected for low solid phase concentrations. In this way, the fully mechanistic model can be written as:

$$\frac{\partial c}{\partial t^*} = -\frac{1-\varepsilon}{\varepsilon} g_i(q, q^*) \tau_s - \frac{\partial c}{\partial x^*} + \frac{1}{Pe} \frac{\partial^2 c}{\partial x^{*2}} \quad (\text{A.1})$$

$$\frac{\partial q}{\partial t^*} = g_i(q, q^*) \tau_s \quad (\text{A.2})$$

$$\frac{\partial c(x^* = 1, \forall t)}{\partial x^*} = 0 \quad (\text{A.3})$$

$$\frac{\partial c(x^* = 0, \forall t)}{\partial x^*} = Pe(c - c_{inlet}) \quad (\text{A.4})$$

$$c(x^* \in (0, 1), t^* = 0) = c_0 \quad (\text{A.5})$$

$$q(x^* \in (0, 1), t^* = 0) = q^*(c_0) \quad (\text{A.6})$$

$$q^* = f_j(c, p) \quad (\text{A.7})$$

$$\tau_s = L/v \quad (\text{A.8})$$

Where $g_i(q, q^*)$ and $f_j(c, p)$ can be:

$$g_{LDF} = 0.22(q^* - q) \quad (\text{A.9})$$

$$g_{\text{Vermeulen's}} = 0.22 \frac{q^{*2} - q^2}{2.0q} \quad (\text{A.10})$$

$$g_{\text{iLDF}} = 0.22(q^* + 0.2789q^* e^{\frac{-q}{2q^*}} - q) \quad (\text{A.11})$$

$$f_{\text{Langmuir}} = \frac{55.54 \times 1.8c}{1.0 + 1.8c} \quad (\text{A.12})$$

$$f_{\text{Sips}} = \frac{55.54 \times 1.8c^{1.5}}{1.0 + 1.8c^{1.5}} \quad (\text{A.13})$$

in which, $t^* = t/\tau_s$ is the dimensionless time, $x^* = x/L$ is the fixed bed dimensionless length, the axial distance from column inlet, c is the liquid phase concentration, q is the solid phase concentration, $q^*(c, p)$ is the solid phase equilibrium concentration (calculated from a known isotherm f with known parameters p), ε is the bed porosity, v is the interstitial velocity, Pe is the Peclet number.

A.2 UDE Training and test set performance - Sips isotherm and Langmuir isotherm

This section of the appendix contains the train and test set performance for Sips and Langmuir isotherm for 5.5-0.75-9.33 inlet concentrations.

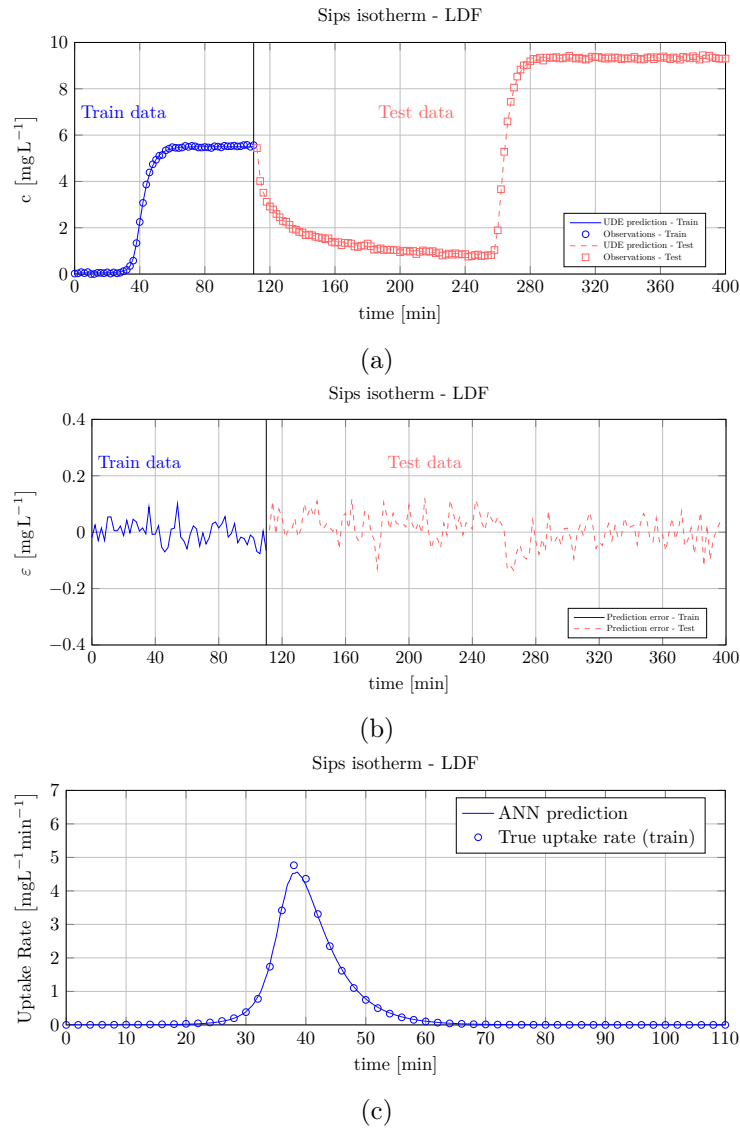


Figure A.1: Results for Sips isotherm and LDF kinetics - 0.5 min^{-1} sampling rate (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets. (c) UDE prediction of uptake rate ($x^* = 1$) compared to ground truth.

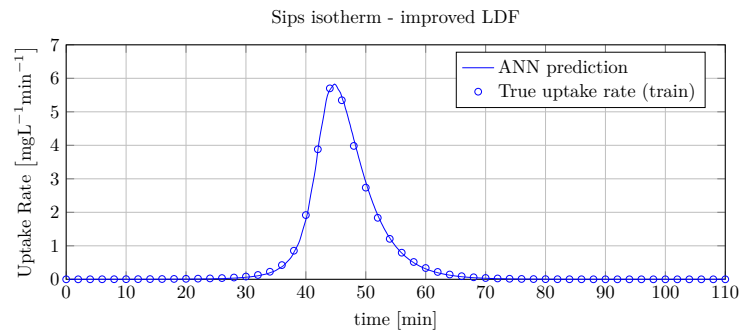
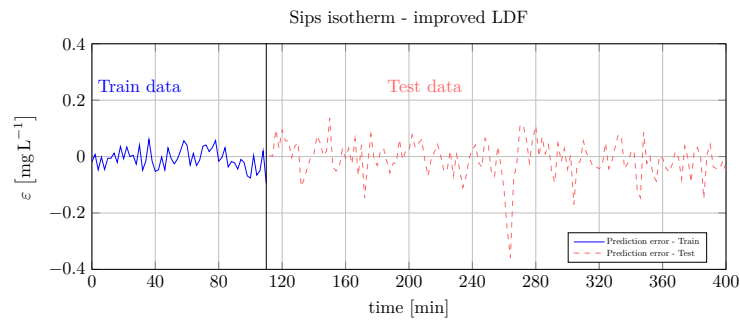
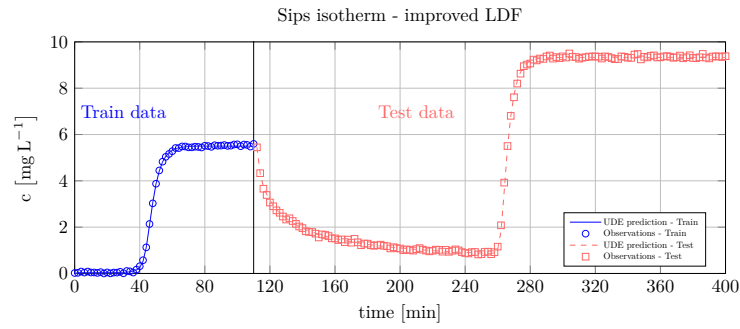


Figure A.2: Results for Sips isotherm and improved LDF kinetics - 0.5 min^{-1} sampling rate (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets. (c) UDE prediction of uptake rate ($x^* = 1$) compared to ground truth.

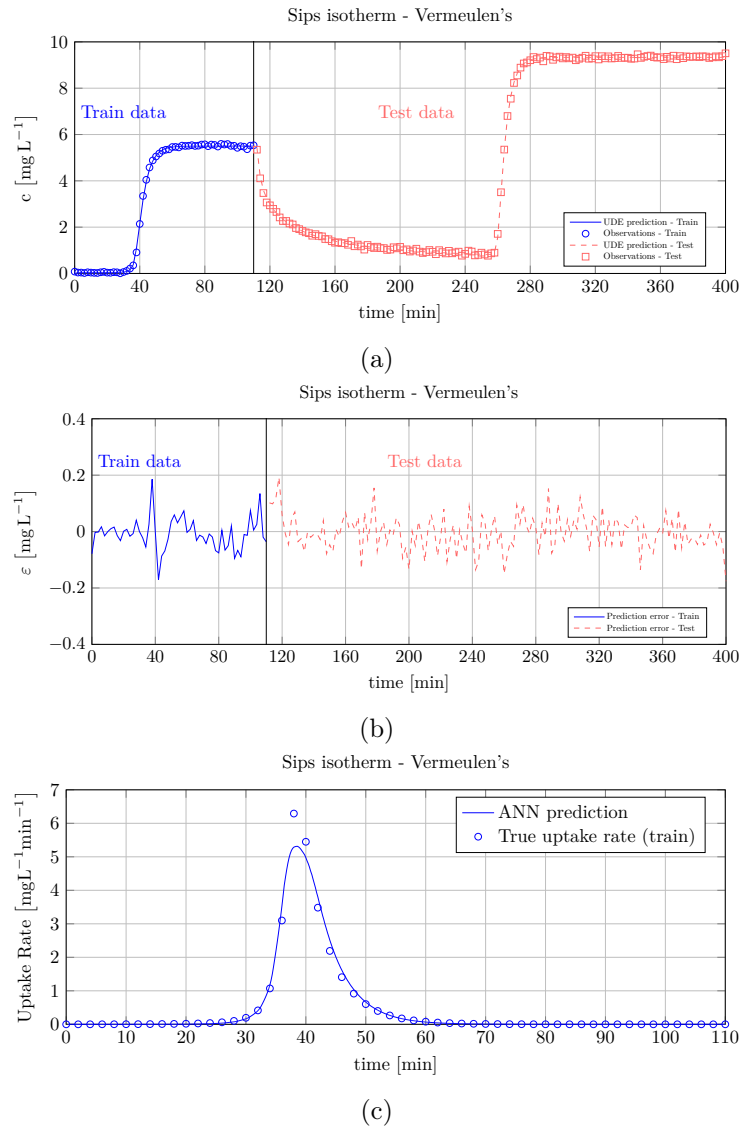


Figure A.3: Results for Sips isotherm and Vermeulen's kinetics - 0.5 min⁻¹ sampling rate (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets. (c) UDE prediction of uptake rate ($x^* = 1$) compared to ground truth.

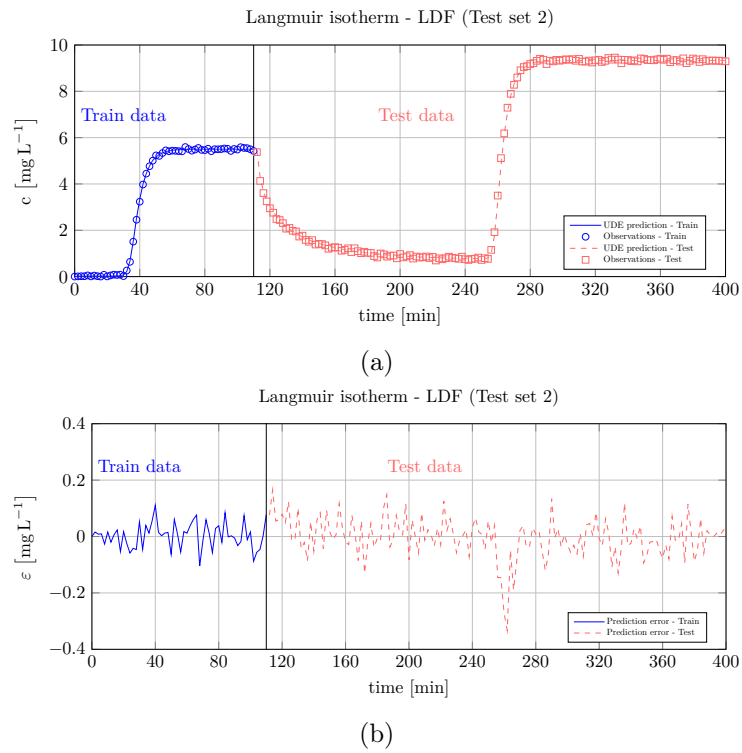


Figure A.4: Results for Langmuir isotherm and LDF kinetics - 0.5 min^{-1} sampling rate and 5.5-0.75-9.33 test set inlet concentrations (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets.

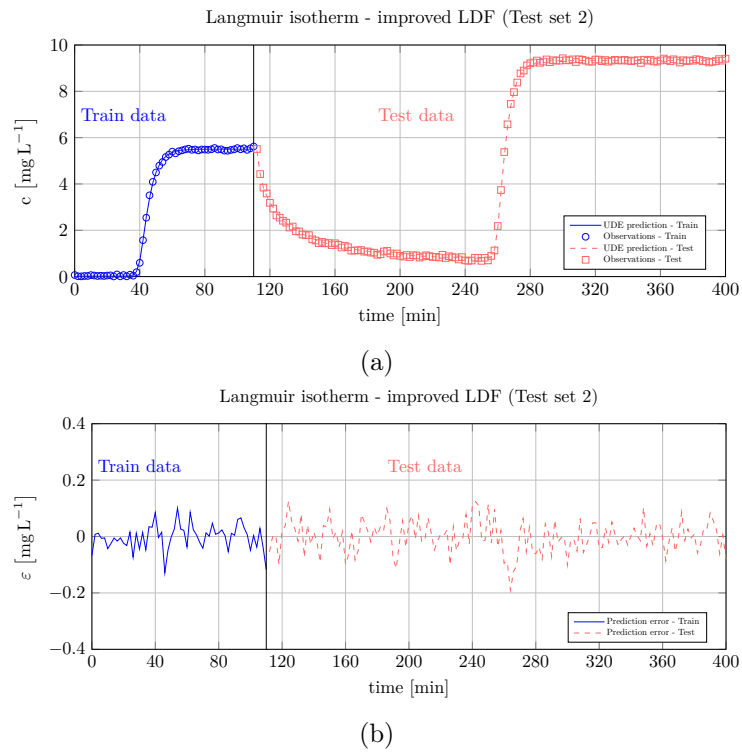


Figure A.5: Results for Sips isotherm and improved LDF kinetics - 0.5 min^{-1} sampling rate and 5.5-0.75-9.33 test set inlet concentrations (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets.

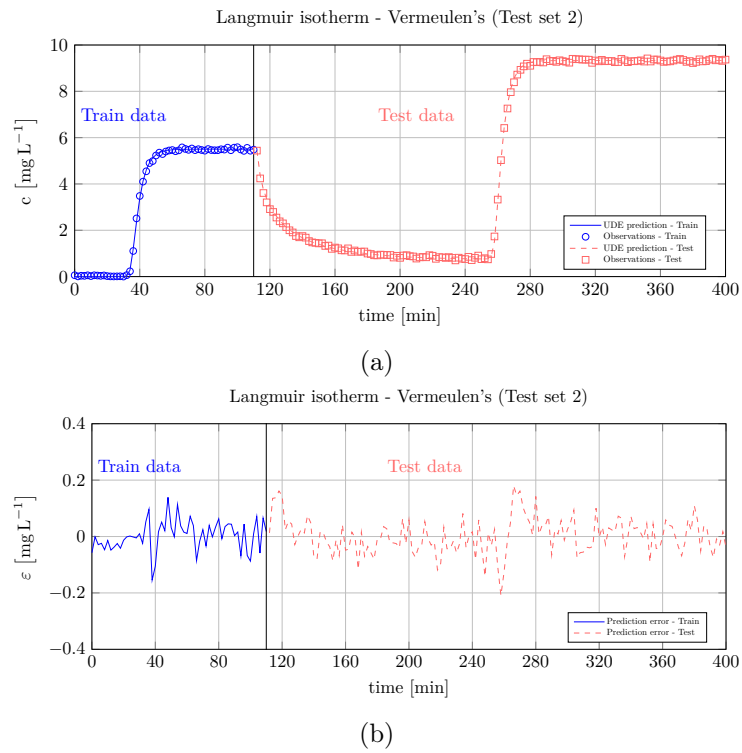


Figure A.6: Results for Sips isotherm and improved Vermeulen's kinetics - 0.5 min^{-1} sampling rate and 5.5-0.75-9.33 test set inlet concentrations (a). Breakthrough predictions and observations for training and test sets. (b) Prediction error for training and test sets.

A.3 Sparse regression - Sips isotherm

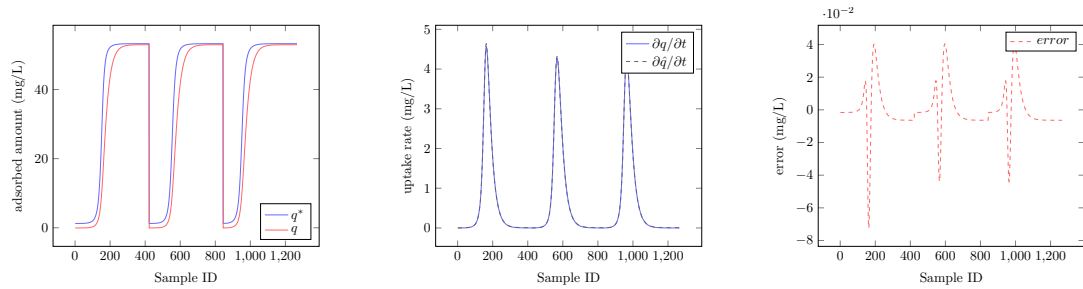


Figure A.7: Sparse regression results for the best number of active terms (3) - Sips isotherm + LDF kinetics at three positions in the bed.

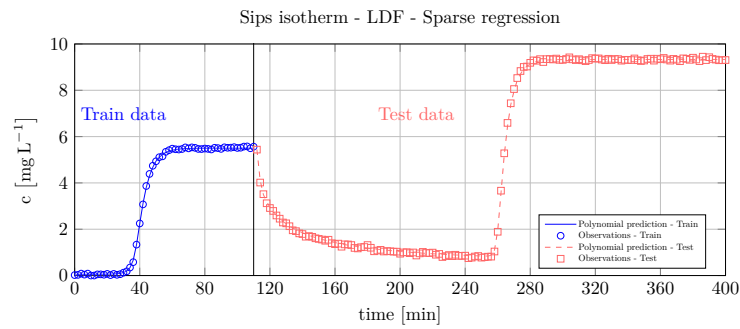


Figure A.8: Results for Sips isotherm and LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.

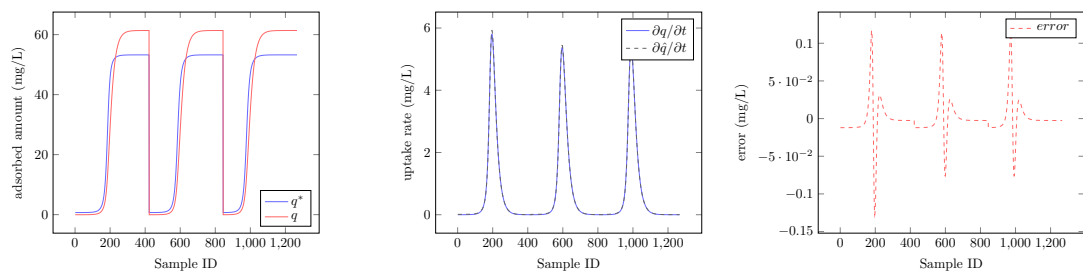


Figure A.9: Sparse regression results for the best number of active terms (4) - Sips isotherm + improved LDF kinetics at three positions in the bed.

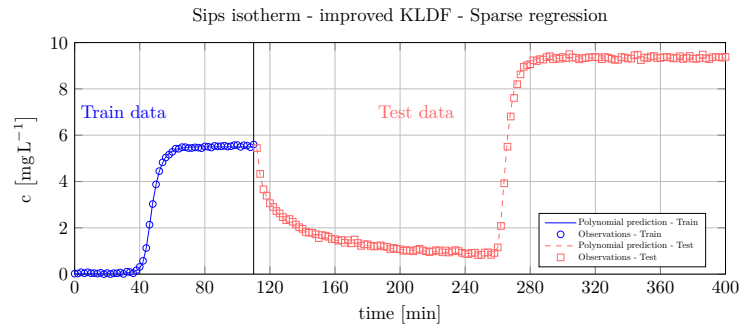


Figure A.10: Results for Sips isotherm and improved LDF kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.

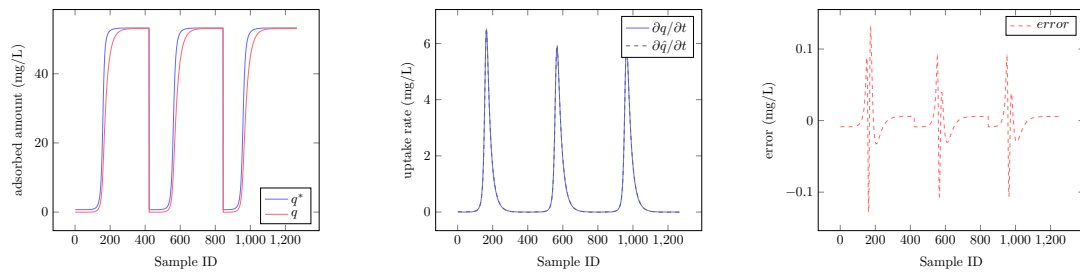


Figure A.11: Sparse regression results for the best number of active terms (6) - Sips isotherm + Vermeulen's kinetics at three positions in the bed.

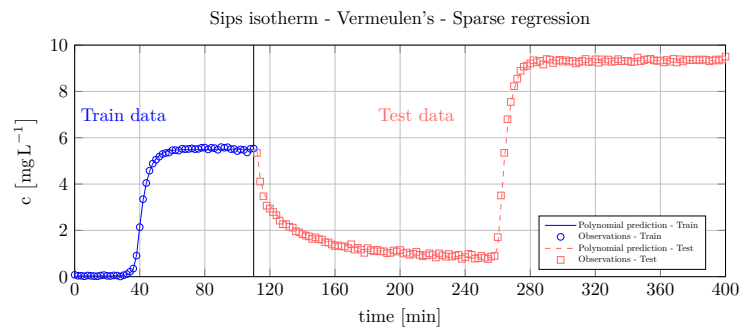


Figure A.12: Results for Sips isotherm and Vermeulen's kinetics - Breakthrough predictions and observations for training and test sets using the identified polynomial.

Appendix B

Chapter 7

B.1 Average concentration and quantile ribbons

Figures B.1 and B.2 show the mean concentration over time and the 99% credible intervals (between 5% and 99.5% quantiles) at 0.05 m and 0.5 m from the releasing source.

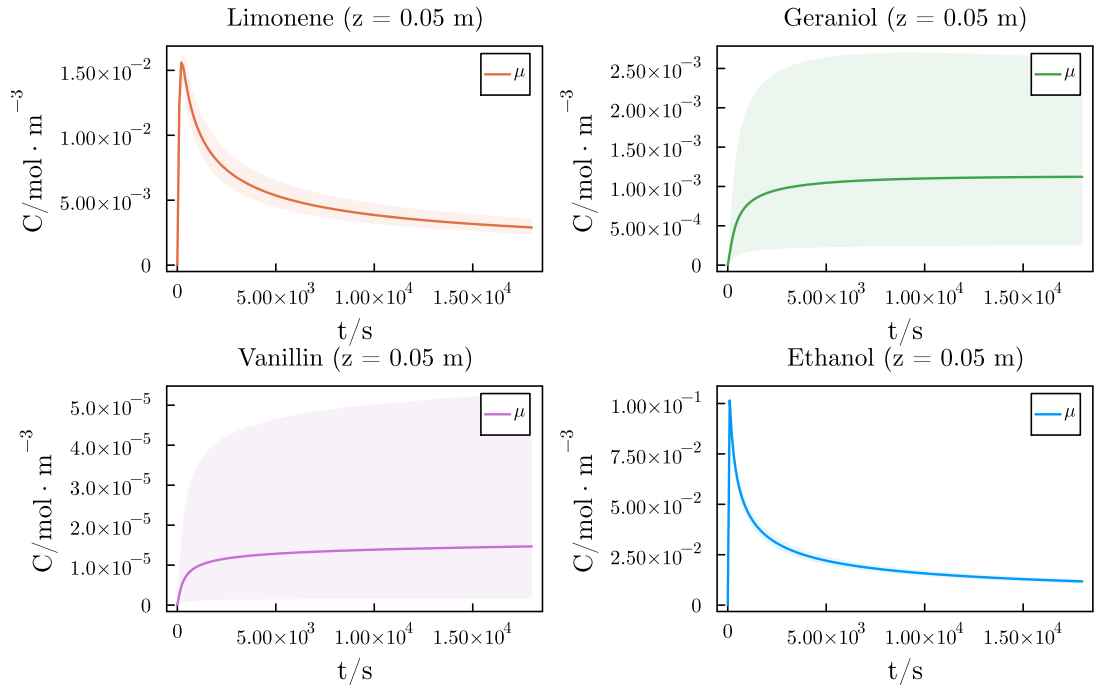


Figure B.1: Mean gas phase concentrations and 99% credible intervals for all components at $z = 0.05$ m

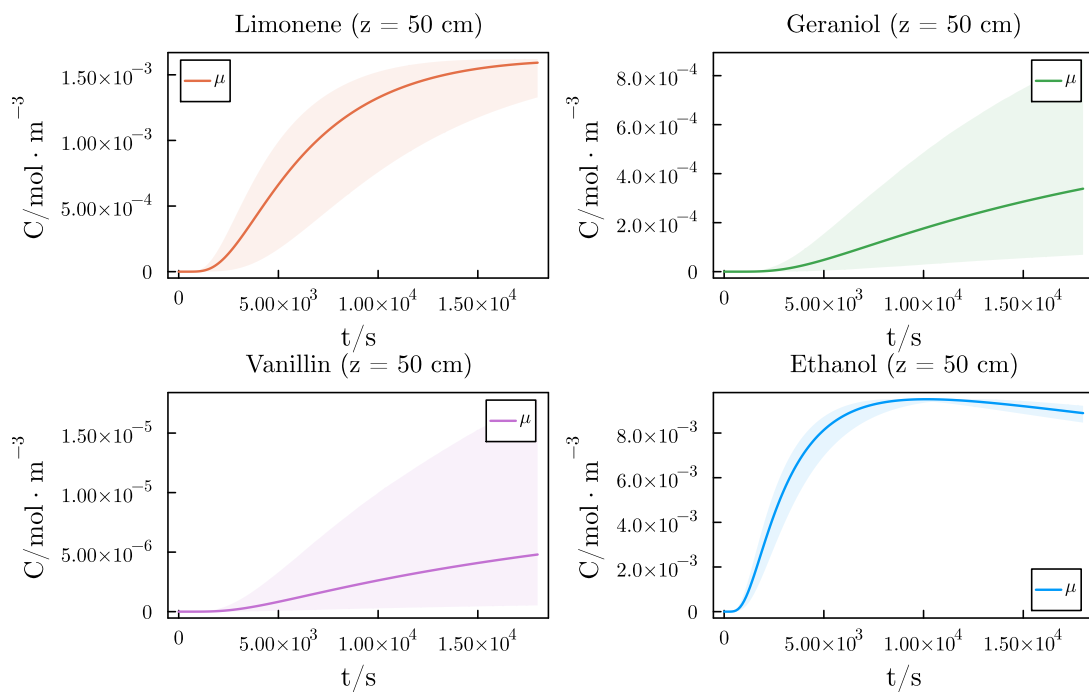


Figure B.2: Mean gas phase concentrations and 99% credible intervals for all components at $z = 50$ cm