UNIVERSITÉ DU LUXEMBOURG

PhD-FSTM-2024-083
Faculty of Science, Technology and Medicine

National Technical University of Athens
School of Chemical Engineering

DISSERTATION

Defence held on 15 November 2024 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN SCIENCES DE L'INGÉNIEUR

AND

DOCTOR OF PHILOSOPHY IN ENGINEERING

OF THE NATIONAL TECHNICAL UNIVERSITY OF ATHENS

by

Paris PAPAVASILEIOU
Born on 18 July 1997 in Athens (Greece)

HYBRID EQUATION-BASED AND DATA-DRIVEN COMPUTATIONAL WORKFLOWS FOR ANALYSIS AND PREDICTION OF INDUSTRIAL DEPOSITION PROCESSES

Dissertation defence committee

Prof. Dr Stéphane BORDAS, Supervisor
Full professor in Computational mechanics, UNIVERSITÉ DU LUXEMBOURG / Esch-sur-Alzette / Luxembourg

Prof. Dr Andreas G. BOUDOUVIS, Co-Supervisor
Professor, NATIONAL TECHNICAL UNIVERSITY OF ATHENS / Athens / Greece

Prof. Dr Alexander SKUPIN, Chair
Associate professor, UNIVERSITÉ DU LUXEMBOURG / Esch-sur-Alzette / Luxembourg

Prof. Dr Dimitrios GEROGIORGIS, Vice-Chair
Professor, UNIVERSITY OF EDINBURGH / Edinburgh / United Kingdom

Prof. Dr Georgios STEFANIDIS, Member
Professor, NATIONAL TECHNICAL UNIVERSITY OF ATHENS / Athens / Greece

Prof. Dr Luca MAGRI, Member
Professor, IMPERIAL COLLEGE LONDON / London / United Kingdom

Στον παππού μου Γιώργο, τον καλύτερο δάσκαλο που γνώρισα ποτέ.

To my grandfather George, the best teacher I have ever met.

Η έγκριση της διδακτορικής διατριβής από την Ανωτάτη Σχολή Χημικών Μηχανικών του Ε.Μ.Πολυτεχνείου δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα. (Ν. 5343/1932, Άρθρο 202)

Acknowledgements

The journey of this PhD. would not have been possible without the support of several people, and I would like to take this opportunity to thank them.

First, this endeavor would not have been possible without my supervisors, Prof. Stéphane Bordas and Prof. Andreas Boudouvis. Thank you for your support, your guidance, and your willingness to allow me to pursue the several ideas and opportunities that I came across during my PhD.

Second, I would like to express my deepest gratitude to Dr. Eleni Koronaki. Eleni, I would like to thank you wholeheartedly for allowing me to start a PhD within your project. Our discussions were a major (if not the greatest) factor in shaping this thesis and the overall trajectory of my PhD. Thank you for your unwavering support and "having my back" from day one.

In CERATIZIT, our industrial partner for the main project of this thesis, I found three excellent scientists and collaborators. Dr. Martin Kathrein, Dr. Christoph Czettl, and Dr. Gabriele Pozzetti thank you for all the fruitful exchanges and the valuable insights of the process you provided. Your contribution to this work was priceless.

Subsequently, I would like to thank the members of the thesis defense jury. Prof. Dimitrios Gerogiorgis and Prof. Luca Magri, I am honored that you agreed to review my thesis and be present in the defense. Prof. Georgios Stefanidis, thank you for our discussions and your pivotal contribution to setting off this PhD. Finally, Prof. Alexander Skupin, thank you for being part of our engaging collaboration (alongside Dr. Sofia Farina) and agreeing to chair my defense (also for the excellent Berlin travel suggestions).

I express my sincere gratitude to Dr. Dimitris Giovanis, as part of this work would not have been possible without his contributions.

To my officemates, Aravind (I still owe you that guitar solo), Mingo, and Chintan, I am grateful for all the time we spent together, our stimulating conversations, and all the inside jokes. It was a pleasure to work alongside you and get destroyed in Foosball. My experience at the University of Luxembourg would not have been the same without you. Keep delivering!

Thanks to all the members of the Legato team, my fellow travelers on this journey. Arnaud,

Saurabh, Zhaoxiang, Diego, Meryem, Surendran, Lars, Jack, and all the others, many thanks. This journey would not have been the same without you. Many thanks to Geremy and Sofia for our fruitful collaborations.

Special thanks to Dr. Giorgos Gakis and Dr. Ioannis Aviziotis from NTUA for all their help and our productive discussions in the initial stages of my PhD.

Many thanks to Ricardo, May, and the rest of the team. May our paths cross again!

Dimitri and Eddie, Vyrona and Gianni, thank you for all the memories. Hanging out with you never failed to brighten my day and make me feel close to home.

Alex and Tessa, thanks for being the best roommates I could have asked for. I am honored to call you my dear friends. As a bonus, I found two more dear friends in your partners, Angela and Ruben. In all honesty, when reminiscing about my time in Luxembourg, my mind will always travel to the times we spent together.

To all my friends in Greece (I sincerely hope that I am not forgetting someone): Mitso, Thano, Pantelo, Stavraki, Spiro, Angelo, Kots, Simo, Theodora, Niko, Ioanna, Nektaria, Vlachos brothers, Maria, Ellie and Chrysoula, Alex, Klaoudio, Thodori, and Byron, thank you for your support. Whether physical or (more often) virtual, it was vital throughout these years.

Lemonia, thank you for always being there for me. Thank you for your support and patience through all the highs and lows of this PhD, even when it was not easy for you. Thank you for believing in me even in times when I did not.

Finally, I would like to thank my family. To my father, Giorgos, and my mother, Efi, this thesis is as much yours as it is mine. Thank you for *everything*. To my aunt, Tzeni, my uncle, Dimitris, and my cousin Ioanna, thank you for your enduring support. Παππού (μεγάλε εσύ δάσκαλε), γιαγιά, τα καταφέραμε!

Financial Support

This work was financially supported by the Fonds National de la Recherche (FNR) Luxembourg (BRIDGE grant HybridSimCVD) and the Faculty of Science, Technology, and Medicine (FSTM) of the University of Luxembourg.

Abstract

Despite the leaps in quality and quantity of industrial data along with the increased interest in data-driven approaches brought about by Industry 4.0, there are still processes that are too complex to be accurately modeled via traditional first-principles methods, yet lack the necessary data for a purely data-driven approach.

Taking an industrial chemical vapor deposition (CVD) process as a key example, this work proposes a hybrid computational workflow involving equation-based (computational fluid dynamics - CFD) and machine learning (ML) methods for the modeling, investigation, and prediction of such complex processes. First, this work aims to provide a way of predicting process outcomes while also allowing the exploration of the process and obtaining insights regarding the several interplaying physical and chemical phenomena that govern it. The proposed CFD model can help with the exploration of the process and the prediction of the quality quantity of interest, which is the thickness of the deposited alumina. It can also shed light on the governing phenomena of the process. However, it comes with a high computational cost, which makes its use in everyday applications prohibitive. To overcome this, a purely data-driven predictive model which offers improved predictive and computational performance is proposed. A way of combining process data and the results of the CFD model is also proposed, via the GappyPOD method.

Subsequently, this work proposes a purely data-driven approach to identify potential critical process parameters based on a blend of supervised and unsupervised learning approaches. Following an initial clustering of the available process outcome data and the analysis of the resulting clusters, the differences between them can be matched to the differences in their respective process inputs, allowing the identification of potential key parameters. These parameters allow for deeper insight into the process and can then be used to develop data-driven models for the qualitative and quantitative prediction of the process. The versatility of this approach is then highlighted by its application to a vastly different process; the metabolism of astrocyte cells.

Υβριδικές υπολογιστικές προσεγγίσεις βασισμένες σε εξισώσεις και οδηγούμενες από δεδομένα για την ανάλυση και πρόβλεψη βιομηχανικών διεργασιών απόθεσης

Εκτενής Ελληνική περίληψη

Η παρούσα διατριβή πραγματεύεται την συνδυασμένη μοντελοποίηση διεργασιών με μεθόδους βασισμένες σε εξισώσεις και με μεθόδους βασισμένες στα δεδομένα. Ο απώτερος σκοπός είναι η πρόταση υπολογιστικών προσεγγίσεων που θα επιτρέπουν την ανάλυση και την καλύτερη κατανόηση των διεργασιών, αλλά και την έγκυρη πρόβλεψή των εκβάσεών τους εντός των πιθανών χρονικών (π.χ. ανάγκη πρόβλεψης εντός συγκεκριμένου χρονικού περιθωρίου) και υλικών περιορισμών (π.χ. χαμηλή διαθέσιμη υπολογιστική ισχύς) που υπάρχουν σε βιομηχανικές συνθήκες.

Στην μηχανική αλλά και σε άλλους τομείς της επιστήμης, οι φυσικές και χημικές διεργασίες που λαμβάνουν χώρα εντός ενός συστήματος μπορούν να είναι καθοριστικές για την έκβασή τους. Μπορούν να κάνουν την διαφορά μεταξύ προϊόντων που είναι εντός και εκτός προδιαγραφών και, σε βιολογικά συστήματα, την διαφορά μεταξύ της συντήρησης της ζωτικής λειτουργίας και της κρίσιμης αστοχίας του συστήματος. Η δυνατότητα ανάλυσης, κατανόησης και πρόβλεψης αυτών των διεργασιών είναι άκρως σημαντική. Η μαθηματική μοντελοποίηση έχει αυτές τις δυνατότητες.

Μέσω της μαθηματικής μοντελοποίησης, γίνεται προσπάθεια εύρεσης συσχετίσεων μεταξύ των «εισόδων» του συστήματος και του αποτελέσματος της διεργασίας. Για την επιτυχή μοντελοποίηση του οποιουδήποτε συστήματος, η ύπαρξη δεδομένων είναι απαραίτητη. Χωρίς δεδομένα σχετικά με την διεργασία, τις παραμέτρους της και τις ποσότητες ενδιαφέροντος (Quantities of Interest – QoIs), οποιοδήποτε προτεινόμενο μοντέλο δεν θα μπορούσε ένα επικυρωθεί και η αντιστοιχία του με την πραγματικότητα θα παρέμενε απροσδιόριστη. Η ποσότητα των δεδομένων είναι εξαιρετικά σημαντική και το πλήθος των διαθέσιμων δεδομένων είναι καθοριστικό για την επιλογή του τρόπου μοντελοποίησης.

Η «παραδοσιακή» ανάλυση βασίζεται σε εξισώσεις που διατυπώνουν θεμελιώδεις αρχές. Συνήθως πρόκειται για αρχές διατήρησης ορμής, μάζας και ενέργειας στα συνεχή μέσα όποτε οι εξισώσεις που προκύπτουν είναι μερικές διαφορικές , δηλαδή διαφορικές εξισώσεις με μερικές παραγώγους ως προς χρόνο και χώρο. Αυτή η προσέγγιση μπορεί να ονομαστεί προσέγγιση βασισμένη στις εξισώσεις, προσέγγιση βασισμένη στην φυσική, ή προσέγγιση οδηγούμενη από υποθέσεις. Όπως προδίδει και το όνομά τους, αυτές οι προσεγγίσεις ξεκινούν από κάποιες υποθέσεις για την διεργασία, βασισμένες σε ήδη υπάρχουσα γνώση (θεμελιώδεις αρχές). Οι

υποθέσεις αυτές ακολουθούνται από πειράματα, το αποτέλεσμα των οποίων τις υποστηρίζει ή τις καταρρίπτει, αναδεικνύοντας έτσι την ανάγκη για διαφορετικές υποθέσεις. Οι απαιτήσεις αυτών των μεθόδων για δεδομένα ειναι χαμηλές.

Εξίσου χρήσιμες για την ανάλυση διεργασιών αποτελούν οι προσεγγίσεις που είναι βασισμένες αμιγώς στα δεδομένα. Αυτές οι προσεγγίσεις δεν χρειάζονται κάποια προκαθορισμένη υπόθεση σχετικά με την διεργασία, καθώς δύνανται να εφαρμοστούν απευθείας και έχουν την δυνατότητα εύρεσης συσχετίσεων και μοτίβων εντός του διαθέσιμου συνόλου δεδομένων. Με βάση τα μοτίβα και τις συσχετίσεις που ανιχνεύονται μέσω της χρήσης μεθόδων αμιγώς οδηγούμενων από δεδομένα, είναι δυνατόν να σχηματιστούν νέες υποθέσεις για την διεργασία. Να σημειωθεί πως όσο πιο «πλούσιο» το σύνολο των διαθέσιμων δεδομένων, τόσο πιο εύκολη η εύρεση ουσιαστικών μοτίβων και συσχετίσεων.

Τι συμβαίνει όμως όταν τα διαθέσιμα δεδομένα δεν είναι ούτε λιγοστά ούτε άφθονα; Τότε, εμφανίζεται η ανάγκη για υβριδικές προσεγγίσεις, οι οποίες προσπαθούν να συνδυάσουν τα προτερήματα των δύο προαναφερθέντων προσεγγίσεων. Η «ενδιάμεση» αυτή ποσότητα δεδομένων χαρακτηρίζει πολλές εφαρμογές μηχανικής, οπού η έγκυρη και έγκαιρη μοντελοποίηση είναι αναγκαία, ανεξαρτήτως της πιθανώς υψηλής πολυπλοκότητας των διεργασιών. Η μοντελοποίηση τέτοιου τύπου διεργασιών, συνήθως απαιτεί αρκετές παραδοχές. Επίσης τα διαθέσιμα δεδομένα είναι «πραγματικά» και συνεπώς ενδέχεται να περιέχουν λάθη, ανακρίβειες και ελλείψεις, κάτι που ενδέχεται να δημιουργήσει προβλήματα στην μοντελοποίηση.

Μια τέτοια διεργασία είναι και η Χημική Απόθεση από Ατμό (ΧΑΑ). Στην παρούσα διατριβή, η βιομηχανική απόθεση της πολυστρωματικής επίστρωσης Ti(C,N)/α-Al_2O_3 πάνω σε τσιμεντοποιημένο καρβίδιο για την παραγωγή ανθεκτικών κοπτικών εργαλείων χρησιμοποιείται ως μελέτη περίπτωσης, καθώς αποτελεί μια πολύπλοκη διεργασία που δεν χαρακτηρίζεται ούτε από παντελή έλλειψη δεδομένων, άλλα ούτε και από αφθονία δεδομένων.

Η ΧΑΑ αποτελεί μια διεργασία κατά την οποία μια επίστρωση αποτίθεται σε ένα θερμό υπόστρωμα από πρόδρομες ενώσεις που βρίσκονται στην αέρια φάση, μέσω μιας σειράς ομογενών και ετερογενών χημικών αντιδράσεων. Στην διεργασία συνεισφέρει πληθώρα φυσικών (π.χ. διάχυση, συναγωγή, εκρόφηση) και χημικών φαινομένων και εάν σε όλα αυτά τα φαινόμενα συνυπολογισθεί και ο μεγάλος αριθμός ενδιάμεσων προϊόντων και παραπροϊόντων που δημιουργούνται κατά την διάρκεια της διεργασίας, η πολυπλοκότητά της γίνεται προφανής.

Οι προσεγγίσεις βασισμένες στις εξισώσεις, και ως επί το πλείστον η υπολογιστική ρευστοδυναμική (Computational Fluid Dynamics – CFD) έχουν χρησιμοποιηθεί με επιτυχία στο παρελθόν για την μοντελοποίησης διεργασιών ΧΑΑ. Ο λόγος πίσω από αυτή την επιλογή είναι η ικανότητα των μοντέλων υπολογιστικής ρευστοδυναμικής να διατυπώσουν τον τρόπο κατά τον οποίο τα διάφορα φυσικά και χημικά φαινόμενα που λαμβάνουν ταυτόχρονα χώρα κατά την

διάρκεια της διεργασίας επηρεάζουν το τελικό αποτέλεσμα. Παρόλα αυτά, η ανάπτυξη ενός έγκυρου μοντέλου ρευστοδυναμικής για μια διεργασία ΧΑΑ δεν αποτελεί εύκολο έργο. Πολύ συχνά, οι διάφορες παράμετροι που καθορίζουν τα φυσικά και χημικά φαινόμενα είναι άγνωστες και οι τιμές τους πρέπει να βρεθούν μέσω επίπονων διαδικασιών δοκιμής και σφάλματος. Επίσης, σε ιδιαίτερες περιπτώσεις η γεωμετρία του αντιδραστήρα δύναται να αλλάζει από μέρα σε μέρα, κάτι που δυσχεραίνει την χρήση ενός μόνο μοντέλου για όλες τις περιπτώσεις. Τέλος, η ΧΑΑ αποτελεί μια διεργασία εξαιρετικά ευαίσθητη στις συνθήκες ροής, θερμοκρασίας, και πίεσης εντός του αντιδραστήρα. Έτσι, πολύ συχνά, υπάρχει έλλειψη αισθητήρων εντός του αντιδραστήρα και συνεπώς έλλειψη καίριων δεδομένων για την ανάπτυξη μοντέλων ακριβείας της διεργασίας.

Παρόλα αυτά, ο ερχομός της τέταρτης βιομηχανικής επανάστασης (Industry 4.0) έχει βοηθήσει στην αύξηση των διαθέσιμων δεδομένων στον βιομηχανικό τομέα. Η ποσότητα των βιομηχανικών δεδομένων, αν και εμφανώς αυξημένη, σε καμία περίπτωση δεν μπορεί να συγκριθεί με την ποσότητα των δεδομένων που είναι διαθέσιμα στα μέσα κοινωνικής δικτύωσης ή στον χρηματοοικονομικό τομέα. Πέρα από την αύξηση των δεδομένων, η τέταρτη βιομηχανική επανάσταση έχει συνεισφέρει και στην γενικότερη αύξηση του ενδιαφέροντος για τα δεδομένα και για τις μεθόδους που είναι αμιγώς βασισμένες σε αυτά. Η αύξηση της ποσότητας δεδομένων αλλά και το μεγαλύτερο ενδιαφέρον για αυτά οδήγησε στην εφαρμογή μεθόδων αμιγώς βασισμένων στα δεδομένα και για την μελέτη και πρόβλεψη της προαναφερθείσας διεργασίας ΧΑΑ. Παρόλα αυτά, μιας και η κατανόηση των φυσικών και χημικών φαινόμενων που εμπλέκονται στην διεργασία είναι εξίσου σημαντική, η μοντελοποίηση μέσω μεθόδων οδηγούμενων από δεδομένα συνοδεύεται από την ανάπτυξη ενός μοντέλου υπολογιστικής ρευστοδυναμικής για την διεργασία.

Εν τέλει, η διεργασία μοιάζει πολύ περίπλοκη για να μοντελοποιηθεί επαρκώς από κάποιο μοντέλο υπολογιστικής ρευστοδυναμικής, ενώ το πλήθος των διαθέσιμων δεδομένων δεν επιτρέπει την βέλτιστη μοντελοποίηση της μέσω μεθόδων αμιγώς βασισμένων στα δεδομένα. Αυτή η δυσκολία της μοντελοποίησης, σε συνδυασμό με τις ιδιαιτερότητες που χαρακτηρίζουν τα δεδομένα που προέρχονται από τέτοιου τύπου διεργασίες οδηγεί στο κύριο ερώτημα που καλείται να απαντήσει η παρούσα διατριβή, το οποίο είναι το πώς πρέπει να προσεγγίζονται διεργασίες οι οποίες δεν διαθέτουν τα απαραίτητα δεδομένα για να μοντελοποιηθούν βέλτιστα από μεθόδους αμιγώς βασισμένες στα δεδομένα, άλλα είναι επίσης πολύ υπερβολικά περίπλοκες για να μοντελοποιηθούν αποκλειστικά από μεθόδους βασισμένες σε εξισώσεις. Η διατριβή προσπαθεί να απαντήσει σε αυτό το ερώτημα, απαντώντας σε δύο επιμέρους ερωτήματα:

1. Χρησιμοποιώντας αυτήν την διεργασία ΧΑΑ ως μελέτη περίπτωσης, ποια είναι η κα-

λύτερη υπολογιστική προσέγγιση που επιτρέπει την ακριβή πρόβλεψη της διεργασίας κάνοντας αποδοτική χρήση των (χρονικών ή υλικών) πόρων;

2. Με βάση τα διαθέσιμα δεδομένα, μπορούν να προσδιοριστούν πιθανές καίριες παράμετροι της διεργασίας;

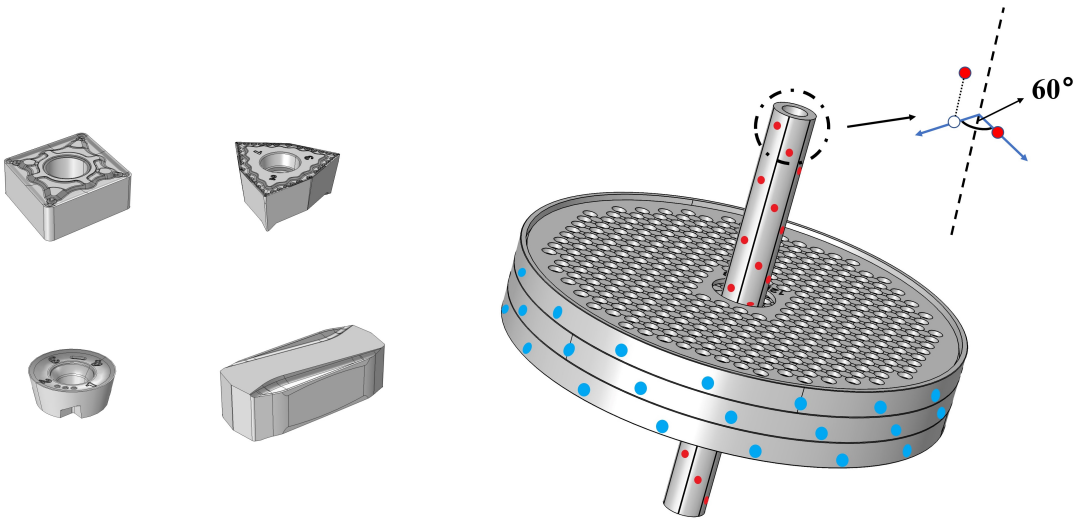
Η απάντηση του πρώτου ερωτήματος έρχεται σε δύο μέρη. Το πρώτο μέρος αφορά την απόπειρα μοντελοποίησης της διεργασίας ΧΑΑ της αλούμινας μέσω υπολογιστικής ρευστοδυναμικής, ενώ το δεύτερο αφορά την μοντελοποίηση της διεργασίας από μεθόδους αμιγώς οδηγούμενες από δεδομένα, κάτι που συνδυάζεται με την παράλληλη σύγκριση των δύο προσεγγίσεων.

Η μοντελοποίηση της βιομηχανικής κλίμακας απόθεσης της αλούμινας (α-Al_2O_3) μέσω υπολογιστικής ρευστοδυναμικής έρχεται με πολλές προκλήσεις, οι οποίες οφείλονται στις ιδιαιτερότητες του βιομηχανικού αντιδραστήρα, στις απαιτήσεις τις παραγωγής σε συνδυασμό φυσικά με τις ιδιαιτερότητες της ίδιας της διεργασίας ΧΑΑ.

Ο αντιδραστήρας που χρησιμοποιείται για την διεργασία (Sucotec SCT600TH) χαρακτηρίζεται από μια ιδιαίτερη γεωμετρία, η οποία περιλαμβάνει 40-50 διάτρητους δίσκους που τοποθετούνται ο ένας πάνω από τον άλλον, γύρω από έναν περιστρεφόμενο σωλήνα ο οποίος διανέμει τα αέρια αντιδρώντα στον αντιδραστήρα μέσω δύο αντιδιαμετρικών οπών στο επίπεδο του κάθε δίσκου. Η περιστροφή του σωλήνα προσδίδει στην διεργασία έναν περιοδικό χαρακτήρα. Με βάση τις απαιτήσεις της παραγωγής, η γεωμετρία του αντιδραστήρα μπορεί να αλλάξει ακόμα και σε καθημερινή βάση, λόγω των διαφορετικών γεωμετριών υποστρωμάτων (κοπτικών εργαλείων στην συγκεκριμένη περίπτωση) που χρειάζεται να επικαλυφθούν, αλλά και επειδή η γεωμετρία και ο τύπος των διάτρητων δίσκων που χρησιμοποιούνται εξαρτώνται από την γεωμετρία των υποστρωμάτων που επικαλύπτονται.

Πέρα από τις γεωμετρικές και παραγωγικές ιδιαιτερότητες, υπάρχει και το πρόβλημα του χημικού δικτύου που χρησιμοποιείται για την απόθεση της αλούμινας. Στην παρούσα εφαρμογή, η αλούμινα αποτίθεται από ένα μείγμα αντιδρώντων $AlCl_3$, CO_2, HCl, H_2, H_2S. Η πρόδρομη ένωση $AlCl_3$ αντιδρά με νερό που παράγεται εντός του αντιδραστήρα μέσω της αντίστροφης water gas shift αντίδρασης από διοξείδιο του άνθρακα και υδρογόνο. Διάφορα συστήματα αντιδράσεων έχουν προταθεί στην βιβλιογραφία, τα οποία συνδυάζουν έως και 104 αντιδράσεις μεταξύ 35 ενδιάμεσων προϊόντων. Η χρήση ενός τόσο λεπτομερούς συστήματος αντιδράσεων σε ένα μοντέλο ρευστοδυναμικής θα οδηγούσε σε μεγάλα υπολογιστικά κόστη τόσο κατά την ανάπτυξή του (επαναπροσδιορισμός των παραμέτρων των αντιδράσεων), όσο και κατά την χρήση του (αυξημένα υπολογιστικά κόστη λόγω αύξησης των αγνώστων).

Με βάση τα παραπάνω, μπορεί να γίνει αντιληπτό πώς ένα πολύ λεπτομερές μοντέλο υπολογιστικής ρευστοδυναμικής για την συγκεκριμένη διεργασία θα περιελάμβανε μια τρισδι-



(α′) (β′)

Σχήμα 1: (α′) Ενδεικτικές γεωμετρίες των προς επίστρωση κοπτικών εργαλείων. (β′) Τρισδιάστατη αναπαράσταση ενός τμήματος του αντιδραστήρα με 3 δίσκους. Τα υποστρώματα τοποθετούνται σε καθέναν από αυτούς τους δίσκους. Με κόκκινο χρώμα: οπές στον περιστρεφόμενο σωλήνα διανομής των αέριων αντιδρώντων. Με μπλε χρώμα: διατρήσεις εξόδου για κάθε δίσκο. Οι οπές και η περιστροφή του σωλήνα εισόδου επιτρέπουν την ομοιόμορφη ροή των αντιδρώντων αερίων στο εσωτερικό του αντιδραστήρα. Η απόθεση μπορεί να λάβει χώρα σε όλες τις επιφάνειες εντός του αντιδραστήρα (δίσκοι, ένθετα, τοιχώματα κ.λπ.).

άστατη γεωμετρία και όλους τους 40-50 δίσκους της διεργασίας. Επίσης, θα ήταν εξαρτώμενο από τον χρόνο. Επίσης, θα χρησιμοποιούσε ένα λεπτομερές και περίπλοκο σύστημα αντιδράσεων. Όλα τα παραπάνω θα οδηγούσαν σε ένα εξαιρετικά ακριβό μοντέλο, καθιστώντας την χρήση του σε καθημερινή βάση απαγορευτική.

Με σκοπό την υπέρβαση των παραπάνω προβλημάτων που σχετίζονται με το υπολογιστικό κόστος, προτείνεται ένα δισδιάστατο μοντέλο ρευστοδυναμικής (βλ. Κεφάλαιο 3) το οποίο λαμβάνει υπόψιν κομμάτια του αντιδραστήρα που αποτελούνται από επτά δίσκους. Επίσης, το μοντέλο περιλαμβάνει ένα απλουστευμένο σύστημα χημικών αντιδράσεων, κάτι που σε συνδυασμό με τις γεωμετρικές απλουστεύσεις μειώνει περαιτέρω το απαιτούμενο υπολογιστικό κόστος. Τέλος, για την προσομοίωση της περιστροφής του σωλήνα παροχής των αντιδρώντων σε μια δισδιάστατη γεωμετρία, επιλέγονται κατάλληλες περιοδικές οριακές συνθήκες για την παροχή.

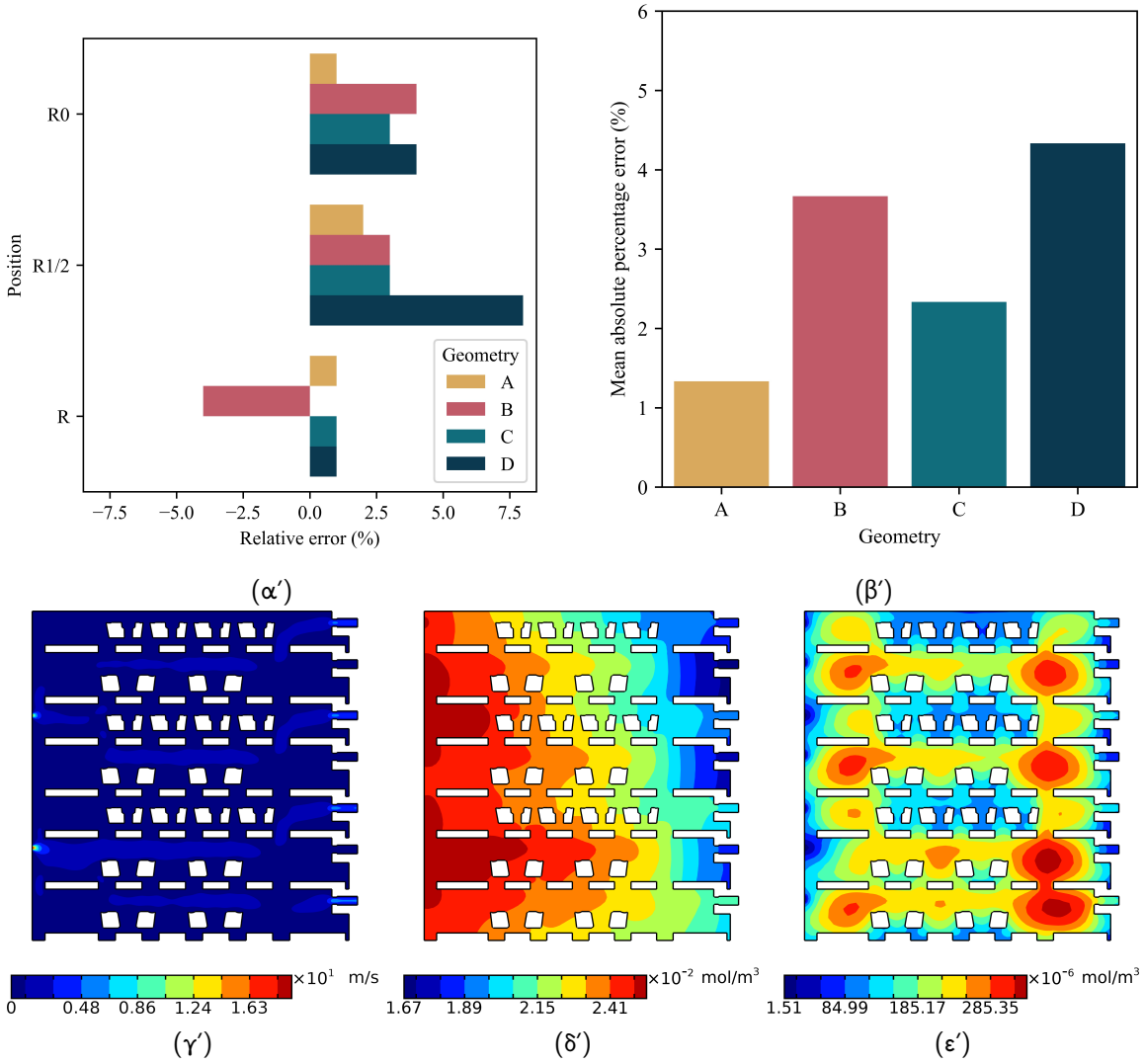
Το προτεινόμενο μοντέλο αρχικά προσαρμόζεται στο πρόβλημα μέσω της προσαρμογής των κινητικών παραμέτρων των προτεινόμενων χημικών αντιδράσεων, και στην συνέχεια επικυρώνεται χρησιμοποιώντας τα διαθέσιμα δεδομένα παραγωγής, τα οποία περιλαμβάνουν

μετρήσεις του πάχους της αλούμινας σε 15 διαφορετικά σημεία του αντιδραστήρα για κάθε σειρά παραγωγής. Να σημειωθεί πως το πάχος αποτελεί εξαιρετική ένδειξη της ποιότητας της παραγωγής, μιας και ο στόχος είναι το ομοιόμορφο πάχος απόθεσης για τα υποστρώματα όλου του αντιδραστήρα. Παρά τον μεγάλο αριθμό παραδοχών και απλουστεύσεων, το μοντέλο είναι ικανό να δώσει προβλέψεις του πάχους της αλούμινας, με ένα μέσο απόλυτο ποσοστιαίο σφάλμα της τάξεως του 4%. Πέρα από τις προβλέψεις του πάχους της επίστρωσης της αλούμινας, το μοντέλο παρέχει πληροφορίες για το προφίλ της ροής εντός του αντιδραστήρα, αλλά και για τις επιμέρους συγκεντρώσεις των διαφόρων αντιδρώντων εντός του αντιδραστήρα. Μία προσομοίωση για 60 δευτερόλεπτα της διεργασίας (2 περίοδοι – λόγω της περιστροφής του σωλήνα εισόδου) για ένα κομμάτι 7 δίσκων, απαιτεί περίπου 3 ώρες σε έναν τυπικό επεξεργαστή. Επιπροσθέτως, το μοντέλο επιτρέπει και υπολογιστικά πειράματα, για αναλογίες πρόδρομων ενώσεων για τις οποίες δεν υπάρχουν διαθέσιμα δεδομένα. Με βάση αυτά τα υπολογιστικά πειράματα, λήφθηκαν ενδείξεις για το κινητικό καθεστώς το οποίο καθορίζει την ταχύτητα της απόθεσης, το οποίο φαίνεται να είναι ελεγχόμενο από τις χημικές αντιδράσεις και όχι από φαινόμενα διάχυσης.

Παρόλη την επιτυχή μοντελοποίηση της διεργασίας μέσω υπολογιστικής ρευστοδυναμικής, μιας προσέγγισης βασισμένης σε εξισώσεις, το απαιτούμενο υπολογιστικό κόστος για την προσομοίωση ενός ολόκληρου αντιδραστήρα (περίπου 20 ώρες) καθιστά την χρήση του μοντέλου αδύνατη για καθημερινή χρήση (π.χ. βελτιστοποίηση της αλληλουχίας των δίσκων για την παραγωγή της ημέρας). Για αυτό το λόγο, εξετάζεται η προοπτική ενός μοντέλου αμιγώς οδηγούμενου από τα δεδομένα (βλ. Κεφάλαιο 4).

Τα διαθέσιμα δεδομένα παραγωγής, τα οποία περιέχουν πληροφορίες σχετικά με την διάταξη και τις συνθήκες λειτουργίας του αντιδραστήρα, αλλά και σχετικά με το αποτέλεσμα της διεργασίας (δηλαδή το πάχος της επίστρωσης της αλούμινας), μπορούν να αξιοποιηθούν για την αμιγώς οδηγούμενη από τα δεδομένα μοντελοποίηση της διεργασίας. Αυτό δύναται να γίνει μέσω της χρήσης αλγορίθμων μηχανικής μάθησης (Machine Learning - ML) . Οι αλγόριθμοι μηχανικής μάθησης μπορούν να «μαθαίνουν» από τα διαθέσιμα δεδομένα αλλά και να ανακαλύπτουν μοτίβα. Στην περίπτωση της πρόβλεψης του πάχους της αλούμινας σε κάποια συγκεκριμένη σειρά παραγωγής, οι προαναφερθείσες λεπτομέρειες διάταξης και λειτουργίας του αντιδραστήρα μπορούν να χρησιμοποιηθούν σαν μεταβλητές «εισόδου» σε ένα μοντέλο μηχανικής μάθησης το οποίο έχει μεταβλητή εξόδου το πάχος της αλούμινας σε κάποιον δίσκο του αντιδραστήρα. Μιας και οι μεταβλητές εξόδου είναι γνωστές, το πρόβλημα αυτό θα λυθεί με αλγορίθμους που ανήκουν στην κατηγορία της επιτηρούμενης μάθησης.

Φυσικά, τα διαθέσιμα δεδομένα παραγωγής είναι δεδομένα από τον πραγματικό κόσμο. Συνεπώς, περιέχουν σφάλματα αλλά και παραλήψεις. Επίσης, είναι δεδομένα που περιέχουν

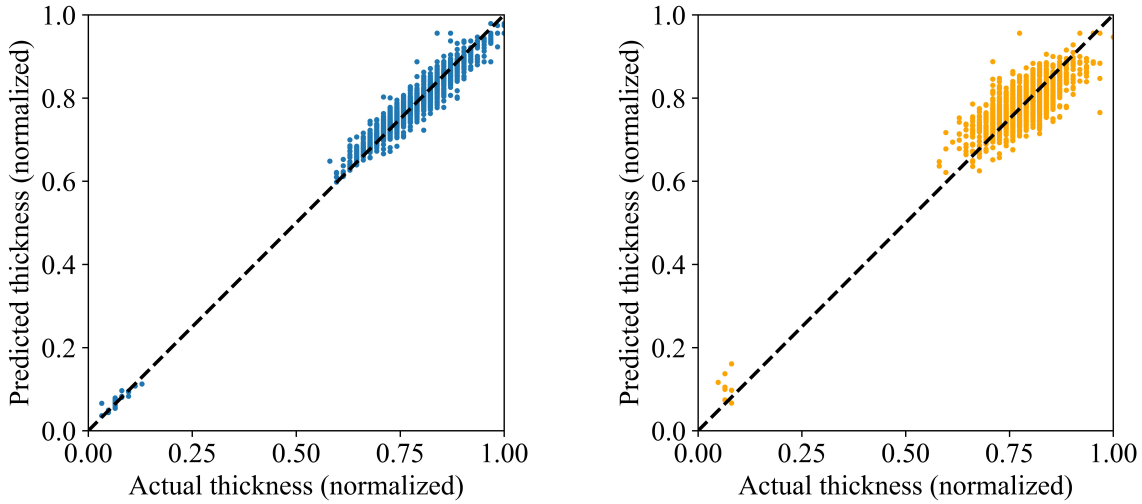


Σχήμα 2: (α′) Σχετικό σφάλμα για τις προβλέψεις του προτεινόμενου μοντέλου ρευστοδυναμικής για 3 διαφορετικές θέσεις με διαθέσιμα δεδομένα πάχους αλούμινας. Οι προσομοιώσεις πραγματοποιούνται για τέσσερις διαφορετικές γεωμετρίες 7 δίσκων συνολικά. (β′) Μέσο απόλυτο ποσοστιαίο σφάλμα (μέσος όρος για τις 3 θέσεις για τις οποίες υπάρχουν διαθέσιμα δεδομένα) για τις προσομοιώσεις του προτεινόμενου μοντέλου υπολογιστικής ρευστομηχανικής για τις 4 διαφορετικές γεωμετρίες του αντιδραστήρα. (γ′) Μέτρο ταχύτητας, (δ′) συγκέντρωση της πρόδρομης ένωσης και (ε′) συγκέντρωση νερού στο εσωτερικό του αντιδραστήρα σε μια συγκεκριμένη χρονική στιγμή κατά τη διάρκεια της απόθεσης.

πληροφορία με διαφορετικές μορφές (π.χ. ακέραιοι αριθμοί, πραγματικοί αριθμοί αλλά και κείμενο). Όλη αυτή η πληροφορία πρέπει να μετασχηματιστεί σε μορφή εύκολα διαχειρίσιμη από

τους αλγορίθμους μηχανικής μάθησης. Επίσης, είναι πολύ σύνηθες να μην χρειάζονται όλες οι διαθέσιμες μεταβλητές εισόδου για την επίτευξη έγκυρων προβλέψεων. Πολλές φορές, η χρήση «μη κατατοπιστικών» μεταβλητών εισόδου ενδέχεται να αλλοιώσει την προγνωστική απόδοση του μοντέλου. Επιπροσθέτως, το μοντέλο που θα χρησιμοποιηθεί και η πολυπλοκότητά του παίζουν επίσης σημαντικό ρόλο στην εγκυρότητα των προβλέψεων.

Έτσι, μετά το «καθάρισμα» των δεδομένων και την επιλογή κατάλληλων μετασχηματισμών για της μεταβλητές εισόδου (αφαίρεση του μέσου όρου και διαίρεση με την τυπική απόκλιση για της αριθμητικές μεταβλητές και δυαδική κωδικοποίηση για τις κατηγορικές μεταβλητές), ακολουθεί η επιλογή των μεταβλητών εισόδου που θα χρησιμοποιηθούν. Στην παρούσα περίπτωση, καταλήξαμε σε 13 μεταβλητές σχετικές με την γεωμετρία και την διάταξη του αντιδραστήρα. Καθώς εστιάζουμε στην πρόβλεψη του πάχους της αλούμινας ανά δίσκο, είναι σημαντικό να συμπεριληφθούν πληροφορίες για τα χαρακτηριστικά των γειτονικών δίσκων. Ένα μέρος των επιλεγέντων μεταβλητών δημιουργήθηκε βασισμένο στις αρχικά διαθέσιμες μεταβλητές και φαίνεται να προσδίδει στην βελτίωση της προβλεπτικής ικανότητας του μοντέλου. Ενδεικτικά, 2 από τις μεταβλητές που δημιουργήθηκαν, ήταν η συνολική επιφάνεια των υποστρωμάτων εντός του αντιδραστήρα και η τυπική απόκλιση της ανά δίσκο επιφάνειας απόθεσης εντός του αντιδραστήρα. Τέλος, απομένει η επιλογή μοντέλου μηχανικής μάθησης. Μετά από προκαταρκτικές δοκιμές, φάνηκε πως το μοντέλο που δύναται να περιγράψει καλύτερα τα διαθέσιμα δεδομένα είναι ένα μοντέλο βασισμένο σε συστοιχίες δέντρων απόφασης και πιο συγκεκριμένα ένα XGBoost μοντέλο. Η πολυπλοκότητα του μοντέλου μπορεί να επηρεάσει την προβλεπτική του ικανότητα, ειδικά αν αυτό είναι περισσότερο ή λιγότερο περίπλοκο από όσο πρέπει. Εδώ πρέπει να σημειωθεί ότι η προβλεπτική ικανότητα του μοντέλου υπολογίζεται πάντοτε για προβλέψεις του μοντέλου για ένα σύνολο δεδομένων το οποίο δεν έχει χρησιμοποιηθεί για την «προπόνηση» του. Η πολυπλοκότητα καθορίζεται από τις διάφορες υπερπαραμέτρους του μοντέλου (π.χ. το μέγιστο βάθος των δέντρων, ο συνολικός αριθμός δέντρων στην συστοιχία). Έτσι, οι αρχικές δοκιμές οφείλουν να ακολουθηθούν από μια διεργασία βελτιστοποίησης αυτών των υπερπαραμέτρων. Εν τέλει, επιλέγεται ένα μοντέλο παλινδρόμησης XGBoost , ικανό να προβλέψει το πάχος της επίστρωσης της αλούμινας ανά δίσκο με μέσο απόλυτο ποσοστιαίο σφάλμα της τάξεως του 3%, δείχνοντας βελτίωση σε σύγκριση με το μοντέλο υπολογιστικής ρευστοδυναμικής. Πέρα από την βελτίωση στην προβλεπτική ικανότητα, το μοντέλο μπορεί να κάνει προβλέψεις για έναν ολόκληρο αντιδραστήρα σε περίπου ένα δευτερόλεπτο, πετυχαίνοντας έτσι μια τεράστια βελτίωση στον απαιτούμενο χρόνο πρόβλεψης της τάξεως του 99.99% συγκριτικά με το μοντέλο υπολογιστικής ρευστοδυναμικής. Το μοντέλο αυτό όμως, σε αντίθεση με το μοντέλο υπολογιστικής ρευστοδυναμικής, δεν δίνει καμία πληροφορία σχετικά με την ροή και τις συγκεντρώσεις των αντιδρώντων εντός του αντιδραστήρα.



(α′) (β′)

Σχήμα 3: Απόδοση του προτεινόμενου μοντέλου μηχανικής μάθησης (α′) στο σύνολο δεδομένων εκπαίδευσης: Μέσο τετραγωνικό σφάλμα: 0.005 ¸ Μέσο απόλυτο σφάλμα:0.051 ¸ Μέσο απόλυτο ποσοστιαίο σφάλμα: 0.9% ¸ R^2: 0.980. (β′) απόδοση στο σύνολο δεδομένων δοκιμής: Μέσο τετραγωνικό σφάλμα: 0.059 ¸ Μέσο απόλυτο σφάλμα: 0.187 ¸ Μέσο απόλυτο ποσοστιαίο σφάλμα: 3.1% ¸ R^2: 0.753.

Στην συνέχεια, σε μια προσπάθεια να συνδυαστούν τα διαθέσιμα δεδομένα παραγωγής και τα δεδομένα που παρέχονται από το μοντέλο υπολογιστικής ρευστοδυναμικής, προτείνεται μια προσέγγιση μέσω της μεθόδου GappyPOD, η οποία επιτρέπει την ακριβή ανακατασκευή των αποτελεσμάτων του μοντέλου υπολογιστικής ρευστοδυναμικής (τα προφίλ ροής και τα προφίλ συγκεντρώσεων) με την χρήση περιορισμένου αριθμού δεδομένων τόσο από το μοντέλο ρευστοδυναμικής, όσο και από τα διαθέσιμα δεδομένα παραγωγής.

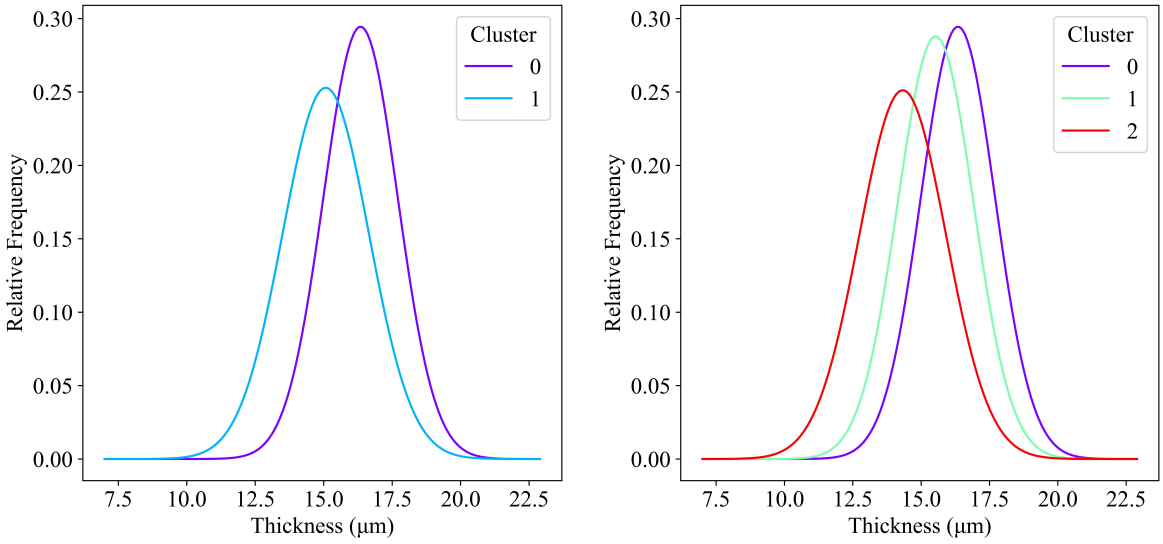
Με την ανάπτυξη αυτών των δύο προσεγγίσεων για την μοντελοποίηση της διεργασίας, έχουν δοθεί δυο διαφορετικοί τρόποι πρόβλεψης για το αποτέλεσμα της διεργασίας, ο καθένας με τα δικά του προτερήματα και μειονεκτήματα. Διαθέτοντας και τους δύο όμως, υπάρχει η δυνατότητα και για γρήγορες προβλέψεις, αλλά και για βαθύτερη ανάλυση της διεργασίας. Έτσι, με τα δύο προτεινόμενα μοντέλα, θεωρούμε ότι η ερώτηση 1 που τέθηκε παραπάνω έχει απαντηθεί.

Στο υπόλοιπο μέρος της διατριβής, ο στόχος είναι η απάντηση της δεύτερης ερώτησης. Συγκεκριμένα, χρησιμοποιώντας μεθόδους μηχανικής μάθησης, γίνεται προσπάθεια προσδιορισμού μεταβλητών που πιθανώς να είναι καίριες για το αποτέλεσμα της διεργασίας (βλ.

Κεφάλαιο 5). Η προσπάθεια αυτή ξεκινάει, με την χρήση των διαθέσιμων δεδομένων για κάθε σειρά παραγωγής. Οι 15 διαθέσιμες μετρήσεις πάχους αλούμινας για τις 603 διαθέσιμες σειρές παραγωγής περνάνε από μια διαδικασία ομαδοποίησης (clustering). Η ομαδοποίηση, είναι μια μέθοδος μηχανικής μάθησης που ανήκει στις μεθόδους μη επιτηρούμενης μάθησης, μιας και δεν υπάρχει μεταβλητή εξόδου. Σκοπός της ομαδοποίησης είναι να ανακαλύψει ομάδες μεταξύ των διαθέσιμων δεδομένων. Τα μέλη των ομάδων που δημιουργούνται μοιάζουν πιο πολύ με τα μέλη της ομάδας τους παρά με τα μέλη άλλων ομάδων. Η ομοιότητα αυτή καθορίζεται με βάση κάποια κριτήρια, ένα παράδειγμα των οποίων αποτελεί η Ευκλείδεια απόσταση. Στην παρούσα δουλειά, χρησιμοποιείται αλγόριθμος συσσωρευτικής ιεραρχικής ομαδοποίησης (agglomerative hierarchical clustering), βασισμένος στην ευκλείδεια απόσταση. Στον συγκεκριμένο αλγόριθμο, ο χρήστης καλείται να επιλέξει τον αριθμό των ομάδων.

Κατόπιν της ομαδοποίησης των δεδομένων από τις 603 σειρές παραγωγής, η οποία είναι βασισμένη μόνο στα δεδομένα του πάχους απόθεσης, μπορούμε να κάνουμε κάποιες παρατηρήσεις, με βάση την ποιοτική πληροφορία που περιέχεται στο πάχος απόθεσης. Αφού υπολογισθεί ο μέσος όρος και η τυπική απόκλιση του πάχους απόθεσης για την κάθε ομάδα που προέκυψε, φαίνεται πως οι ομάδες που προκύπτουν κατέχουν διαφορετικά ποιοτικά χαρακτηριστικά. Εάν ζητηθεί από τον αλγόριθμο να βρεθούν 2 ομάδες, προκύπτει πως αυτές έχουν σημαντικές διαφορές, τόσο στον μέσο όρο (ομάδα 0: 16.35 μικρόμετρα, ομάδα 1: 15.08 μικρόμετρα), όσο και στην τυπική απόκλιση (ομάδα 0: 1.355 μικρόμετρα, ομάδα 1: 1.578 μικρόμετρα) του πάχους απόθεσης. Αντίστοιχα, στην περίπτωση των τριών ομάδων, η ομάδα 1 από την πρώτη περίπτωση φαίνεται να «σπάει» σε 2 επιμέρους. Οι τρείς ομάδες που προκύπτουν επίσης παρουσιάζουν σημαντικές διαφορές στον μέσο όρο (ομάδα 0: 16.35 μικρόμετρα, ομάδα 1: 15.53 μικρόμετρα, ομάδα 2: 14.32 μικρόμετρα) και την τυπική απόκλιση όρο (ομάδα 0: 1.354 μικρόμετρα, ομάδα 1: 1.386 μικρόμετρα, ομάδα 2: 1.588 μικρόμετρα) του πάχους απόθεσης. Στις ΧΑΑ, είναι επιθυμητό να υπάρχει όσο το δυνατόν μεγαλύτερο πάχος απόθεσης, με την προϋπόθεση να διατηρείται η ομοιομορφία της επίστρωσης μεταξύ όλων των υποστρωμάτων. Χάρη σε αυτό το ποιοτικό χαρακτηριστικό του πάχους, είναι εμφανές ότι οι ομάδες που προκύπτουν από την ομαδοποίηση έχουν ξεκάθαρες ποιοτικές διαφορές, με την ομάδα 0 να αποτελεί την «καλύτερη» ομάδα και την ομάδα 2 να αποτελεί την «χειρότερη».

Εάν στην συνέχεια εστιάσουμε στις μεταβλητές εισόδου της διεργασίας για αυτές τις ομάδες, μπορούμε να παρατηρήσουμε έντονες διαφορές. Η ομάδα 0, αποτελείται σχεδόν αποκλειστικά από σειρές παραγωγής που χρησιμοποιούν μια νεότερη εκδοχή της «συνταγής» παραγωγής. Στην «συνταγή» κωδικοποιούνται διάφορα βήματα που πρέπει να ληφθούν κατά την διάρκεια της παραγωγής, τα οποία σχετίζονται με τις συνθήκες εντός του αντιδραστήρα και τις αναλογίες των αντιδρώντων. Οι ομάδες 1 και 2, αποτελούνται από σειρές παραγωγής



Σχήμα 4: Κατανομή πάχους στην περίπτωση: (α′) 2 ομάδων και (β′) 3 ομάδων. Το υψηλό μέσο πάχος και η χαμηλή τυπική απόκλιση αποτελούν μέτρο της αποτελεσματικότητας της διαδικασίας και της ποιότητας του προϊόντος. Οι σειρές παραγωγής της «μώβ» ομάδας επιδεικνύουν ανώτερα ποιοτικά χαρακτηριστικά.

που χρησιμοποιούν παλιότερες εκδοχές της «συνταγής».

Έπειτα, αναζητώντας το τί διαφέρει στις μεταβλητές εισόδου μεταξύ των ομάδων 1 και 2, παρατηρούμε πώς υπάρχουν διαφορές σε μια μεταβλητή που αναδεικνύει το πόσο σωστά έγινε η διάταξη του αντιδραστήρα με βάση την ονομαστική επιφάνεια απόθεσης (η οποία επιλέγεται από τον χειριστή του αντιδραστήρα) και την πραγματική επιφάνεια απόθεσης εντός του αντιδραστήρα. Οι τιμές αυτών των δύο επιφανειών δεν συμπίπτουν πάντα, καθώς οι επιλογές που έχουν οι χειριστές για την ονομαστική επιφάνεια, δίνονται σε βήματα του ενός τετραγωνικού μέτρου. Στην πράξη, ο χειριστής στρογγυλοποιεί την πραγματική επιφάνεια προς τα πάνω κατά την επιλογή της ονομαστικής. Οι ομάδες 1 και 2 λοιπόν, παρουσιάζουν σημαντική διαφορά σε μια μεταβλητή που αντικατοπτρίζει αυτήν την διαφορά μεταξύ ονομαστικής και πραγματικής επιφάνειας. Αυτή η μεταβλητή είναι η απόλυτη τιμή της διαφοράς. Η μεταβλητή αυτή, μαζί με την «συνταγή» διαφέρουν μεταξύ των 3 ομάδων και ενδέχεται να είναι κρίσιμες για το αποτέλεσμα της διεργασίας.

Στην συνέχεια, χρησιμοποιώντας τις παραπάνω προσδιορισμένες μεταβλητές, μπορούν να αναπτυχθούν μοντέλα επιτηρούμενης μάθησης. Εάν σαν μεταβλητή εξόδου για αυτά τα μοντέλα χρησιμοποιηθεί η ομάδα στην οποία θα ανήκει μια παρτίδα, τότε το πρόβλημα

αποτελεί ένα πρόβλημα ταξινόμησης. Εάν σαν μεταβλητή εξόδου χρησιμοποιηθεί το μέσο πάχος της απόθεσης, τότε έχουμε μπροστά μας ένα πρόβλημα παλινδρόμησης.

Για το πρόβλημα ταξινόμησης, χρησιμοποιείται ένας αλγόριθμος τυχαίων δασών. Με την χρήση των μεταβλητών που προσδιορίστηκαν παραπάνω, σε συνδυασμό με κάποιες σχετικές με τον αύξοντα αριθμό του αντιδραστήρα και την χρονιά παραγωγής αλλά και την τυπική απόκλισης της επιφάνειας εντός του αντιδραστήρα, μπορούμε να προβλέψουμε την ομάδα στην οποία θα ανήκει μια παρτίδα με ικανοποιητική ακρίβεια. Η ακρίβεια για την περίπτωση των δύο ομάδων είναι 96.7%, ενώ για την περίπτωση των τριών ομάδων είναι 79.3%.

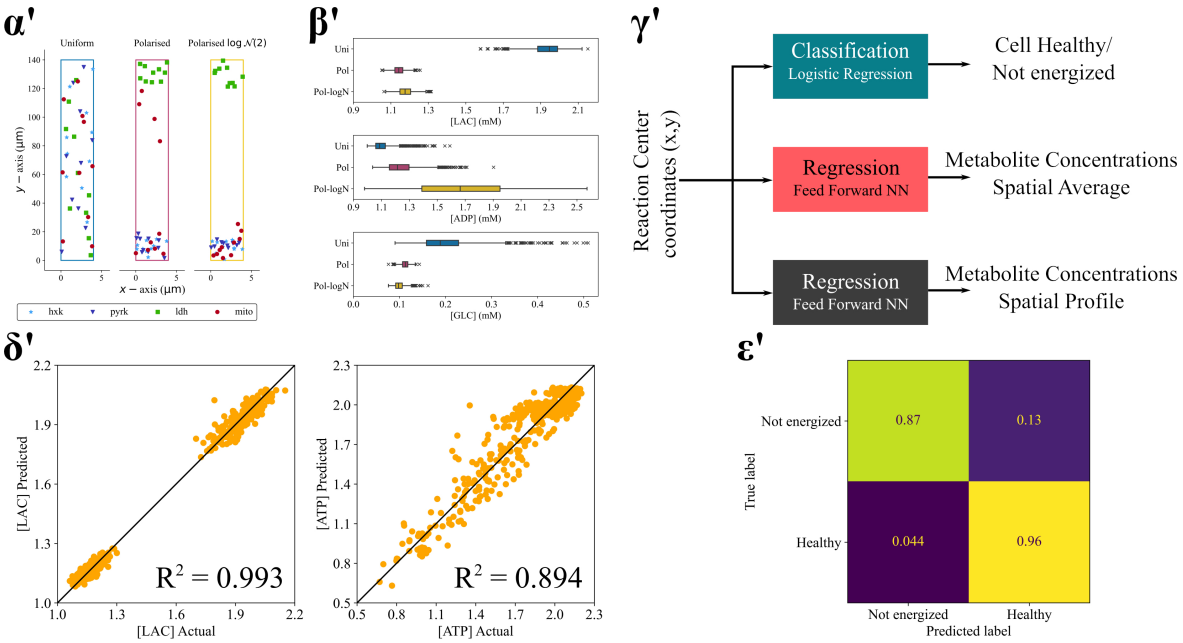
Για το πρόβλημα παλινδρόμησης, χρησιμοποιείται ένας αλγόριθμος XGBoost. Με την χρήση των μεταβλητών που χρησιμοποιήθηκαν στο πρόβλημα της ταξινόμησης, μαζί με 5 από τις 15 διαθέσιμα πάχη επίστρωσης, καταφέρνουμε να προβλέψουμε το μέσο πάχος επίστρωσης με ικανοποιητική ακρίβεια, ελαττώνοντας έτσι τις απαραίτητες μετρήσεις πάχους κατά 66.7%, και δίνοντας την δυνατότητα βελτίωσης στα βήματα που λαμβάνονται μετά το πέρας της διεργασίας για ποιοτικό έλεγχο.

Τέλος, χρησιμοποιώντας τις τιμές SHAP, οι οποίες έχουν την βάση τους στην θεωρία παιγνίων, μπορούμε να αποδώσουμε σε κάθε μεταβλητή εισόδου μια τιμή, η οποία δηλώνει τη μέση συμβολή της μεταβλητής στην πρόβλεψη, βοηθώντας έτσι στην κατανόηση του τρόπου με τον οποίο οι αλλαγές σε μια μεταβλητή μπορεί να επηρεάσουν την τελική απόδοση του μοντέλου.

Με αυτόν τον τρόπο, αναδεικνύεται πώς μια αλληλουχία μη επιτηρούμενων και επιτηρούμενων μεθόδων μηχανικής μάθησης μπορεί να επισημάνει πιθανές κρίσιμες μεταβλητές της διεργασίας. Η προσέγγιση αυτή εφαρμόζεται στην συνέχεια και σε ένα εντελώς διαφορετικό πρόβλημα· το πρόβλημα του μεταβολισμού στα αστροκύτταρα.

Εκτός του κεντρικού θέματος της διεργασίας ΧΑΑ, στην διατριβή περιλαμβάνεται η μελέτη της διεργασίας του μεταβολισμού στα αστροκύτταρα (βλ. Κεφάλαιο 6), με σκοπό την ανάδειξη σημαντικών παραμέτρων μέσω μεθοδολογίας αντίστοιχης αυτής που εφαρμόστηκε για την διεργασία ΧΑΑ. Στην περίπτωση των αστροκυττάρων, γίνεται χρήση ενός συνόλου δεδομένων που προκύπτει από υπολογιστικά πειράματα. Οι μεταβλητές εισόδου είναι οι θέσεις των κέντρων όπου λαμβάνουν χώρα χημικές αντιδράσεις εντός του κυττάρου. Οι μεταβλητές εξόδου, είναι οι συγκεντρώσεις των διάφορων μεταβολιτών, με σημαντικότερους το ATP και το ADP. Συμπεράσματα για την υγεία και την μεταβολική κατάσταση του κυττάρου μπορούν επίσης να αντληθούν από τον λόγο ATP προς ADP.

Χρησιμοποιώντας αντίστοιχη αλληλουχία μη επιτηρούμενων και επιτηρούμενων μεθόδων μηχανικής μάθησης, καταφέρνουμε να προσδιορίσουμε τις θέσεις των κέντρων αντιδράσεων που επηρεάζουν τις συγκεντρώσεις των μεταβολιτών, με τις θέσεις των μιτοχονδρίων να φα-



Σχήμα 5: (α′) Δείγματα των τριών διαφορετικών τύπων διαμορφώσεων κέντρων αντίδρασης: ομοιόμορφη, λογαριθμοκανονική και πολωμένη λογαριθμοκανονική. Η γλυκόζη εισέρχεται στην κυτταρική περιοχή από την αρχή του άξονα, ενώ το γαλακτικό οξύ εξέρχεται από την αντίθετη κορυφή. (β′) Κατανομές τριών μεταβολιτών ενδιαφέροντος για κάθε τύπο διαμόρφωσης των κέντρων αντιδράσεων. (γ′) Οι τρεις υλοποιημένες προσεγγίσεις επιβλεπούμενης μάθησης. (δ′) Διαγράμματα ισοτιμίας εκτός δείγματος για την πρόβλεψη του χωρικού μέσου όρου των [LAC] και [ATP]. (ε′) Πίνακας σύγχυσης εκτός δείγματος για την πρόβλεψη της ενεργειακής κατάστασης του κυττάρου (μη επαρκώς διεγερμένη ή υγιής).

ίνονται να παίζουν τον σημαντικότερο ρόλο για την ενεργειακή κατάσταση του κυττάρου (βασισμένη στον προαναφερθέντα λόγο ATP προς ADP). Προτείνεται επίσης ένα μοντέλο λογιστικής παλινδρόμησης για ταξινόμηση, το οποίο λαμβάνει σαν μεταβλητές εισόδου τις θέσεις των κέντρων αντιδράσεων και προβλέπει εάν το κύτταρο είναι υγιές ή αν δεν έχει επαρκή ενέργεια με εξαιρετική ακρίβεια (95.5%) και ανάκληση (97.53%). Στην συνέχεια προτείνονται διάφορα νευρωνικά δίκτυα για παλινδρόμηση, τα οποία λαμβάνοντας υπόψιν τις θέσεις των κέντρων αντιδράσεων, μπορούν να προβλέψουν είτε τις μέσες συγκεντρώσεις των μεταβολιτών στο κύτταρο, είτε ολόκληρα τα προφίλ συγκεντρώσεων με ικανοποιητικές προβλεπτικές επιδόσεις. Στο τέλος χρησιμοποιείται και μια ανάλυση με τιμές SHAP, δίνοντας έτσι και τη μέση συμβολή της κάθε μεταβλητής στην πρόβλεψη.

Με αυτόν τον τρόπο θεωρούμε πως η ερώτηση 2 που τέθηκε παραπάνω απαντάται επαρκώς, και σε συνδυασμό με την απάντηση στην ερώτηση 1 που έχει ήδη δοθεί, προτείνεται μια

αξιόλογη απάντηση στην ερώτηση του «πώς πρέπει να προσεγγίζονται διεργασίες οι οποίες δεν διαθέτουν τα απαραίτητα δεδομένα για να μοντελοποιηθούν βέλτιστα από μεθόδους αμιγώς βασισμένες στα δεδομένα, άλλα είναι επίσης υπερβολικά περίπλοκες για να μοντελοποιηθούν αποκλειστικά από μεθόδους βασισμένες σε εξισώσεις.

Table of Contents

List of Figures

List of Tables

List of abbreviations

ANN	Artificial Neural Network
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
BO	Bayesian Optimization
CFD	Computational Fluid Dynamics
CVD	Chemical Vapor Deposition
GLC	Glucose
DML	Double/Debiased Machine Learning
DNN	Deep Neural Network
EDA	Exploratory Data Analysis
FEM	Finite Element Method
FVM	Finite Volume Method
GLY	Glyceraldehyde 3-phosphate
LAC	Lactate
ML	Machine Learning
MAE	Mean Absolute Error

MAPE Mean Absolute Percentage Error

MSE Mean Squared Error

RMSE Root Mean Squared Error

PCA Principal Component Analysis

POD Proper Orthogonal Decomposition

PYR Pyruvate

QoI Quantity of Interest

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

VAE Variational Autoencoder

Chapter 1

Introduction

1.1 Motivation

In engineering and other scientific disciplines, the chemical and physical processes that occur within a system can dictate outcomes that range from success to failure, from products that meet specifications to scrap, or even from sustained life to critical failure in biological systems. Understanding, analyzing, and predicting the behavior of these processes is essential, and mathematical modeling serves as a powerful tool to achieve these goals.

Mathematical modeling aims to identify relationships between key quantities (inputs and outputs) in a system. To do this effectively, data are essential. Without data on processes, their parameters, outcomes, and other relevant factors, models cannot be validated or produce accurate quantitative results. This makes it difficult to assess how closely a model reflects reality. The choice of modeling approach depends on the amount of data available.

When data are limited, an equation-based approach (also known as a physics-based or hypothesis-driven approach) is commonly used. This method starts with a clear hypothesis and assumptions, often grounded in existing theories or prior knowledge. First-principles models, which rely on the fundamental equations of a system, fall into this category. These models require experiments to test and validate their hypotheses. However, incorrect assumptions in an equation-based model can lead to significant errors. Furthermore, as modeled systems become more complex and require greater accuracy, the modeling process becomes more challenging. First-principles models typically depend on parameters to produce quantitative results, and if these parameters are unknown, they

must be derived from available data. Additionally, as our understanding of a process evolves or new factors are introduced, these models must be updated, which requires ongoing maintenance.

When data are plentiful, data-driven approaches can be applied directly. These methods typically operate without a predefined hypothesis and aim to uncover insights and detect patterns directly from the available data through computational techniques and statistical analysis. The insights and patterns discovered can then serve as the basis for generating new hypotheses about the process. The larger the dataset and the more relevant the information it contains, the easier it becomes to identify meaningful patterns and relationships.

But what happens when the quantity of data is neither "big" nor "small"? Then hybrid approaches, which try to get the best of both worlds by combining elements from equation-based and data-driven approaches, become necessary. This is quite often the case in many engineering applications, where process insights and outcome predictions are required regardless of the complexity of the process. The modeling of such processes might require many assumptions. Furthermore, real-world data can be "dirty", which means that available process data may lack information or even contain errors.

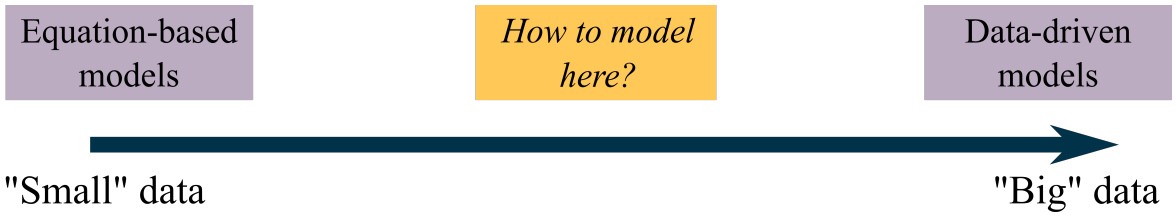


Figure 1.1: Modeling approaches for varying data quantities. When data is scarce, equation based modeling is the most appealing approach. When data is abundant, data-driven modeling can surpass its equation-based counterpart by extracting process insights directly from the data. When data are neither "big" nor "small", hybrid approaches combining both equation based and data-driven modeling can be used to attempt to get the best of both worlds.

An example of such an application is the chemical vapor deposition (CVD) of multilayer Ti(C,N)/α-Al_2O_3 coatings on cemented carbide substrates [1]. This process aims to provide cutting tools with excellent heat and wear resistance properties, together with excellent cutting performance. However, as with all CVD processes, it comes with a high complexity, associated with the interplay of several physical and chemical phenomena (e.g., reactions, diffusion, and con-

vection) that occur in CVD. Furthermore, CVD processes can be highly sensitive to operating conditions, making them challenging to model and, in extension, control, optimize, and predict with accuracy and consistency.

CVD processes, like all processes, can greatly benefit from modeling. Modeling can provide insight into the process and means to improve it. To this end, several approaches can be taken. Traditionally, equation-based methods such as Computational Fluid Dynamics have been used for the modeling of CVD processes [2]–[5], as they can shed light on the way in which the various phenomena contribute to the outcome of the process. Nevertheless, the development of accurate CFD models can be a daunting task, as oftentimes the several parameters required for the introduction of the numerous phenomena are unknown. In certain cases, the geometry of the reactor can change from production run to production run, making the use of a single model less viable. Additionally, as the process can be very sensitive to perturbations, the use of sensors within the reactor is often avoided, and therefore, dynamic process data might not be available.

Apart from the constraints enforced by difficulties associated with model development, conducting computational experiments and making predictions using such models can also be undesirable, especially in cases where time is of the essence. CFD models of CVD processes that aim to take into account all of the interplaying phenomena can be highly resource intensive, if not computationally intractable. Several approaches have been proposed in the literature to overcome these problems, usually by developing surrogates or reduced-order models [6]–[9]. However, the acquisition of an adequate amount of CFD simulations is necessary for the development of these surrogate models. This is a cost (either in resources or time) that needs to be considered before development.

Since the beginning of the Industry 4.0 era, data availability has increased greatly in industrial settings [10]. Although this increase is evident, the scale of the data still remains orders of magnitude lower than the scale of social media or finance data. Regardless, this increase in quantity has led to an increase in the number of applications of data-driven methods in many aspects of industry and an increase in industrial interest in data (the motto "data is the new oil" comes to mind). This led to a large number of research works, with several applications in industrial processes [11]. The topics of focus include maintenance management [12]–[14], quality management [15]–[19], production planning [20]–[22] and control [23], [24], logistics and supply

chain management [25]–[28], and engineering design [29], [30]. Applications have even emerged in management [31].

This increased quantity of data, along with the interest in data-driven approaches and the limitations of equation-based methods, led to the pursuit of data-driven methods for the analysis and prediction of a CVD process in the present work. However, because the interpretation of the interplay between the aforementioned physical and chemical phenomena is of interest, the implementation of data-driven approaches is accompanied by the development of a CFD model of the process. However, it appears that the process is too complex to be efficiently modeled by traditional CFD methods, while also lacking the data required to be optimally modeled by data-driven approaches. This raises the following question: *"How should we approach processes that lack the necessary data to be optimally modeled by data-driven methods, but are also too complex to be optimally modeled solely by equation based methods?"*.

1.2 Objectives

This work aims to answer the above question by answering the following.

1. Using this CVD process as a case study, what is the best computational workflow that allows us to make accurate predictions in a resource-efficient and low-cost way?

2. Can we determine critical process parameters from the available production data?

Question number 1 is answered in Chapters 3 and 4, where both CFD and ML models of the process are proposed. Using this CVD process as a case study, we propose ways of developing an accurate model of the process, using both CFD (Chapter 3) and machine learning (ML) approaches (Chapter 4). Question number 2, is answered first in Chapter 5, where a blend of supervised and unsupervised learning techniques is proposed as a way to identify potential critical process parameters in real-world data. In an attempt to explore and highlight the versatility of the approach proposed in Chapter 5, our second question is answered again in Chapter 6, this time for a completely different application that involves data from a computational model; the metabolism of astrocyte cells.

1.3 Thesis Structure

Chapter 2 presents the interdisciplinary scientific background where CVD, the process of interest, is introduced. Furthermore, a brief introduction to Computational Fluid Dynamics, the equation-based method implemented for the modeling of the CVD process, is given. Finally, Chapter 2 ends with a slightly more detailed introduction to machine learning methods and the various steps necessary for the development of data-driven models. This chapter only briefly mentions concepts that are further analyzed in the following chapters, which are based on the journal publications derived from my research project.

Chapter 3 presents the first step taken in the investigation of the CVD process. It addresses the development of a chemistry-enhanced Computational Fluid Dynamics (CFD) model that allows the prediction of α-Al_2O_3 thickness along with the analysis of several phenomena, which will be presented in detail in Section 2.1.1, involved in the deposition process. The developed model is also used for a series of computational experiments, aiming to get insights regarding the rate-determining mechanism of the α-Al_2O_3 deposition. This chapter was published as a journal article in [32].

Chapter 4 presents a comparison between the developed CFD model and data-driven approaches that can also provide accurate α-Al_2O_3 thickness predictions. Here, several supervised learning algorithms were implemented using only production data, with ensemble tree-based methods [33] and more specifically XGBoost [34] demonstrating the best performance on the dataset. The predictive performance for the α-Al_2O_3 coating thickness is better for the proposed data-driven approach, with a reduction of more than 99.99% in the required resources, compared to the CFD approach. This chapter is finalized with the implementation of another data-driven approach, namely GappyPOD [35]. GappyPOD allows for the reconstruction of the full state-space of the CFD simulations, using only limited data from the simulations, along with certain thickness measurements. This chapter was published as a journal article in [36].

Focusing on entire production runs, Chapter 5 proposes a combination of ML methods for the analysis of the process using historical production run data. We show that when using *only* process outputs that are indicative of quality for clustering, the resulting clusters demonstrate clear

qualitative differences. By analyzing the differences between the process inputs that correspond to each of the resulting clusters, we can identify inputs that are potentially critical to the outcome of the process. Following this step and using the identified inputs, supervised learning algorithms can be proposed for the prediction of the quality of a production run. Finally, a SHAP value analysis [37] is included to quantify the impact of each input on the output of the process. This chapter was published as a journal article in [38].

Subsequently, in Chapter 6 we demonstrate the versatility of the framework developed in Chapter 5, by applying it to an entirely different process; the metabolism of astrocyte cells. Using a computational model developed in the work of Farina *et al.* [39], [40], we show that we are able to determine inputs, in this case reaction centers, that are highly influential on the energy state of the cell. A classification model is developed for the prediction of the state of the cell (healthy vs. non-energized), while regression models allowing for the prediction of metabolite concentrations are also proposed. This chapter is currently in submission and is available as a preprint in [41].

Finally, the results obtained from the different chapters are summarized in Chapter 7, and the thesis is concluded with a discussion of possible future directions.

1.4 Dissemination

My project has led to the following scientific contributions.

Articles

Peer-reviewed journal articles:

- P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, T. J. Mountziaris, and S. P. A. Bordas, "An efficient chemistry-enhanced CFD model for the investigation of the rate-limiting mechanisms in industrial Chemical Vapor Deposition reactors," *Chemical Engineering Research and Design*, vol. 186, pp. 314–325, 2022. DOI: 10.1016/j.cherd.2022.08.005

- P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "Equation-based and data-driven modeling strategies for industrial coating

processes," *Computers in Industry*, vol. 149, p. 103 938, 2023. DOI: 10.1016/j.compind.2023.103938

- P. Papavasileiou, D. G. Giovanis, G. Pozzetti, M. Kathrein, C. Czettl, I. G. Kevrekidis, A. G. Boudouvis, S. P. A. Bordas, and E. D. Koronaki, "Integrating supervised and unsupervised learning approaches to unveil critical process inputs," *Computers & Chemical Engineering*, vol. 192, p. 108 857, 2025. DOI: 10.1016/j.compchemeng.2024.108857

Submitted preprints:

- P. Papavasileiou, S. Farina, E. D. Koronaki, A. G. Boudouvis, S. P. A. Bordas, and A. Skupin, *Machine Learning-based Predictions of Spatial Metabolic Profiles Demonstrate the Impact of Morphology on Astrocytic Energy Metabolism*, 2024. DOI: 10.1101/2024.09.18.613725

Conferences

Oral presentations:

- P. Papavasileiou, E. D. Koronaki, A. G. Boudouvis, S. P. A. Bordas, G. Pozzetti, M. Kathrein, and C. Czettl, "Development of an efficient CFD model for an industrial scale CVD reactor with rotating gas feeding system," presented at the 10th GRACM International Congress on Computational Mechanics, Virtual conference, 2021

- P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, S. P. A. Bordas, and A. G. Boudouvis, "Assessment of CFD and ML modelling strategies for industrial chemical vapor deposition reactors," presented at the 13th PESXM (Panhellenic Scientific Conference on Chemical Engineering), Patras, Greece, 2022

- P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "A comparison of equation-based and machine learning models of industrial scale deposition processes," presented at the ECCOMAS2022 International Congress on Computational Mechanics, Oslo, Norway, 2022

- P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "An efficient CFD model of an industrial scale CVD reactor allowing accurate

coating thickness predictions," presented at the WCCM-APCOM 15th World Congress on Computational Mechanics and 8th Asian Pacific Congress on Computational Mechanics, Yokohama, Japan (virtual congress), 2022

Poster presentations:

- P. Papavasileiou, E. D. Koronaki, D. G. Giovanis, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "Investigation of defining process inputs using unsupervised machine learning," presented at the 14th European Congress of Chemical Engineering and 7th European Congress of Applied Biotechnology, Berlin, Germany, 2023

Chapter 2

Scientific background

2.1 Chemical vapor deposition

2.1.1 Overview

Chemical vapor deposition (CVD) is a widely used chemical process for the production of solid thin film coatings. It involves the reaction of a solid substance on a heated substrate via a series of homogeneous chemical reactions occurring in the gas phase and heterogeneous reactions taking place between the gas and solid phases.

CVD and its various subcategories are widely applied in various thin film technologies, including the creation of semiconductors [47], dielectrics [48], conductive oxides [49], passivation layers [50], [51], and oxidation barriers [52]. They are also crucial in producing coatings that resist heat [53], corrosion [54], and wear [55], [56]. In microelectronics [57], CVD plays a key role in the growth of epitaxial layers [58], [59]. Beyond these uses, CVD is employed in the manufacturing of high-temperature materials such as ceramics [60] and tungsten [61] and the development of solar cells [62], [63]. Furthermore, CVD is used for the fabrication of high temperature fiber composites [64], [65] and the generation of particles with precisely controlled sizes [66].

Despite its wide range of applications, CVD can be considered a complex process, as it involves several competing physical and chemical phenomena. As shown in Fig. 2.1, the main phenomena that occur are homogeneous (gas-gas) and heterogeneous chemical reactions (gas-solid), gas diffusion, adsorption, desorption, and convection. If the numerous intermediate species and byproducts

produced from the aforementioned chemical reactions are also considered, then the complexity of the process becomes evident. As a consequence, CVD processes can be very sensitive to operating conditions and quite difficult to accurately and reliably model, control, optimize, and predict.

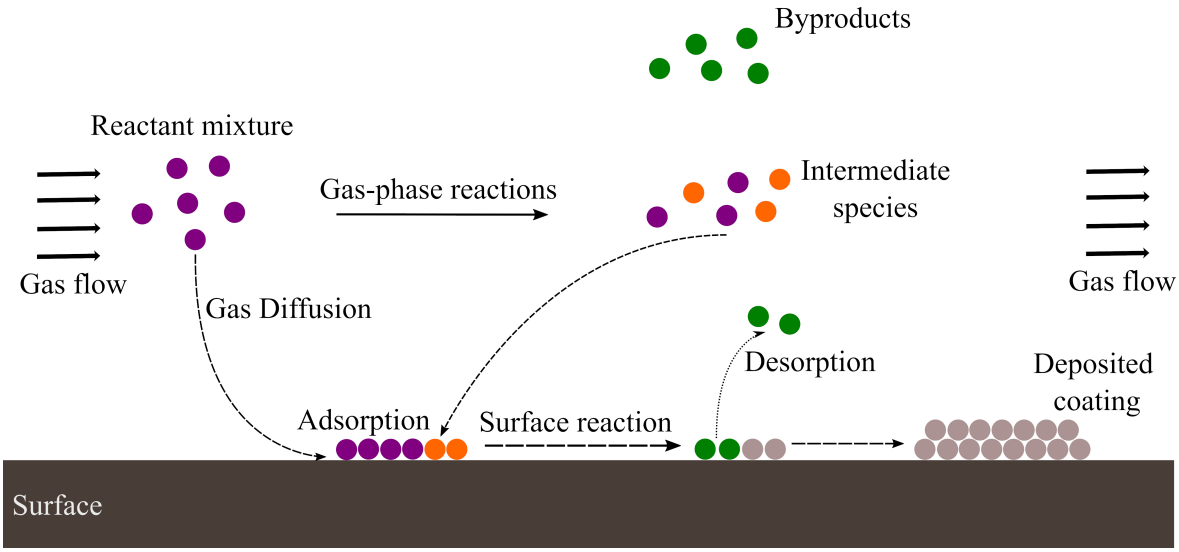


Figure 2.1: An overview of the various interplaying phenomena occurring in a CVD process (Taken from [32]).

2.1.2 Production of wear resistant Ti(C,N)/α-Al_2O_3 coatings for cutting tools

The process investigated in the present work is an industrial-scale process of our collaborators at CERATIZIT. The goal of the process is the deposition of multilayer Ti (C,N)/α-Al_2O_3 coatings on cemented carbide substrates via CVD. A simple example of the coating is presented in Fig. 2.2. The resulting product improves the properties of carbide cutting tools, providing excellent resistance to heat and wear, as well as improved high-temperature hardness [67]–[69].

Figure 2.2: An overview of the material produced. A multilayer coating of Ti(C,N)/α-Al_2O_3 is deposited on a cemented carbide substrate using a two-step coating process.

This multilayer coating is deposited in a commercial industrial CVD reactor (Sucotec SCT600TH), using a sophisticated two-step coating approach. First, a base layer of Ti (C,N) of around 9 μm is deposited using a chemical system consisting of $TiCl_4$–CH_3CN–H_2–N_2–CO at a temperature of 900°C and a pressure of 100 mbar [70], [71]. From the precursors, $TiCl_4$ is used as the Ti source and CH_3CN is used as the source of C and N.

Following the deposition of Ti(C,N), the reactor operating conditions change to T=1005°C and p=80 mbar, in order to accommodate the deposition of α-Al_2O_3. The chemical system used for the deposition of alumina consists of $AlCl_3$–CO_2–HCl–H_2–H_2S. This step takes around 3 hours to complete [1]. $AlCl_3$ acts as the source of Al, while H_2O produced in situ via the reverse water-gas shift reaction [72], [73] acts as the source of O for the coating. A more detailed review of the α-Al_2O_3 coating step is presented in Chapter 3.

The produced cutting tools are consistently required to maintain cutting capacity for the prescribed time indicated by the manufacturer. For this reason, the uniformity of the coating thickness in all production runs, reactors, and production sites is essential, as it contributes to the uniform longevity of the products [74]. Therefore, it is clear that the main quality metric of the process is coating thickness.

For the analysis of the process and the prediction of process outcomes, it is essential to develop models. To this end, it is possible to implement equation-based methods (Chapter 3) or data-driven methods (Chapters 4 and 5), all centered on obtaining a better understanding of the process and the important quality metric of the process, namely the deposition thickness. However, developing such models comes with several challenges, as explained in Chapter 1.

2.2 Equation-based approaches

2.2.1 Computational Fluid Dynamics

Computational Fluid Dynamics (CFD) is a subcategory of fluid mechanics that uses numerical analysis to analyze and solve problems related to flowing fluids. Through the use of computers, CFD aims to solve the fundamental equations of fluid flow (Navier-Stokes) along with conservation equations, conservation equations for energy, species and electric potential among others. CFD

can make all the required calculations that allow for the simulation of the free-flow of fluids as well as the interactions between fluids and surfaces, which are defined by boundary conditions. Taking the example of pouring a glass of water, CFD would allow us to simulate the flow of water, along with the interaction of the water with the inner surface of the glass.

When both fluids and chemical reactions are involved, CFD can still provide answers [75]–[77], taking into account chemical reaction schemes and conservation of species equations. As will be shown in Chapter 3, CFD is widely used for modeling and investigating CVD processes, with great success.

Governing equations

The governing equations for a CFD model usually include the conservation equations of mass, momentum, and energy. When modeling a chemical process, such as CVD, where several heterogeneous and homogeneous reactions determine the outcome of the process, including equations for the transport of chemical species and chemical reactions becomes necessary. This adds a layer of complexity to the model. Of course, with some realistic assumptions, these equations can be simplified, as will be shown in Chapter 3.

- Conservation of mass

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \tag{2.1}$$

  where $\mathbf{u}$ denotes the gas velocity vector and ρ the density of the gas mixture.

- Conservation of momentum

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho \mathbf{u} \cdot \nabla(\mathbf{u}) = -\nabla p + \nabla \cdot \left[\mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T) - \frac{2}{3}\mu(\nabla \cdot \mathbf{u})\mathbf{I} \right] + \mathbf{F} \tag{2.2}$$

  where p is the pressure, μ is the dynamic viscosity of the gas mixture, $\mathbf{I}$ is the identity tensor and $\mathbf{F}$ is the volume force vector.

- Conservation of energy

$$\rho C_p \left(\frac{\partial T}{\partial t} + (\mathbf{u} \cdot \nabla) T \right) = - (\nabla \cdot \mathbf{q}) - \frac{T}{\rho} \frac{\partial \rho}{\partial T}\bigg|_p \left(\frac{\partial p}{\partial t} + (\mathbf{u} \cdot \nabla) p \right) + Q + \left[\mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T) - \frac{2}{3}\mu(\nabla \cdot \mathbf{u})\mathbf{I} \right] : \left[\frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T) \right] \quad (2.3)$$

  where C_p is the specific heat capacity of the gas mixture at constant pressure, T is the absolute temperature, $\mathbf{q}$ is the heat flux vector and Q contains the heat sources. The : symbol denotes a double dot product.

- Transport of chemical species

$$\rho \frac{\partial \omega_i}{\partial t} + \rho(\mathbf{u} \cdot \nabla)\omega_i = R_i - \nabla \cdot \mathbf{j_i} \quad (2.4)$$

  where ω_i denotes the mass fraction of the i species and $\mathbf{j_i}$ and R_i are the diffusion flux and the reaction rate of species i, respectively. The diffusion flux ($\mathbf{j_i}$) for each component i can be calculated using a diffusion model (e.g., Maxwell-Stefan).

For a steady-state solution of the above system of equations, only boundary conditions are required. If solved in transient conditions, initial conditions are also necessary.

Solving the above system of equations requires discretization of both the spatial and temporal domains. For the discretization of the spatial domain, methods such as the Finite Element Method (FEM) or the Finite Volume Method (FVM) can be used. For transient problems, there are implicit discretization options such as BDF and generalized-α and explicit methods such as Adams–Bashforth or Rugge-Kutta. The selected method always depends on the nature of the problem and potential resource limitations.

2.3 Machine learning

Taken from the book of Mitchell [78], we can find a definition of machine learning:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by

P, improves with experience E.

Let us say that we want to predict the outcome of a football match our favorite team is playing. The possible outcomes are victory, draw, or defeat. An outcome predictor could be a machine learning algorithm that, given examples of previous matches the team has played (e.g., match statistics, squad selections, weather, etc.) and the outcome of each match, can learn to predict the outcome of the upcoming match. The football matches that the algorithm uses to learn are called the training set. Each training example is called a training sample. In this case, the task T is to predict the outcome of the match our favorite team is playing, the experience E is the training data, consisting of the historical performance of the team, and the performance measure P must be defined; for example, we can use the ratio of correctly predicted match outcomes. This particular performance metric is called accuracy and is often used in classification tasks.

To perform this task without machine learning, one must intently focus on the historical data, try to find patterns or rules that result to each of the potential outcomes, and then evaluate the developed rules based on their performance. It is obvious that this task would be very difficult and would require immense time resources.

In contrast, ML can autonomously identify patterns in historical data (e.g., a specific player making a huge difference, the team showing an advantage in adverse weather conditions) and evaluate them using historical data, eventually ending up with rules that provide the optimal performance for the given dataset.

Based on the discovered rules, and by interpreting the developed model, we can gain insight on our favorite team, understand what makes the team win or lose and even appreciate players or coach decisions that we previously couldn't fathom.

Football fans, statisticians, and perhaps gamblers alike understand that this task is not as simple as the example makes it appear. This is mainly due to the stochasticity associated with football. However, it serves as a great example that shows the strengths of machine learning approaches in cases where the data is abundant.

Apart from classification tasks, such as the one described above, there are also regression tasks, where continuous values are predicted. An example of such a task, still relevant to football, would be the prediction of attendance to the team's next game. Based on ticket price, attendance in

previous matches, and current form of the team, the algorithm could output a real number, which would then be rounded up or down to the nearest integer).

There are also clustering [79] tasks. Let us say that the football team wants to identify groups of similar fans within their fan base, in order to target them with special offers, events, and promotions. The algorithm would take the characteristics of the fans as input and provide a label for each fan, indicating to which cluster they belong. In contrast to the classification and regression tasks, where validating the output of the algorithm is easy (the outcome of the game is readily available, as well as the attendance), the output of clustering algorithms cannot be directly validated and requires further analysis. This lack of actual outcome (or label, as is more commonly known in ML), classifies clustering as an unsupervised learning [80] method. In contrast, regression and classification belong to the group of supervised learning methods [81].

To accommodate the different types of tasks, different performance metrics must be implemented. Accuracy is a popular metric for classification, root mean square error is a popular metric for regression, and within-cluster sum of squares is a popular metric for clustering. Nevertheless, many more performance metrics are available (or can be constructed) based on the task at hand.

Machine learning is very versatile and can be used in a wide variety of applications where data are available. Its different disciplines (supervised/unsupervised learning) allow its implementation for tackling various tasks and gaining insight for several different processes, be it football or industrial chemical vapor deposition processes.

2.3.1 Unsupervised learning

Clustering

Clustering algorithms function by evaluating the dissimilarity or similarity between data points. Through this process, they form clusters by grouping similar observations, helping to uncover patterns or structures in the dataset.

Clustering algorithms usually use a similarity matrix as input, which includes pairwise dissimilarities between all observations. For quantitative variables, the Euclidean distance is the most commonly used metric, measuring the straight-line distance between two points in feature space. However, selecting an alternative distance metric may lead to different clustering outcomes [80],

[82].

Clustering algorithms can be grouped into different families based on their approaches. There are partitional methods (i.e., k-means clustering), which assign observations to clusters iteratively, considering distances from centroids. These require an a priori decision on the number of clusters and are sensitive to initialization [83].

Density-based methods, such as OPTICS and DBSCAN, detect clusters by examining regions of high density separated by low-density areas. The key parameters for these methods include the minimum number of points required to form a cluster and the minimum distance between core points [84]–[86].

Another family of clustering methods are hierarchical methods. These methods link data points according to specific criteria. Agglomerative clustering starts with a number of clusters equal to the number of observations and progressively combines clusters until a single cluster remains. Divisive clustering starts with all observations in one cluster and splits them iteratively until there are as many clusters as the number of observations. Distance metrics are used to calculate the dissimilarity between clusters, and the merging (or splitting) of clusters depends on the linkage criterion used. The choice of distance metric and linkage criterion can significantly influence the results [87], [88].

2.3.2 Supervised learning

Supervised learning algorithms differ from unsupervised ones as they use labeled data, where the model inputs (or features) x_i are paired with their respective outputs (or responses) y_i. These models leverage data to forecast outcomes for future observations. Supervised learning includes regression for continuous variables and classification for binary or ordinal outputs [81].

The available supervised learning methods are analyzed in depth in Chapter 4. However, for the sake of completeness, a brief overview of the methods evaluated for this work is presented here. Several common supervised learning methods include the following categories: (a) Linear methods, such as linear regression, lasso regression [89], ridge regression [90], and logistic regression, which is primarily used for classification tasks. (b) Support vector machines (SVMs) [91], which can be linear or non-linear depending on the chosen kernel, are commonly applied to classification

problems. (c) Tree-based methods, which encompass classification and regression trees [92], as well as ensemble techniques such as random forests [93], gradient-boosted trees [94], extra trees [95], and XGBoost [34], all of which combine multiple trees to enhance performance [96]. (d) Artificial neural networks (ANNs), with their various architectures [97], offer flexible approaches suitable for classification and regression tasks.

Artificial neural networks (ANNs) are a subset of machine learning models that has recently drawn a lot of attention. ANNs are inspired by the intricate organization of biological neural networks in the brain. ANNs consist of interconnected nodes known as neurons, each loosely modeling its biological counterpart. These neurons are linked by edges, similar to synapses in the brain. The network architecture involves layers of neurons, with signals flowing from the input layer to the output layer. Intermediate layers, sandwiched between input and output, are termed hidden layers. When the number of hidden layers is two or more, the network is classified as a deep neural network (DNN).

Neurons in each layer receive signals from the preceding layer, process them, and transmit an output signal to the subsequent layer. The output computation of each layer (with the exception of the output layers) employs a non-linear activation function. Each neuron has associated parameters, weights and biases, that are optimized during network training. In supervised learning scenarios with labeled data, these parameters are fine-tuned to minimize the discrepancy between predicted and actual target values. ANNs excel at modeling intricate relationships due to their architectural complexity, rendering them a powerful tool in various machine learning tasks.

The TensorFlow [98] and Keras Python libraries [99] are used for the development and training of the ANN models presented in this work, and the scikit-learn library [100] is used for the application of the rest of the ML methods.

2.3.3 Data-driven pipeline

Developing a predictive model is not as simple as training the model. There are several steps that must be taken to ensure optimal model performance. These steps involve cleaning the data, obtaining an initial understanding of the data, and determining potential cases for predictive models. After determining potential cases where supervised learning approaches can be used to

make predictions, the available features and their relevance to the output must be considered. Not all features might be relevant to the output. Furthermore, new features based on the available once might need to be created.

When the time for training the model comes, one must ensure that the model is properly trained and is able to generalize well on unseen data. Avoiding overfitting is a common goal when training predictive models and involves several different methods related to the data, the model, and the learning process, as discussed in Section 2.3.3.

Data "clean-up"

Data, and especially real-world data can often be "dirty". This means that the data can have peculiar formats, certain observations might miss certain entries, or certain entries might contain errors. The model developer must first bring the data in a format that can easily be manipulated in an environment that allows for model development. Furthermore, the modeler has to decide whether observations containing missing entries need to be discarded or if the missing entries should be imputed. Finally, when it comes to potential errors or mistypes within the data, the modeler should first detect the errors and then decide on taking corrective action or discarding the observations.

Exploratory data analysis

Following the "clean-up" of the data, it is time to look and *see what it seems to say* [101]. Exploratory data analysis (EDA) is just that. It involves data visualization, through histograms, box plots, dot plots, etc., along with basic statistical analysis for the variables present in the dataset (e.g., calculation of the mean, the median, and standard deviation alongside the quantiles of continuous variables) [102]. Dimensionality reduction or clustering could also be considered part of EDA, as they allow for the identification of potential patterns in our data.

EDA allows us to process the data, detect abnormalities, and identify potential patterns in our data. It allows us to get a comprehensive grasp of our data before jumping into more rigorous data analysis and modeling.

Feature engineering

Feature engineering is another crucial step that precedes the training of the predictive model. The performance of most machine learning models greatly depends on the feature vector representation. Consequently, data scientists invest significant effort in designing preprocessing pipelines and data transformations [103].

Feature engineering usually takes place in parallel with EDA. It involves the creation of new features, based on preexisting ones, in an attempt to find the best feature representation for the predictive model that we wish to develop [104]. These engineered features could, for example, be the product of two features, the inverse of a feature, the sum of the absolutes of two features, or the maximum/minimum value between several features.

Last but not least, feature engineering also involves the encoding of categorical variables [105] and the transformation (e.g., standardization, min-max transformation) of numeric variables, thus bringing the data in a model-friendly format.

Overfitting and how to avoid it

A very common problem supervised learning approaches suffer from is overfitting. When we say that a model is overfitting, we mean that the model performs very well on the training data but nevertheless performs poorly on the unseen test data [106]. The overfitted model effectively "learns" the data of the training set, without capturing the underlying principles [107], and thus failing to generalize on the test set.

Overfitting can occur for several reasons. First, it can be caused by the selection of a very complicated model. Furthermore, it can also be due to the selection of non-informative features or the selection of more features than necessary for satisfactory prediction performance [106].

Strategies to avoid overfitting can be employed when it comes to data, model characteristics, and the learning process of the model.

Cross-validation On the data side, cross-validation [108], [109] is one of the most popular approaches. Cross-validation is used for estimating the prediction error of a model. For example, in k-fold cross-validation, the dataset is split into k groups. One of the groups is kept unseen by

the model during training and acts as a test set. The rest of the groups make up the training set. This procedure is repeated until each of the k groups has been used as the test set. The predictive performance of the model is then evaluated considering all k test sets, using the mean and standard deviation of each performance metric. In this way, the performance metrics provided by cross-validation are more representative of the model's performance than the performance metrics of the model on a single held-out test set. In general, cross-validation is crucial and should be incorporated in other steps of model development, where predictive accuracy is of interest (e.g. feature selection or hyperparameter optimization).

Feature selection and feature extraction Two further measures against overfitting are feature selection and feature extraction. When machine learning algorithms are applied to high-dimensional data, the phenomenon of the curse of dimensionality comes into play. This refers to the sparsity of data in high-dimensional spaces, which negatively affects algorithms designed for low-dimensional spaces [110], [111].

Feature extraction is based on dimensionality reduction techniques such as PCA [112] or LDA [113] and reduces the dimensionality of the data by projecting high-dimensional features to a new low-dimensional feature space. Feature selection, on the other hand, aims at selecting an important feature subset directly from the high-dimensional space. There are three main categories of feature selection methods: a) Filter methods which evaluate the importance of features by examining statistical measures, focusing on their correlation with the target variable. This approach selects features based on their standalone attributes, independently of any specific machine learning algorithm used. Examples of filter methods are the calculation of correlation between a variable and the target output or the chi-square score in the case of categorical outputs. b) Wrapper methods iteratively take advantage of the learning performance of the predictive model to assess the quality of selected subset of features. Any algorithm that can actively search for optimal subsets of features in the complete data set can be considered a wrapper method, from exhaustive search [114] to evolutionary algorithms [115]. c) Embedded methods take advantage of the structure of the predictive model to select relevant features. Examples of embedded methods are random forests [93] or Lasso regression [89].

Data augmentation Data augmentation is another approach on the data side, with the aim of tackling overfitting and improving model performance and robustness [116], [117]. In principle, the goal is to artificially enhance the dataset by adding slightly modified versions of existing observations [118]. In the example of image recognition, the images present in the dataset might be randomly rotated, mirrored, resized, and cropped in an attempt to improve model performance [119]. For tabular data, the synthetic minority oversampling technique (SMOTE) [120] and its variants [121] are widely used in cases of imbalanced classification datasets. Variational Autoencoders (VAEs) [122] have also been proposed as a data augmentation method for tabular data [123].

Early stopping In supervised learning, model training involves the minimization of a cost function. When an iterative method (such as gradient descent) is used to determine the model parameters (e.g. weights and biases in a neural network) that minimize the cost function, early stopping [124] can be used as a tool to avoid overfitting. Early stopping, in principle, stops the model training if the predictive performance on an unseen dataset does not improve for a certain number of iterations. In this way, the model is deterred from "memorizing" the data of the training set, and its generalizability is retained.

Regularization Regularization is another concept related to the objective function. In general, the output of the model can be determined by several features. However, not all of these features might have limited effect on the model output and might hinder the model's predictive performance. One way of limiting this negative influence of these features on the model is feature selection (discussed above). Another way of limiting the effect of these useless features is through the use of regularization [125]. A common way to do this is to add a "penalty" term in the cost function relative to the variable weights of the model. This term is usually the l1- (sum of absolute values) or l2-norm (sum of squared values) of the variable coefficients. There is usually a user-defined tuning parameter that multiplies this norm based on the desired regularization effect.

Another regularization method used exclusively in ANNs is dropout. The concept of dropout refers to the exclusion of certain units, both hidden and visible, within a neural network. This exclusion involves temporarily eliminating a unit from the network along with its associated incoming and outgoing connections. The selection of units to be dropped is performed randomly with

a predefined probability. The application of dropout to a neural network is equivalent to sampling a pruned network from it. At the time of testing, the pruned networks are combined in order to make the prediction [126].

Hyperparameter optimization The hyperparameters of the predictive models play an important role in its predictive performance. Optimizing a machine learning model to suit various tasks requires fine-tuning of its hyperparameters. The selection of an optimal hyperparameter setup for machine learning models critically influences the overall performance of the model [127]. The hyperparameters of the model can be related to its architecture (e.g. number of layers and number of neurons in an ANN, or the maximum depth and number of estimators in a random forest model). They can also be related to the type of regularization used (dropout, l1 or l2 regularization) or the learning process (e.g. learning rate) of the model. It is evident that the variable space for these hyperparameters is not continuous; therefore, the application of gradient-based methods for the optimization of the model can be challenging, if not forbidding.

Grid search or random search are two widely used gradient-free methods for hyperparameter optimization. In grid search, the selected numerical hyperparameters are dispersed at uniform intervals across the user-defined domain, ensuring an exhaustive examination of the parameter space. For categorical hyperparameters, all values are considered (unless a specific subset is selected by the user). In grid search, all possible combinations are considered, which can lead to prohibitive costs, especially as the number of tested hyperparameters increases. Random search aims to address this problem by randomly selecting combinations of hyperparameters within the domain, given resource constraints. In these two methods, each hyperparameter configuration is independently evaluated [128].

Another category of gradient-free methods used for hyperparameter optimization is population-based methods, such as genetic algorithms and particle swarm optimization [129]. These algorithms sustain a population, which is essentially a collection of hyperparameter configurations, and enhance this population by employing local modifications, also known as mutations. In addition, they combine distinct members through a process known as crossover to generate a subsequent generation composed of superior configurations. These algorithms could be particularly suitable for

hyperparameter optimization problems with large configuration space due to their high efficiency [128].

Another popular hyperparameter optimization method is Bayesian optimization (BO) [130]. It is an iterative algorithm whose strategy is to model the relationship between the hyperparameters and the performance of the model (or the cost function). This relationship is modeled using Gaussian processes [131] or random forest [132] surrogate models. The cost function is evaluated for a randomly sampled set of hyperparameters. Subsequently, BO then uses these data to train the surrogate model, which outputs estimates for the cost function, as well as the prediction uncertainty. These estimates can produce a predictive distribution for the different hyperparameter configurations. Based on the predictive distribution, BO establishes a cheap-to-evaluate acquisition function that balances between exploitation and exploration.Exploration involves the systematic sampling of instances in regions that have not yet been examined, in order to uncover potentially significant areas which may have been previously overlooked. In contrast, exploitation focuses on sampling within the currently identified promising regions. These regions are determined based on the posterior distribution and are believed to have a higher likelihood of containing the global optimum [128]. BO models aim to strike a balance between exploration and exploitation to identify the most probable optimal regions while ensuring that better configurations in unexplored areas are not overlooked. BO approaches can usually detect near-optimal hyperparameter combinations in a few iterations [133].

In general, hyperparameter optimization is an indispensable part of model development, as it can address the problem of overfitting by adjusting several characteristics of the model and the learning process simultaneously and without requiring a lot of user input. The user is only asked to select a hyperparameter optimization method and to provide instructions and constraints regarding the search space.

Chapter 3

Development of an efficient chemistry-enhanced CFD model

This chapter is reproduced from P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, T. J. Mountziaris, and S. P. A. Bordas, "An efficient chemistry-enhanced CFD model for the investigation of the rate-limiting mechanisms in industrial Chemical Vapor Deposition reactors," *Chemical Engineering Research and Design*, vol. 186, pp. 314–325, 2022. DOI: 10.1016/j.cherd.2022.08.005. As the first author of the article, I participated in the development of the proposed CFD model, the validation and visualization of the results, and the writing of the original draft of the manuscript.

According to the Contributor Role Taxonomy (CRediT), this corresponds to the following roles: Investigation, Methodology, Validation, Visualization and Writing – original draft.

Abstract

An efficient CFD model for the deposition of alumina from a gas mixture consisting of $AlCl_3$, CO_2, HCl, H_2 and H_2S in an industrial CVD reactor with multiple disks and a rotating feeding tube, is proposed. The goal is twofold: (i) to predict the thickness of the deposited material, (ii) to investigate whether the process rate is determined by the reaction rate or by diffusion. A reaction model that consists of a gas-phase homogeneous reaction and a heterogeneous reaction is implemented, with a proposed kinetics rate that includes the effect of the H_2S concentration. The latter has a catalytic effect, but the mechanism is not entirely understood. The entire reactor geometry (consisting of 40-50 perforated disks) is divided into appropriately chosen 7-disk sections. The 2D, time-dependent CFD model is validated using production data for the deposition thickness. The proposed computational tool delivers accurate predictions (average relative error 5%) for different geometries corresponding to real reactor set-ups. Extending the functionality beyond prediction, a computational experiment is performed to illuminate the interplay between species diffusion and chemical reaction rates, which determines the rate-limiting mechanism. The results indicate that species diffusion is fast enough and therefore reaction kinetics determine the overall deposition rate.

3.1 Introduction

Chemical Vapor Deposition (CVD), where a solid coating is deposited on a heated surface from a mixture of gas reactants, is used for various applications, including microelectronics [57], polymers for microfluidics, sensors, and membranes [134] and wear resistant coatings [55], [56]. It is a complex process involving competing physical phenomena, such as convection, diffusion, and chemical reactions. The balance established between transport phenomena and chemistry is critical for determining the efficiency of the process and the quality of the produced material.

Computational Fluid Dynamics (CFD) models of CVD processes that account for the transport of mass, momentum, and species inside reactors have been proposed in order to elucidate the interplay between the different mechanisms and its effect on the process [135]–[138]. Such models have also been used for optimizing the design of CVD reactors [139], [140], as well as for predicting

reaction rates inside the reactor [141]. Zou *et al.* [142] successfully tried to analyze industrial CVD reactors using a porous media approach. This approach was used in order to tackle the difficulty of explicitly modelling the large amount of substrates in the system by modelling the substrate-packed drawers of the reactor as porous media. Others have thoroughly investigated the reactant gas flow regimes inside of the reactors [143], as well as the effect of the flow on the produced coatings [4], [144]. Mitrovic *et al.* in a series of publications [3], [145], [146], analyzed the flow inside a rotating disk reactor for different process parameters using CFD, determined the optimal parameters for the application and then optimized the reactor design by using the results of the simulations. Nevertheless, their work did not include a chemistry model and hence it was not possible to assess the effect of the flow on the deposited film.

Despite the progress in computer-aided analysis of CVD reactors, important challenges remain, especially in industrial-scale processes:

1. Industrial CVD reactors have a complex geometry in order to increase the coated surface and the throughput of the process. This translates into time-dependent models involving three-dimensional computational geometries, often with moving mesh and therefore, increased level of computational complexity and cost.

2. The actual network of gas-phase and heterogeneous reactions that ultimately lead to deposition, are often not completely known. For example, in the chemical system studied here, the role of hydrogen sulphide is not entirely understood, although its positive effect of the deposition rate has been widely observed [147]–[149].

3. Even when there is a well-established chemical network, it often involves dozens of reactions and intermediate species. Integrating such a chemistry pathway in a CFD model, would make it computationally intractable. Moreover, even when the chemical system is known, ie. the specific reactions and their kinetic rates, the effective reaction rates have to be determined for the particular application and geometry.

4. The geometry of the reactor changes, even in a day-to-day basis in industrial practice. This is not true for every type of CFD application, but it is particularly true in the industry of cutting

tools and wear resistant coatings. Therefore, it is important for the usability of the model to easily accommodate changes in the computational geometry in an almost automatic way.

Points 2 and 3 have been addressed in the past by developing reduced order models of CVD in conjunction with deposition chemistry models [4], [9], [143], [150], [151]. Nevertheless, these reduced order strategies require large amounts of data from detailed models which are often computationally intractable.

In this work we focus on addressing the combination of points 1, 2 and 3 in an industrial-scale CVD application and illustrate the implementation of an efficient modeling strategy that hinges CFD with an effective deposition model, validated by production data. The geometry of the investigated CVD reactor changes on a day-to-day basis, which is why addressing Point 4 is important and will be the subject of future work. Despite the simplifications of the CFD model, we present its potential not only as a predictive tool but also as a means of suggesting the dominance of reaction kinetics in terms of determining the rate-limiting steps of the process. This is an important contribution, because in the context of an industrial process, it is not always feasible to measure the deposition rate experimentally in different temperatures and produce an Arrhenius plot to map out the diffusion and reaction limited regimes.

The application addressed here, is the deposition of alumina onto three-dimensional cemented carbide cutting tools with a well-established thermal LP-CVD process from a gas mixture consisting of $AlCl_3$, CO_2, HCl, H_2 and H_2S [1] in a commercial reactor consisting of several perforated disks and a rotating inlet tube (Sucotec SCT600TH). Several other suggested CVD processes exist for the deposition of Al_2O_3, such as a MO-CVD process utilizing aluminium tri-isopropoxide (ATI) as a precursor [152] or a PE-CVD process utilizing dimethylaluminum isopropoxide (DMAI) as a precursor [153]. These processes not only require a lower thermal budget, but also utilize a safer gaseous atmosphere. However, for our specific application (i.e. wear resistant coatings for cutting tools) and because of the targeted properties of the alumina coating, the aforementioned thermal LP-CVD process is used.

Alumina is very popular for wear-resistant coatings [55], [56] because of its properties [154], [155] and the improved chemical stability and high temperature hardness it provides in Al_2O_3/TiCN multilayer coatings [67]–[69]. The effect of process conditions on the growth and texture of

α-Al_2O_3 has been studied [149], since both directly influence the final properties of the α-Al_2O_3-coated cutting tools. However, little work has been done on the process of the CVD of hard coatings.

The following sections are structured as follows: The geometry and operation of the studied chemical vapor deposition reactor is presented in Section 3.2. The details of the developed CFD model are discussed in Section 3.3. Subsequently, the results of the CFD model are detailed in Section 3.4 along with an analysis of the rate-determining step of the process, followed by the conclusions in Section 3.5.

3.2 Chemical vapor deposition reactor geometry and operation

3.2.1 Reactor set-up and process conditions

This work focuses on the CVD of alumina on cutting tools, referred to henceforth in the text as inserts. An overview of the phenomena taking place inside a CVD reactor is presented in Figure 3.1. Inserts have various shapes and sizes (Fig. 3.2a) depending on their use in industrial applications but are invariably required to maintain cutting capacity for the prescribed time indicated by the manufacturer [74]. For this reason, the special coatings deposited, such as the alumina coating studied here, not only increase longevity but also ensure the expected usability of the insert.

The deposition of alumina on the inserts is studied in a commercial, industrial CVD reactor (Sucotec SCT600TH) which typically consists of 40-50 perforated disks, stacked one on top of the other shown in Fig. 3.2b. For reasons of clarity, a partial schematic of the reactor, depicting 3 disks is shown in Fig. 3.2c. The inserts are placed on the disks, as shown in Fig. 3.2d, while carefully designed perforations allow for the transport of the gas reactants between the disks and around the inserts. For each type of insert, there is a dedicated design of perforated disk, to accommodate the particular geometric characteristics. The mixture of gas reactants enters the reactor through a cylindrical tube at the center of the disk structure, through two inlet holes per disk, placed antipodally (shown in red in Fig. 3.2c). There is a 60° angle difference between the inlet holes of each disk-level of the reactor and the feeding tube rotates at a constant speed of 2 RPM. The gas mixture exits the reactor through holes in the perimeter of each disk (shown in

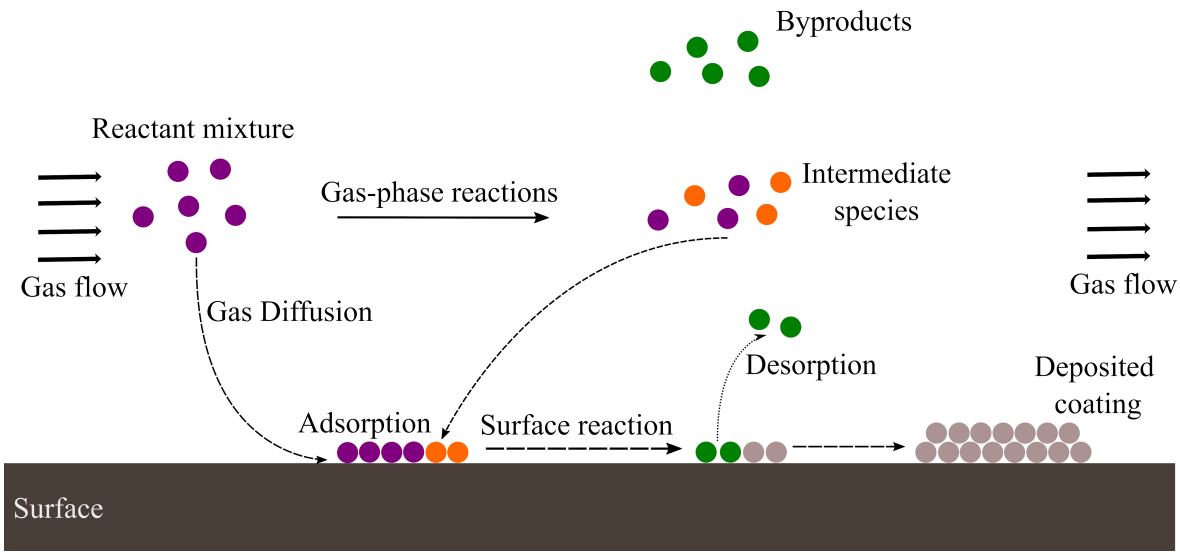


Figure 3.1: Overview of the interplaying mechanisms and phenomena of a CVD process.

blue in the schematic of Fig. 3.2c).

A two step coating process takes place inside the reactor [155]. As a first step, a Ti(C,N) base layer of about 9 µm is grown on the cemented carbide cutting inserts. An α-Al_2O_3 layer is then deposited from $AlCl_3$–CO_2–HCl–H_2–H_2S at $T = 1005°C$ and $P = 80$ mbar. The inlet gas reactant volumetric fractions are 1.7% for $AlCl_3$, 3.7% for CO_2, 2.1% for HCl, 92.2% for H_2 and 0.3% for H_2S. The total inlet gas flow rate is 65 $\mathrm{L} \cdot \mathrm{min}^{-1}$ ($P = 80$ mbar, $T = 1005°C$) [1].

3.2.2 Available production data

The production data available to validate the proposed model, are a total of 15 coating thickness measurements on inserts placed at selected locations inside the reactor, shown in Fig. 3.3.

For each production run, the coating thickness on the inserts at five disks are considered:

1. The top insert-containing disk of the reactor.
2. The 3rd or 4th disk from the top.
3. The middle disk.
4. The 3rd or 4th disk from the bottom.

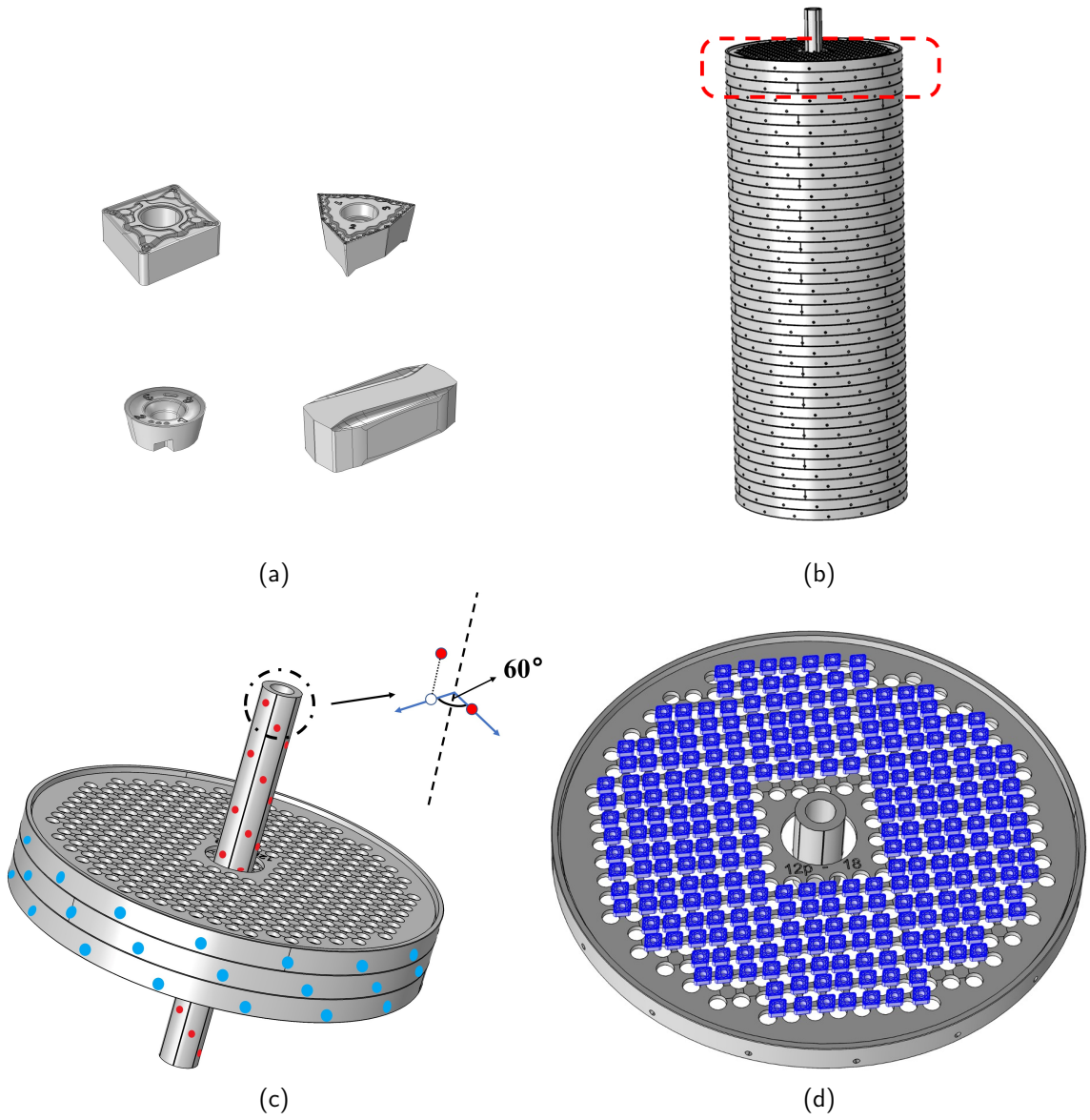


Figure 3.2: (a) Examples of the different cutting tool inserts that are coated inside the reactor. The shapes and sizes of the different inserts coated in the same reactor may differ significantly. (b) A 3D representation of the entire reactor. (c) A close up representation of a 3-disk part of the reactor. The rotating inlet tube passes through the center of the stack of disks. The gas reactants enter through the perforations shown in red. There are two holes per disk level, placed so that there is a 60° angle between the holes in neighbouring disks. The gas outlets are shown in blue. (d) An example of a perforated disk loaded with inserts. The inserts are shown in blue.

5. The lowest insert-containing disk of the reactor.

On each of the aforementioned disks, there are 3 positions of interest. Specifically:

1. The position closest to the inlet, R_0.
2. The position in the mid-distance between the inlet and the outlet, $R_{1/2}$.
3. The position closest to the outlet, R.

All measurements are in μm, with a precision of 0.1 μm.

3.3 Description of the CFD model

3.3.1 Governing equations

The governing equations include the conservation of mass and momentum, as well as the equations for the transport of chemical species and the occurring chemical reactions. A detailed overview of the system of equations can be found in the publication of Gakis *et al.* [4].

The reactor's operating temperature is considered constant in the entire domain. This is due to the fact that the entire reactor set-up is placed inside a furnace and therefore the entire system is heated up to a tightly controlled temperature of 1005°C. The ideal gas assumption is made for calculating the density of the gas mixture. The flow is considered laminar and incompressible. All calculations are made in transient conditions to account for the rotation of the gas inlet tube.

The equations were discretized with the finite element method and solved using COMSOL Multiphysics®. Linear basis functions are used for the continuity equation and quadratic functions for the rest. The computational geometry is presented in detail in the following paragraph.

3.3.2 Computational geometry

The reactor geometry is inherently non-axisymmetric and time-dependent due to the rotation of the vertical tube and the placement of the inlet holes. Therefore, a fully representative simulation would have to account for the entire 40-50 disks, in 3D, while also being time-dependent with a moving mesh. This, however, would come hand in hand with a significant computational cost,

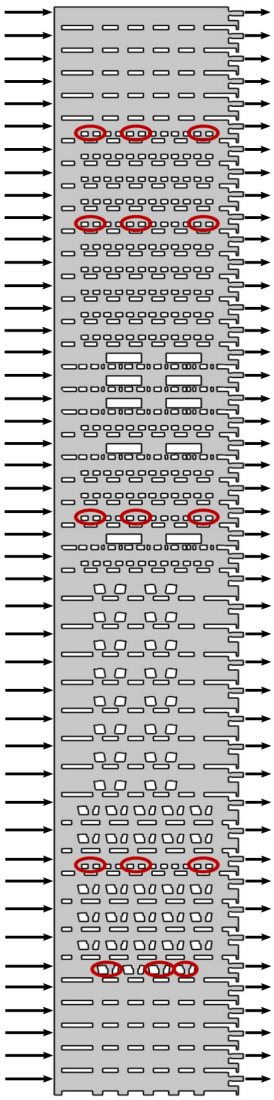


Figure 3.3: A 2D representation of the entire reactor indicating in red the 15 positions with available α-Al_2O_3 coating thickness measurements. The leftmost position is the one closest to the inlet. The arrows indicate the gas reactant inlets and outlets.

even when excluding the mass balances of the species that participate in the multitude of chemical reactions that will be discussed in detail in the following paragraph.

A two-dimensional computational geometry is proposed with appropriately selected boundary conditions. Furthermore, the computational domain does not include all the disks but rather accounts for parts of the reactor, containing 7 disks (cf. Fig. 3.4). The number of disks in the

model is determined by gradually decreasing the number of disks (from 11 to 7) and comparing the deposition thickness in the middle disk with the available production data. By gradually decreasing the number of disks considered, we are able to select the lowest number of disks for which the effects of the top and bottom boundary conditions do not affect the prediction of film thickness at the disk of interest (the middle disk). Another aspect that contributes to efficiency is linked to the fact that different reactor set ups may have several 7-disk parts in common. For example, an alternative reactor configuration could contain the same 7-disk sequence. In this fashion, it is possible to draw conclusions for several combinations of the 7-disk model, that would otherwise require the solution of new entire reactor models each time.

By using this 7-disk, two-dimensional approach and by simulating for 2 periods (or turning cycles) of the process, we can in turn average the deposition rates on each insert and obtain an equivalent deposition rate for several positions of interest inside the reactor.

3.3.3 Boundary conditions

To account for the rotating inlet tube, in the context of a two-dimensional geometry, time-dependent inlet boundary conditions are applied. The perforations of the rotating tube, through which the gases are introduced into the reactor are represented by a fixed inlet boundary in the computational geometry in each disk level. The gas feed velocity is prescribed at each inlet as a time-dependent function that varies between 0 and $V_{\max}$ as a pulse that mirrors the rotation of the inlet tube. The maximum velocity value ($V_{\max}$) is determined based on the experimental conditions and the geometry. Specifically, the following are taken into account:

1. The inlet tube rotates with a rotational speed of 2 RPM.
2. The total inlet gas flow rate is 65 $\mathrm{L} \cdot \mathrm{min}^{-1}$ ($P = 80$ mbar, $T = 1005$°C).
3. There is an average of 35 disks per run.
4. There are two perforations on the inlet tube for each disk. These two perforations are antipodal and their average diameter is 0.002 m.
5. There is a 60° angle difference between the perforations for each disk.

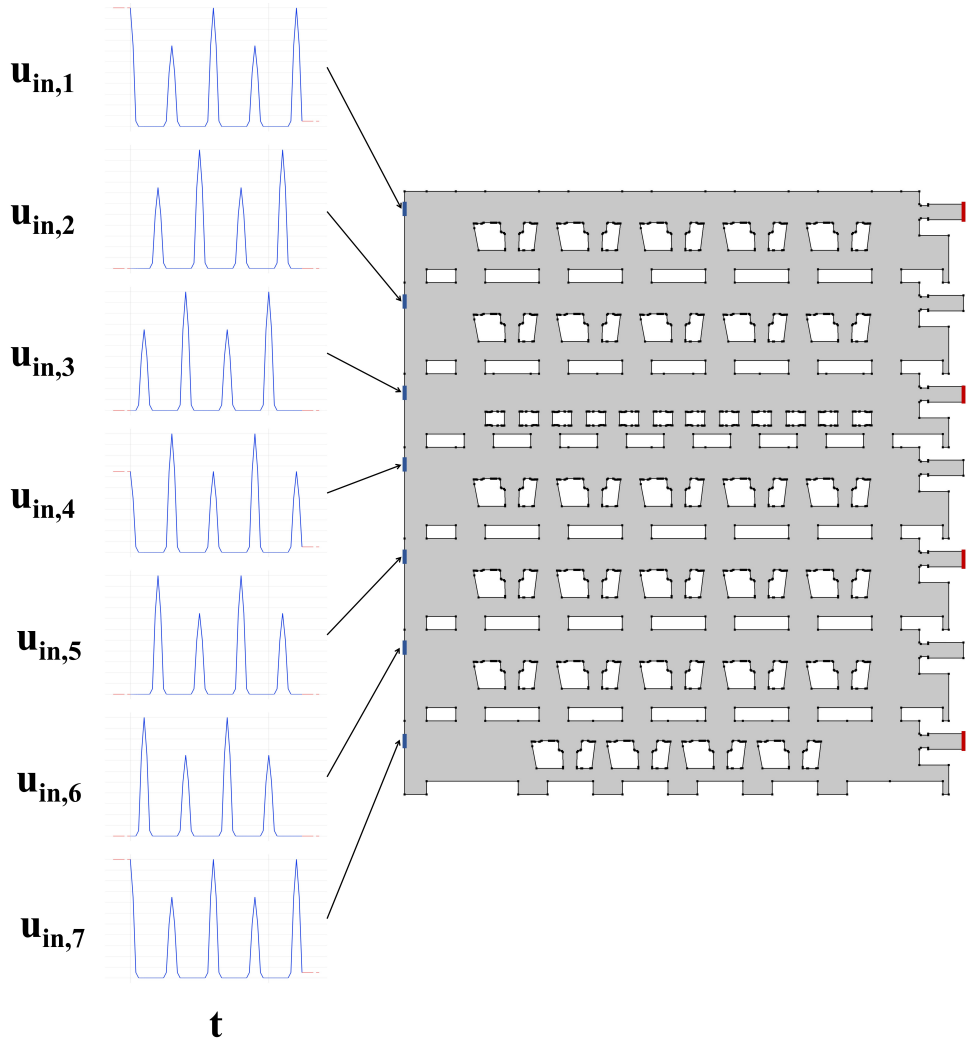


Figure 3.4: Inlets (highlighted in blue) and applied pulse boundary conditions for each one of them. The selected outlet boundaries are presented in red.

For the 7-disk geometry, the inlets along with the pulse boundary conditions applied to them are shown in Fig. 3.4. It should also be noted that the chemical species' concentrations at the inlet are calculated using the volumetric percentages found in Section 3.2.1 and the species' molar fractions at the inlet are set to 0.0385 for CO_2, 0.0169 for $AlCl_3$, 0.0210 for HCl, 10^{-6} for H_2O and CO, 0.9203 for H_2 and 0.0033 for H_2S.

In order to reflect the actual geometry (cf. Fig. 3.2), where the outlet perforations are not

aligned, the prescribed outlet pressure boundary conditions are applied at every other disk level. This means that out of the seven available outlets, only the first, the third, the fifth and the seventh from the top are considered open (marked in red in Fig. 3.4).

3.3.4 Chemistry model - Modeling the α-Al_2O_3 deposition

Several authors have studied the CVD of α-Al_2O_3 from a mixture of $AlCl_3$–CO_2–HCl–H_2–H_2S; for reasons of completeness, a brief overview is presented. The deposition appears to take place due to the hydrolysis of $AlCl_3$ in the presence of H_2O via the following surface reaction [149], [156], [157]:

$$2AlCl_{3(g)} + 3H_2O_{(g)} \rightarrow Al_2O_{3(s)} + 6HCl_{(g)} \tag{3.1}$$

while H_2O is produced in situ in the gas phase via the water-gas shift volumetric reaction [73]:

$$H_{2(g)} + CO_{2(g)} \rightarrow H_2O_{(g)} + CO_{(g)} \tag{3.2}$$

It must be noted, that this direction of the water gas shift reaction is endothermic [158], however, due to the small amount of CO_2 in the gas-phase, we expect no changes in the isothermal profile of the reactor. Another assumption is that the consumption of precursor does not affect the flow, which is reasonable due to its low concentration in the gas phase.

Although the work of Catoire and Swihart [157] highlights the complex mechanisms of the deposition kinetics, implementing this chemistry model that consists of 104 reactions and involves 35 species would inflate the computational cost of the CFD model. Given that the computational geometry is already a discounted representation of the actual reactor, it makes sense to implement the effective kinetics proposed in the work of Schierling *et al.* [156]. In their work, Schierling *et al.* [156] propose a simple reaction scheme, consisting of four reactions with two possible intermediate species in the gas phase, namely $AlCl_2OH$ and AlOCl. The detailed suggested reaction mechanism is the following:

$$AlCl_3 + H_2O \rightleftarrows AlCl_2OH + HCl \tag{3.3}$$

$$AlCl_2OH \rightleftarrows AlOCl + HCl \tag{3.4}$$

$$2AlOCl \rightleftarrows Al_2O_2Cl_2 \tag{3.5}$$

$$Al_2O_2Cl_2 + H_2O \rightleftarrows Al_2O_3 + 2HCl \tag{3.6}$$

According to the authors, the second step (eq. 3.4) is the rate-limiting reaction for the surface reaction, while the first step (eq. 3.3) is in the state of equilibrium. Based on this reaction mechanism, the first suggested deposition rate (R_{dep_1}, eq. 3.7) is derived. The authors then proposed a second empirical rate (R_{dep_2}, eq. 3.8) for an assumed parallel reaction path, with the aim of closely reproducing their experimental data. However, the rate remains empirical since the authors were not successful in searching for a second or third possible reaction sequence. Ultimately, the sum of these two deposition rates ($R_{\text{dep}_1} + R_{\text{dep}_2}$) makes up the total α-Al_2O_3 deposition rate.

$$R_{\text{dep}_1} = k_1 \cdot p_{AlCl_3} \cdot p_{H_2O} \cdot p_{HCl}^{-1} \quad (\text{mol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}) \tag{3.7}$$

$$R_{\text{dep}_2} = k_2 \cdot p_{AlCl_3}^{0.7} \cdot p_{CO_2}^{0.25} \cdot p_{H_2}^{0.2} \cdot p_{HCl}^{-1} \quad (\text{mol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}) \tag{3.8}$$

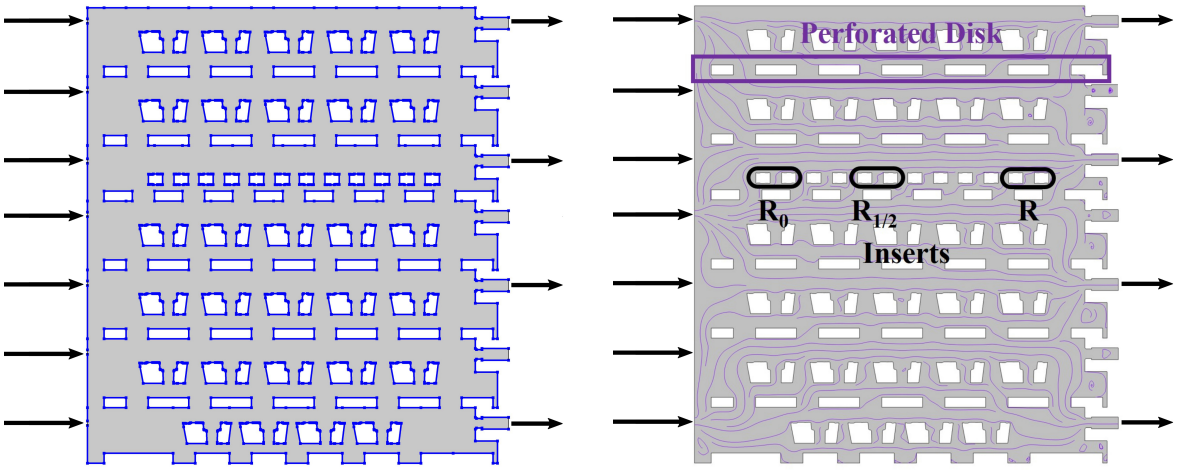
where p_i denotes the partial pressure of each reactant i. The kinetic rate for the water gas shift reaction (eq. 3.2) for a temperature of 1005°C is calculated through equation 3.9 [72]:

$$R_{\text{wgs}} = \frac{d[CO]}{dt} = k_{\text{wgs}} \cdot e^{-E_a/(RT)} [H_2]^{0.5} [CO_2] \quad (\text{mol} \cdot \text{m}^{-3} \cdot \text{s}^{-1}) \tag{3.9}$$

where units in brackets denote the concentration of each reactant in $\text{mol} \cdot \text{m}^{-3}$, the pre-exponential factor (k_{wgs}) is equal to $1.2 \cdot 10^{16} \text{m}^{1.5} \cdot \text{mol}^{-0.5} \cdot \text{s}^{-1}$ and the activation energy (E_a) is equal to $326.36\,\text{kJ} \cdot \text{mol}^{-1}$.

The homogeneous water-gas shift reaction (eq. 3.2) takes place in the domain of the simulation as indicated in Fig. 3.5a. Following experimental evidence, α-Al_2O_3 deposition (eq. 3.1) is considered to take place on all interior surfaces of the reactor, including the reactor walls, the inserts and the disks on which the inserts are placed. The only surfaces excluded are the reactor's inlets and outlets. A visual representation for the boundaries selected for the deposition can be observed in Fig. 3.5a. The α-Al_2O_3 deposition kinetic rate constants, k_1 (eq. 3.7) and k_2 (eq.

3.8) are fitted based on production coating growth data. Due to the lack of production data for different reaction temperatures, it is not possible to fit both a pre-exponential factor ($k_{0,i}$) and an activation energy ($E_{a,i}$) for each deposition rate. Therefore, the entire deposition kinetic constants ($k_i = k_{0,i} exp(-E_{a,i}/RT)$) are fitted all at once. For the WGS reaction, the pre-exponential factor is modified during fitting. However, no modification of the activation energy takes place.



(a) (b)

Figure 3.5: (a) Surfaces where α-Al_2O_3 deposition takes place are shown in purple; the volumetric Water Gas Shift reaction takes place in the area shown in gray. (b) Examples of the 2D representation of the inserts and perforated disks, in the computational domain. The arrows indicate the gas reactant inlets and outlets.

Given the gas flow and reactant concentration profiles inside the reactor, the α-Al_2O_3 deposition (h_{dep}) for the entire production time is given by eq. 3.10, integrating the deposition rates on the deposition boundaries for each insert of interest over the simulated 60s of the deposition process. The result of this integration is the deposition (in $\text{mol} \cdot \text{m}^{-2}$) that took place in the 60 simulated seconds (or 1 minute) of the process. By multiplying this result by the ratio of ($M_{Al_2O_3}/\rho_{Al_2O_3}$) we obtain the deposition thickness (in m) for the simulated 60s of the process. This result is then multiplied by the duration of the deposition process in minutes (t_{dep} - in this implementation, 3h), in order to calculate the deposition thickness for the entire process duration.

$$h_{\text{dep}} = t_{\text{dep}} \frac{M_{Al_2O_3}}{\rho_{Al_2O_3}} \int_{0s}^{60s} (R_{\text{dep}_1} + R_{\text{dep}_2})\, dt \quad (\text{m}) \tag{3.10}$$

where $M_{Al_2O_3}$ and $\rho_{Al_2O_3}$ denote the molecular mass and density of the produced α-Al_2O_3 coating. The molecular mass of α-Al_2O_3 ($M_{Al_2O_3}$) is 101.96 $\mathrm{g \cdot mol^{-1}}$ and the value of density at 1005°C is taken from Munro [159] and is equal to 3891 $\mathrm{kg \cdot m^{-3}}$.

The implementation of the kinetic constants proposed in Tingey [72] for the Water Gas Shift reaction, results on under-prediction of the overall coating deposition, attributed to low water availability. This motivated further investigation into the mechanisms that contribute to the in-situ production of water. Based on the more complex reaction scheme given by Catoire and Swihart [157], the WGS reaction is not the only water-producing reaction. In fact, three different pathways (including the WGS reaction) are responsible for the production of water inside the reactor. All three pathways are able to form water in comparable amounts and are therefore considered competitive and coupled. The authors also suggest that the AlOCl intermediate plays a vital role in one of the aforementioned water production channels. In the publication of Tan *et al.* [160], the effect of the AlOCl intermediate in water production is also acknowledged. However, the authors identified this effect as a catalytic effect on the Water Gas Shift reaction. Based on these previous findings the rate-constant of the Water Gas Shift reaction is fitted to capture the thickness measurements available in the production data.

Finally, Blomqvist *et al.* [148] investigated the effect of H_2S in the deposition of alumina under a chemical system similar to the present one. Although the H_2S appears to have minimal to non-existent effect on the Water Gas Shift reaction in the gas phase, the authors claim that H_2S as a true catalyst on the surface of α-Al_2O_3. However, the true mechanism of this effect is still obscure. For this reason, we also propose a modified version of equation 3.7, which - if given production data for different Hydrogen Sulphide inlet concentrations - could allow for the future investigation of the effect of H_2S in the process. The modified reaction rate equation is (eq. 3.11).

$$R'_{\mathrm{dep}_1} = k'_1 \cdot p_{H_2S} \cdot p_{AlCl_3} \cdot p_{H_2O} \cdot p_{HCl}^{-1} \quad (\mathrm{mol \cdot m^{-2} \cdot s^{-1}}) \tag{3.11}$$

3.4 Results

Since proprietary industrial production data are used for model validation, absolute thickness and deposition rate values cannot be presented. Therefore, only relative values are presented. Two

main metrics are given, considering the predicted ($y_{\text{prediction}}$) and the actual (y_{actual}) deposition thickness values:

1. The relative error (RE), which is calculated by the following formula:

$$RE = \frac{y_{\text{prediction}} - y_{\text{actual}}}{y_{\text{actual}}} \tag{3.12}$$

2. The mean absolute percentage error (MAPE), which is calculated for each geometry by averaging the N absolute values (in our case, $N = 3$) of the relative error per reactor geometry.

$$MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_{\text{prediction,i}} - y_{\text{actual,i}}}{y_{\text{actual,i}}} \right| \tag{3.13}$$

3.4.1 Parameter fitting and model validation

After conducting a mesh independence study for meshes consisting of 129195, 182609 and 287109 elements, a mesh of 129195 elements was used for the discretization of the combined gas flow / α-Al_2O_3 deposition problem. This resulted in a problem consisting of about 10^6 degrees of freedom. Solution time was approximately 3.5 core hours on an 11th Gen Intel(R) Core(TM) i7-1185G7 processor. When compared with the resources required for the 2D, full reactor model ($5 \cdot 10^6$, 66 core hours solution time), an important difference in the required resources can be observed.

An important challenge for this application is the fact that there are no CFD results reported in the literature. To our knowledge, this is the first attempt and therefore the model can only be validated using the available production data. For this reason, four different 7-disk parts of the same reactor are simulated: Geometry A is used for fitting the kinetic parameters of the chemistry model; Geometries B, C, and D (Fig. 3.6) are used for the validation of the model in set-ups, i.e combinations of disks and inserts, representing different parts of the same reactor, where the flow and species concentration distributions are not expected to be the same as in Geometry A. The four geometries are determined by several factors, such as the shape and size of inserts to be coated and the geometry of the disks that carry each type of inserts (each insert has a specific disk geometry). The latter means that the perforations of the disks have a different diameter and the number of inserts in each disk is different, affecting in this way the overall surface area at each

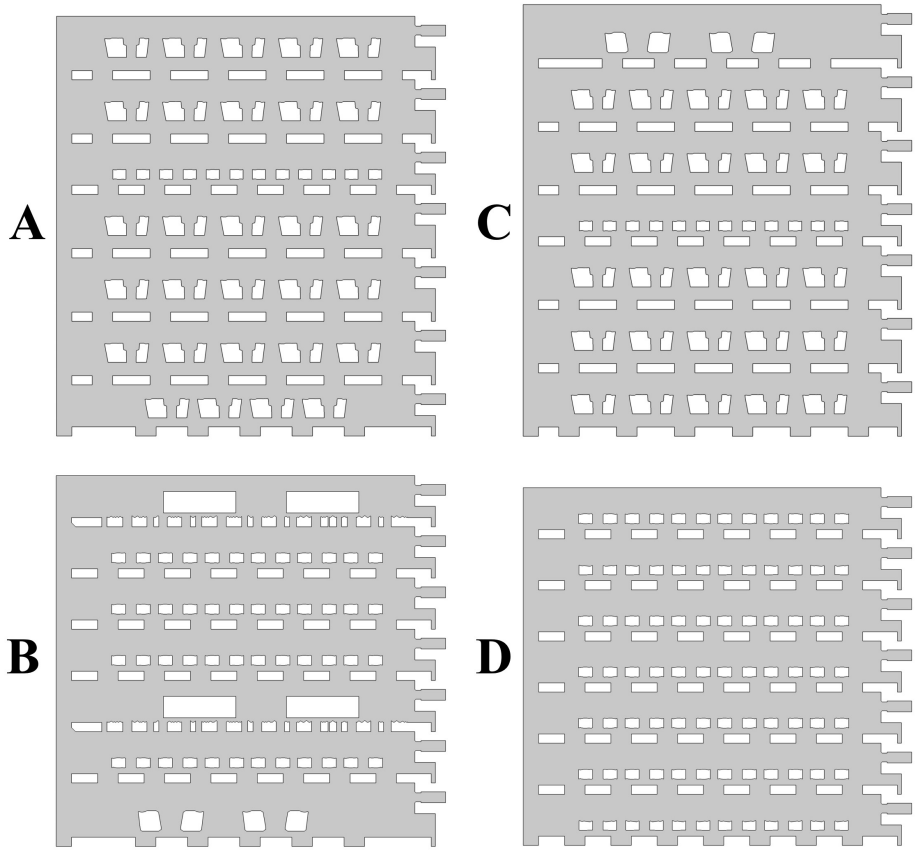


Figure 3.6: The four 7-disk geometries used in the CFD simulations. All cases are different parts of the same reactor.

disk.

The kinetic parameters are adjusted so that the difference between the predicted and the production deposition thickness values is minimized. The comparison between the production and predicted thickness values is done in three different positions (cf. Fig. 3.5b): R_0 which is closest to the inlet, $R_{1/2}$ which is mid-distance between inlet and outlet and R, close to the outlet.

The first step towards fitting the kinetic parameters was choosing initial values. The initial value for the kinetic constant of the Water-gas-shift reaction was taken from the publication of Bustamante *et al.* [73]. For the surface reaction kinetic constants k_1 (eq. 3.7) and k_2 (eq. 3.8), the initial values were set to 0.001 $\mathrm{s \cdot mol \cdot kg^{-1} \cdot m^{-1}}$ and 10^{-6} $\mathrm{mol \cdot m^{-1.85} \cdot s^{-0.7} \cdot kg^{-0.15}}$ respectively. These values led to a great underestimation of the coating thickness at all positions.

After trying different values for the kinetic parameters, it became clear that the reason for this severe underestimation was the value of the kinetic constant of the WGS reaction. This parameter was then fitted (as mentioned in section 3.3.4) to achieve coating thickness predictions in the same order of magnitude as the available production data. An increase of this parameter led to higher coating thickness overall. By making a tenfold increase in the WGS pre-exponential factor, we obtain results comparable to the production data, however, the deposition thickness at the position closest to the inlet is overestimated (RE: 33.8% @ R_0, 7.7% @ $R_{1/2}$, 13.2% @ R).

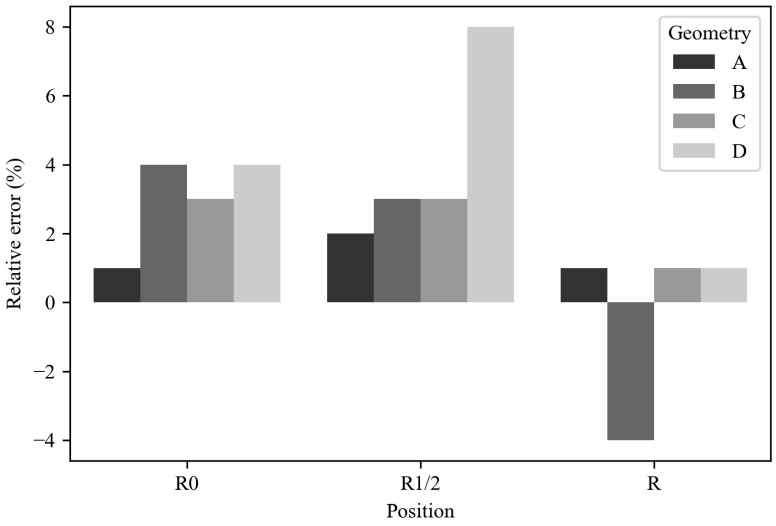
The next step was to reduce the overestimation of the deposition thickness at the position closest to the inlet (R_0). By observing the results of the simulations for different values of k_1, it was clear that this overestimation could be mended by selecting a lower value of the parameter. Therefore, when setting $k_1 = 3 \cdot 10^{-5} \mathrm{s} \cdot \mathrm{mol} \cdot \mathrm{kg}^{-1} \cdot \mathrm{m}^{-1}$ along with a nine-fold increase of the pre-exponential factor of the WGS reaction, slightly underestimating predictions are obtained (RE: -6.2% @ R_0, -4.2% @ $R_{1/2}$, -5.3% @ R).

After obtaining these results, the authors decided to make the transition from k_1 to k_1', trying to include in this way the concentration of H_2S into the α-Al_2O_3 deposition rate (via the proposed rate of eq. 3.11). Based on the average H_2S concentration inside the reactor and the value of k_1 that yielded the previous results, an initial value of $9 \cdot 10^{-7} \mathrm{s}^3 \cdot \mathrm{mol} \cdot \mathrm{kg}^{-2}$ was set for k_1'. This led to underestimation of the coating thickness (RE: -23.8% @ R_0, -20.5% @ $R_{1/2}$, -21.0% @ R). Increasing k_1' to $1.1 \cdot 10^{-6} \mathrm{s}^3 \cdot \mathrm{mol} \cdot \mathrm{kg}^{-2}$, only slightly amended this underestimation (RE: -20.6% @ R_0, -20.4% @ $R_{1/2}$, -20.8% @ R).

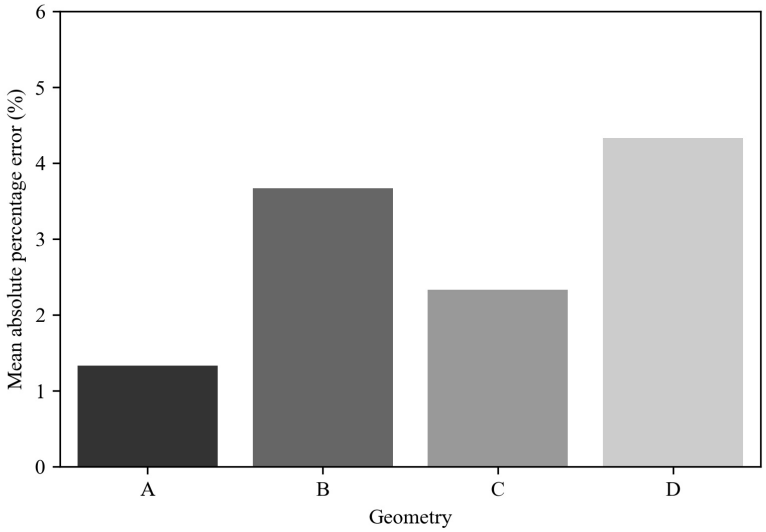
Having observed from previous simulation that an increase WGS reaction pre-exponential factor helps in reducing the underestimation of coating thickness at all positions, an eleven-fold increase was made. This yielded promising results (RE: -4.2% @ R_0, -3.4% @ $R_{1/2}$, -4.3% @ R). With some further fine-tuning, we ended up multiplying the pre-exponential factor of the WGS reaction (k_{wgs}) by a factor of 11.25.

The derived kinetic parameter values shown in Table 3.1, lead to prediction error of 2% at most in each one of the three positions (R_0, $R_{1/2}$ and R) in Geometry A.

The results, for Geometries B, C and D are presented in Fig. 3.7. Overall, the proposed CFD model predicts the actual thickness values within a 4% error (with the exception of the $R_{1/2}$



(a)



(b)

Figure 3.7: The developed model is tested for 4 distinct 7-disk geometries. The coating thickness predicted by CFD simulations is compared to production thickness values in three different positions (R_0, $R_{1/2}$ and R). Geometry A is used for calibration of the chemistry-enhanced CFD model, which is then tested in Geometries B, C and D. (a) Errors relative to the production data per geometry. (b) The mean absolute percentage error (MAPE) for each one of the four geometries. The highest error (observed for Geometry D) does not exceed 5%.

Table 3.1: The fitted kinetic constants used for the simulation of the α-Al_2O_3 deposition.

Parameter	Value	Units	Equation
k_{wgs}	$1.35 \cdot 10^{17}$	$m^{1.5} \cdot mol^{-0.5} \cdot s^{-1}$	3.9
k'_1	$1.1 \cdot 10^{-6}$	$s^3 \cdot mol \cdot kg^{-2}$	3.11
k_2	10^{-6}	$mol \cdot m^{-1.85} \cdot s^{-0.7} \cdot kg^{-0.15}$	3.8

position in Geometry D which has an error of 8%). In terms of the mean absolute error, during fitting it is 1.33% in Geometry A while the highest value is in Geometry D (4.33%). For Geometries B and C, the mean absolute percentage error is 3.67% and 2.33% respectively (Fig. 3.7b).

3.4.2 Investigation of the rate-limiting mechanism

The CFD simulation, allows us to take a closer look at the actual concentration distributions of the reactants, namely of the precursor ($AlCl_3$) and water, in the 4 geometries studied (cf. Fig. 3.6).

When considering the concentration of water (cf. Fig. 3.8), the CFD model predicts almost uniform distribution above the inserts in the disk where the coating thickness is predicted. Some regions of high water concentration are predicted, however they are not located above the inserts. On the contrary, the $AlCl_3$ concentration consistently appears to be higher closer to the inlet of the reactor (cf. Fig. 3.9). This imbalance is not reflected in the thickness of the deposited material either in the simulations or, in fact, in the production data. This observation motivates further investigation into the balance between mass transfer (diffusion) and the reaction kinetics, that ultimately determines the rate-limiting step of the process.

Typically, this study requires altering the temperature and monitoring the change in the deposition rate. For increasing temperature the deposition rate also increases following a linear trend, which is an indication that the reaction rate is the limiting step that determines the overall deposition rate. Past a certain temperature, the deposition rate becomes insensitive to further increase of the temperature, which is a sign that the rate of diffusion of the species on the surface determines the overall deposition rate. This process is typically described in a so-called Arrhenius plot, i.e. the plot of the deposition rate versus the inverse of temperature [161].

In the application studied here, where the data are derived from the production process at a

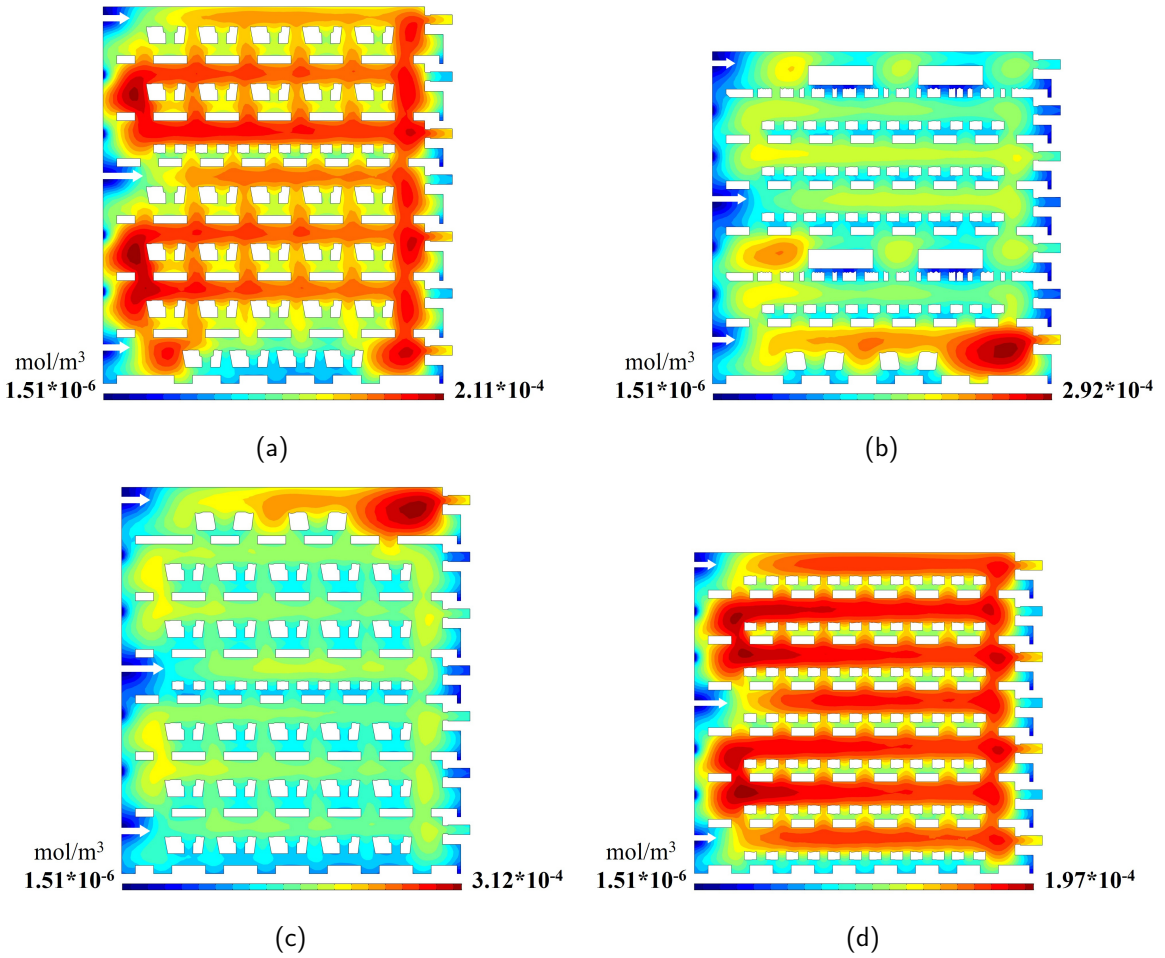


Figure 3.8: Contours of the concentration of H_2O for (a) Geometry A, (b) Geometry B, (c) Geometry C, (d) Geometry D. The white arrows indicate the velocity at each inlet. The CFD results suggest a mostly uniform concentration of water above the inserts.

single temperature (1005 °C), it is not possible to derive an Arrhenius plot based on which to define whether the process is kinetics or diffusion limited. Instead, it is still possible to gain insight into this balance with the proposed CFD model by means of studying the effect of the precursor mass fraction on the coating thickness: two computational experiments are performed, based on Geometry A, one with significantly increased precursor mole fraction at the inlet (by 25%), the second with significantly decreased (by 25%).

If the process were diffusion limited, then the reactions would be very fast and as soon as the precursor molecules reach the surface, they would react forming more α-Al_2O_3 on the sur-

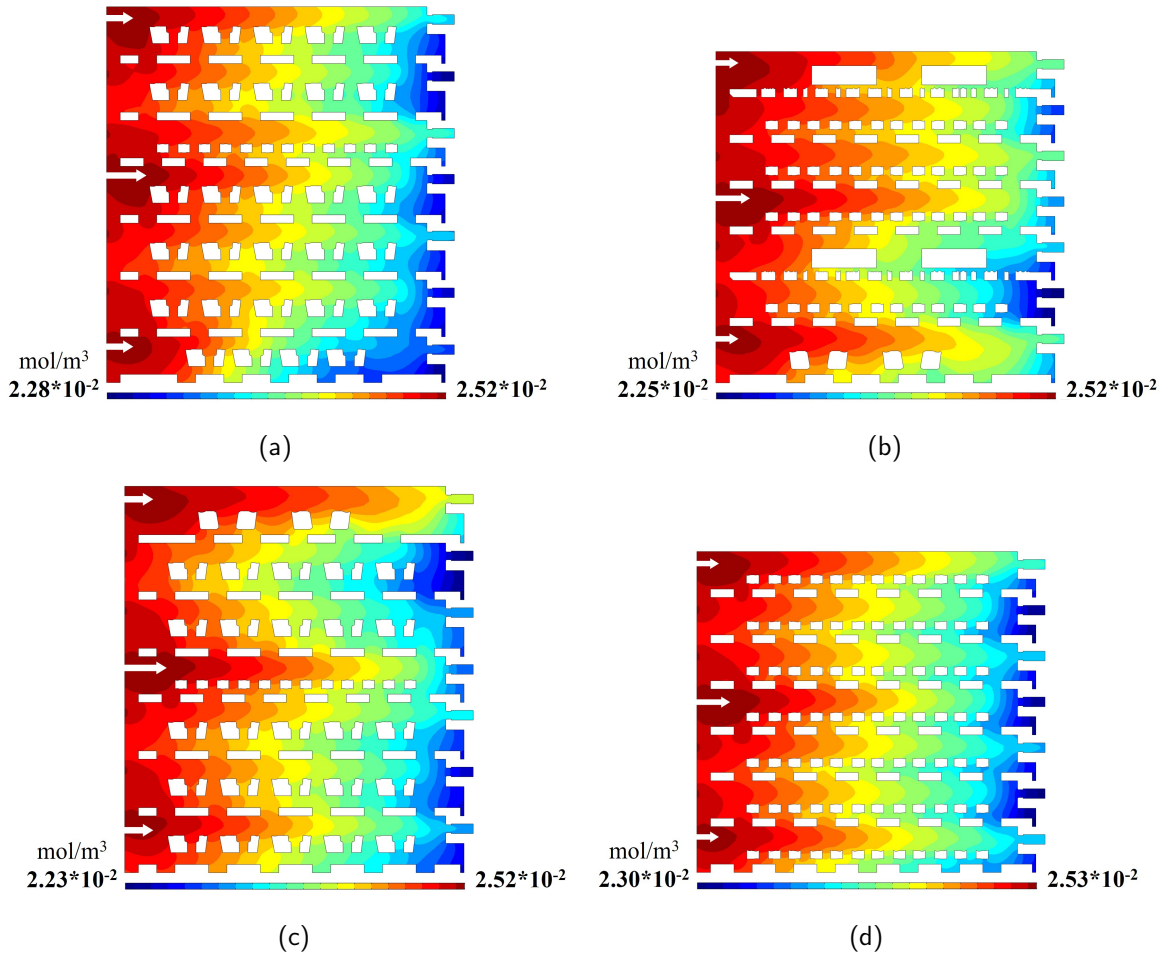


Figure 3.9: Contours of the concentration of $AlCl_3$ for (a) Geometry A, (b) Geometry B, (c) Geometry C, (d) Geometry D. The white arrows indicate the velocity at each inlet. The CFD results suggest highest precursor concentration close to the inlets.

face. Therefore an increase/decrease is expected as an outcome when the $AlCl_3$ mole fraction is increased/decreased respectively. On the contrary, if the deposition rate is affected to a negligible extent, then this would be a valid indication that the process is in the kinetics-limited regime. This comparison is shown in Table 3.2 where the "original" experiment, corresponding to the process conditions in Geometry A, is compared to the CFD results obtained in the same Geometry and conditions with different mole fractions of precursor at the inlet.

The results indicate that the change in the $AlCl_3$ mole fraction leads to negligible fluctuations in the calculated deposition rate for the insert at the $R_{1/2}$ and R position, which corroborates

Table 3.2: Difference in the deposition rate for different precursor concentrations at the inlet - Geometry A. Values are relative to the original simulation.

Position	R_0	$R_{1/2}$	R	$AlCl_3$ inlet mol. frac.
Relative difference (%)	4.5	-0.4	-0.1	1.25×Original
	-7.3	-0.3	-0.1	0.75×Original

the hypothesis that the process is in a kinetics-limited regime. Slight discrepancies appear in the insert at the R_0 position. Specifically, for the experiment with 25% increased precursor mole fraction, there is a 4.5% increase in the deposition rate for the insert at R_0, when compared to the original run. For the experiments with a precursor mole fraction of 25% less than the original, the calculated α-Al_2O_3 deposition rates for the insert at R_0 are 7.3% less than the original experiment. This finding is not contrary to the hypothesis of a kinetics-limited regime and can be explained by considering the concentration of the other reactant, water, in the region above the inserts on interest, summarized in Table 3.3. In the case of increased $AlCl_3$ mole fraction, the concentration of water is higher above the insert in the R_0 position, leading to higher deposition thickness. In contrast, in the case of decreased precursor mole fraction, water concentration is lower above the insert in the R_0 position, leading to decreased deposition rate. Overall though it could be argued that this discrepancy in the R_0 position of 4.5% increase and 7.3% decrease in the deposition rate can still be considered minor, taking into consideration that the alteration to the precursor concentration is by 25%.

Table 3.3: Average H_2O concentrations above the inserts of interest for different inlet precursor concentrations - Geometry A.

Position	R_0	$R_{1/2}$	R	$AlCl_3$ inlet mol. frac.
H_2O conc. $(10^{-4}\mathrm{mol}\cdot\mathrm{m}^{-3})$	1.92	1.95	1.87	Original
	1.66	1.59	1.52	1.25·Original
	2.30	2.53	2.47	0.75·Original

3.5 Conclusions

This work presents an efficient tool for computational analysis of an industrial-scale CVD reactor used for the coating of cutting tool inserts. The proposed CFD model addresses three significant

challenges not only in Chemical Vapor Deposition but also in other processes where chemistry and transport phenomena co-exist: (i) Complex geometries, (ii) Complicated networks of chemical reaction which are not completely known, (iii) Competition between the physical and chemical mechanisms, something that ultimately defines the rate of the overall process.

We demonstrated how this computer-aided approach can predict the thickness of the deposited film with noteworthy accuracy (with a 5% average error). To do so, we implemented a chemistry model that with one homogeneous and one heterogeneous reaction, for the sake of efficiency, which nevertheless takes into account the concentration of hydrogen sulphide. The latter is generally understood to act as a catalyst but to this date there is no consensus on the actual mechanism.

Despite the simplifications introduced for the sake of economizing on the computational effort, the proposed model is still able to illuminate important aspects of the interplay of physical phenomena (mass transport through diffusion) and chemical reaction rates. Results for higher and lower precursor concentrations in the inlet, point to the fact that the process is in the kinetics-limited regime, where the overall deposition rate in determined by the relatively slow reaction rate. Although further investigation is required to determine this fact with certainty, the input of the proposed model is still a useful "hint" to the direction that should be followed experimentally.

Chapter 4

Comparison of equation-based and data-driven modeling strategies for industrial coating processes

This chapter is reproduced from P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "Equation-based and data-driven modeling strategies for industrial coating processes," *Computers in Industry*, vol. 149, p. 103 938, 2023. DOI: 10.1016/j.compind.2023.103938. As the first author of the article, I participated in the development of the proposed CFD and ML models, the analysis of the resulting data, the validation and visualization of the results, and the writing of the original draft of the manuscript. The CRediT contribution statement for the publication is presented below.

CRediT authorship contribution statement

Paris Papavasileiou: Creation of models, Application of mathematical and computational techniques to analyze and synthesize data, Writing the manuscript. **Eleni D. Koronaki**: Development and design of methodology, Methodology, Supervision, Writing the manuscript. **Gabriele Pozzetti**: Supervision, Validation of the results, Providing resources (data). **Martin Kathrein**: Supervision, Validation of the results, Providing resources (data). **Christoph Czettl**: Supervision, Validation of the results, Providing resources (data). **Andreas G. Boudouvis**: Supervision, Formulation of overarching research goals and aims. **Stéphane P.A. Bordas**: Supervision, For-

mulation of overarching research goals and aims.

Abstract

An efficient CFD model for the deposition of alumina from a gas mixture consisting of $AlCl_3$, CO_2, HCl, H_2 and H_2S in an industrial CVD reactor with multiple disks and a rotating feeding tube, is proposed. The goal is twofold: (i) to predict the thickness of the deposited material, (ii) to investigate whether the process rate is determined by the reaction rate or by diffusion. A reaction model that consists of a gas-phase homogeneous reaction and a heterogeneous reaction is implemented, with a proposed kinetics rate that includes the effect of the H_2S concentration. The latter has a catalytic effect, but the mechanism is not entirely understood. The entire reactor geometry (consisting of 40-50 perforated disks) is divided into appropriately chosen 7-disk sections. The 2D, time-dependent CFD model is validated using production data for the deposition thickness. The proposed computational tool delivers accurate predictions (average relative error 5%) for different geometries corresponding to real reactor set-ups. Extending the functionality beyond prediction, a computational experiment is performed to illuminate the interplay between species diffusion and chemical reaction rates, which determines the rate-limiting mechanism. The results indicate that species diffusion is fast enough and therefore reaction kinetics determine the overall deposition rate.

4.1 Introduction

Chemical Vapor Deposition (CVD) processes are popular in a wide range of applications, including microelectronics [57], sensors [134] and wear resistant coatings [55]. The coating process involves the nonlinear interplay of physical mechanisms, such as diffusion and convection, with a plethora of homogeneous and heterogeneous chemical reactions. The competition between the different mechanisms determines the process outcome and the product quality. It is therefore a fine example of a process that is too complicated to study with first-principles models, such as Computational Fluid Dynamics (CFD) and where the data is often not enough to implement sophisticated data-driven strategies. Taking all of the above into consideration, the objective of this work is to investigate the potential benefit of simplified CFD process models, accompanied by purely data-driven predictions using Machine Learning (ML) approaches. Both methods are driven by the size

and type of the available production data. In the CFD case, the data are used for calibration and validation and in the ML case for regression.

Computational Fluid Dynamics is a valuable tool for studying deposition processes [7], [32], [138], [150], [161]–[163], since it allows the investigation of the flow field inside the reactor, as well as the main physical and chemical pathways that lead to the deposition of thin film coatings. Nevertheless, modeling industrial-scale deposition applications using CFD presents several challenges: Firstly, dealing with the complexity of the process, which often has several unknowns and secondly, the large scale of real applications.

Specifically, the actual chemical reactions that lead to deposition, including their rates, are often unknown. Therefore, it is not possible to predict the effect of the interplay between transport phenomena and chemical kinetics on the deposition rate, necessitating the development of a kinetic model [164]. Even when a chemical reaction scheme is available, some of its parameters may need to be fitted for the specific application. This parameter fitting involves an increased computational cost, as it usually requires numerous simulations [4], [8], [143], [151]. Nevertheless, CFD has been applied to several CVD applications, shedding light on previously "opaque" processes [5], [165], [166] while also allowing to predict their outcomes [2]. Although attempts have been made towards increasing the efficiency of CFD models by implementing reduced order modeling methods [8], [9], developing an efficient and accurate model in an industrial setting remains a challenging and time-consuming task.

In the era of Industry 4.0, digitalization has become one of the main drivers of innovation [167] and production data are becoming more and more available. The industry is trying to exploit this data, seeking improvement in several domains, including: maintenance management [168]–[171] quality management [15], [172]–[175], production planning and control [21], [23], [176]–[178], supply chain management [26], process outcome predictions [179]–[182] and process optimization [183], [184]. Furthermore, digital twins [185] are becoming increasingly popular in the process industry [186]–[188], as well as in other, diverse applications [189], [190]. Although the application of sophisticated methods such as Deep Neural Networks (DNNs) [191], [192], Physics Informed Neural Networks (PINNs) [193] and manifold learning [194] has been demonstrated on controlled small scale problems, several challenges still remain when incorporating ML in everyday industrial

practice. Addressing these challenges is one of the main objectives of this work.

The industrial application in this work is the coating of cutting tools with α-Al_2O_3 for increased wear resistance. Concerning CFD, the goal is to propose the best possible simplified model, based on the available data which are necessary for verification and validation. This leads to a 2D, time-dependent CFD model, presented in detail in previous work [32]. The proposed model implements representative boundary conditions and employs a simple reaction scheme for the α-Al_2O_3 deposition with the goal of reducing the computational cost.

Concerning ML, the first task is to pre-treat the available data, upon which the choice of method depends on. Addressing mixed types of data (categorical and numerical) is a common challenge in many applications, not restricted to deposition processes. Several regression models are trained to predict the α-Al_2O_3 coating thickness using characteristics of the reactor set-up and process conditions as inputs. In this work, the focus lies more on tree-based methods [195] which are the best-performing for the given data-set.

The two approaches are initially compared in their ability to accurately and efficiently predict the alumina coating thickness of the cutting tool inserts. Specifically, the advantages and disadvantages of each strategy are assessed in terms of accuracy, interpretability, extrapolation ability and computational cost. As a final step, the two approaches are merged through the implementation of the Gappy Proper Orthogonal Decomposition (GappyPOD) method [35], [196]. The latter, is popular for optimal sensor placement, and here it adapted to propose a sufficient number of known data from which we can infer quantities that are not measurable.

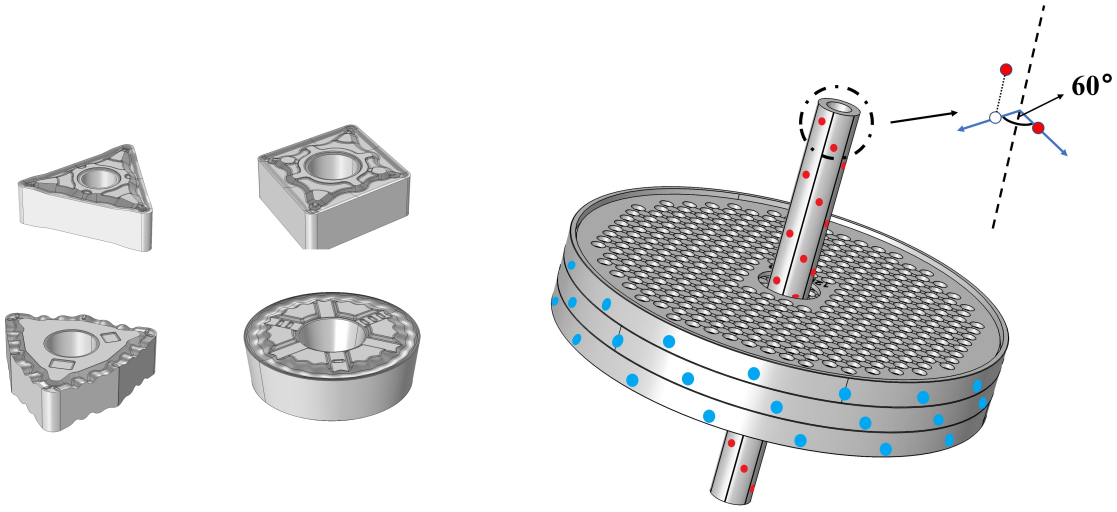
The manuscript is structured as follows: A concise overview of the process and the available production data is given in Section 4.2. The implemented methods (CFD, ML and GappyPOD) are presented in Sections 4.3 and 4.4. The results of each method are analyzed and compared in Section 4.5, followed by the conclusions in Section 4.6.

4.2 Process description

A two-step coating process takes place inside the studied industrial-scale, commercial CVD reactor (Sucotec SCT600TH). First, a Ti(C,N) base layer of about 9 µm is grown on the cemented carbide cutting inserts, such as the ones shown in Fig. 4.1a. Subsequently, an alumina layer is deposited

under a $AlCl_3$–CO_2–HCl–H_2–H_2S chemical system. The temperature and pressure for the alumina coating step are T=1005°C and p=80 mbar, respectively [1]. The alumina coating deposition step of the process takes approximately 3 hours.

The CVD reactor consists of 40-50 perforated disks, stacked one on top of the other, whereon the inserts are placed. In Fig. 4.1b, a schematic of three such disks is shown for clarity. The mixture of gas reactants, enters the reactor via perforations on a rotating cylindrical tube, placed in the center of the structure of the stacked disks. There are two antipodal perforations for each disk level. There is a 60° angle difference between the axis connecting the inlet holes for each disk level. The rotational motion of the inlet tube (rotating with a rotational speed of 2 RPM) causes the process to have an inherent periodic nature. The interior geometry of the reactor changes from production run to production run, since the geometry of the inserts (and the disks on which they are placed), changes based on production requirements.



(a) (b)

Figure 4.1: (a) Indicative geometries of the coated cutting tools. (b) A 3D representation of a 3-disk part of the reactor. The inlet perforations on the rotating inlet tube are shown in red. The outlet perforations for each disk are shown in blue.

The main goal of the process is to achieve uniform coating thickness, since this uniformity also leads to uniform product longevity [74]. Ideally, coating thickness uniformity would be achieved across all production runs, reactors, and production sites. However, this is not always the case.

For this reason, a way of predicting the coating thickness of the inserts given the reactor set-up is needed. Furthermore, coming up with a systematic way of assessing the factors that influence the coating thickness uniformity is also highly important.

4.2.1 Available data

For the Ti(C,N)/α-Al_2O_3 multi-layer coating, the thickness measurements are performed via the Calotest method. A small spherical cavity is ground on the coated inserts using a rotating ball of known geometry, providing a tapered cross-section of the film when viewed under an optical microscope [197]. This way, the thickness of both the Ti(C,N) and α-Al_2O_3 coating layers can be calculated. Measurements are usually taken for 3 positions on 5 disks of interest. Therefore, 15 thickness measurements are available for each production run. A 2D representation of the reactor indicating the points where thickness is typically measured is shown in Fig. 4.2. These measurements allow for not only for the calibration and validation of the CFD model, but also for several ML approaches.

Apart from coating thickness measurements, the dataset also contains several features concerning the process and the reactor setup, which will serve as inputs to the machine-learning model. The production "recipe" used for the coating is the available feature providing information regarding the process. Setup-wise, there is a plethora of available features for each disk of the reactor, including:

1. The position of each disk inside the reactor.

2. The number of inserts placed on each disk.

3. The type of insert placed on each disk. Each type of insert has different geometrical characteristics.

4. The type of disk used. The type of disk used is always relative to the type of insert placed on top of it.

5. The surface area of the inserts placed on the disk.

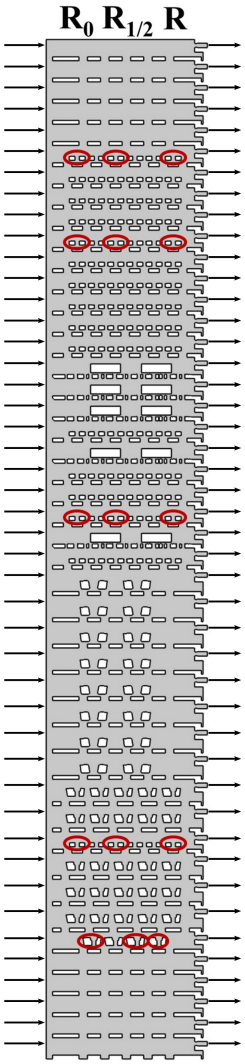


Figure 4.2: Positions with available α-Al_2O_3 thickness values from the production data for our test case. In general, across different production runs, the R position (the one closest to the reactor outlet) is the one with the highest amount of data. For this reason, the ML models are trained to make predictions for inserts placed in this position. The arrows indicate the gas reactant inlets and outlets.

These features allow for the creation of more features, such as the total surface area and the standard deviation of the surface area of the inserts that are coated inside the reactor. Another feature that can be created is the difference between the nominal surface area of the production "recipe" and the actual insert surface area inside the reactor. Furthermore, for each disk, we can exploit the information available for its neighboring disks.

This way, we end up with several features, of which thirteen are used as inputs after being pre-processed. These features are summarized in Table 4.1. Considering the coating thickness measurements as outputs, we can train several supervised learning models to make coating thickness predictions per disk. In this context, during training, a labeled set of inputs is provided and specifically here, the inputs are the aforementioned features and the labels are the α-Al_2O_3 coating thickness measurements.

Table 4.1: Summary of the features included in the training of the regression models.

Feature	Type	Pre-processing
Number of inserts on disk	Numerical (integer)	standardization
Surface area of inserts on disk	Numerical (float)	standardization
Disk position	Numerical (integer)	standardization
Total surface area of inserts inside the reactor	Numerical (float)	standardization
Surface area standard deviation	Numerical (float)	standardization
|Nominal "recipe" surface area - actual surface area|	Numerical (float)	standardization
Production "recipe"	Categorical	binary encoding
Insert geometry	Categorical	binary encoding
Disk geometry	Categorical	binary encoding
Insert geometry – disk above	Categorical	binary encoding
Insert geometry – disk below	Categorical	binary encoding
Disk geometry – disk above	Categorical	binary encoding
Disk geometry – disk below	Categorical	binary encoding

4.3 Computational ingredients

4.3.1 ML methods

For the data-driven approach to the problem, the implementation of an assortment of machine learning methods for the prediction of coating thickness inside the reactor is investigated. All methods implemented fall into supervised learning methods.

In supervised learning, each one of the input variables x_i is associated with a response (or output) y_i [81]. The goal of the ML strategy is to train a model able to relate the input variables x_i to the output y_i. This way, future observations can be predicted and the relationship between the inputs and the output can be interpreted. Here, the goal is to predict the α-Al_2O_3 coating thickness (a continuous target variable) from several inputs, using a regression method. The specific methods include but are not limited to:

- Linear methods, such as linear, lasso or ridge regression.
- Non-linear methods, such as polynomial regression.
- Tree-based methods, such as regression trees and their ensemble versions: random forests, gradient boosted regression trees and extreme gradient boosted regression trees.
- Artificial neural networks.

During the early phases of this research, several techniques were utilized, including linear, lasso, and ridge regression, as well as support vector machines and Gaussian process regression. Preliminary findings indicated that tree-based methods outperformed the other techniques, and as a result, the focus of this study is on tree-based methods.

The models' accuracy will be evaluated via two different metrics, namely the mean absolute error (MAE) and the mean absolute percentage error (MAPE). When the model is trained or tested on N observations and for each observation i the prediction is $\hat{y}_i$ while the actual value is y_i, MAE and MAPE can be written as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{4.1}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{4.2}$$

Two different computational costs pertain to each ML model, the training time (t_{train}) and the prediction time (t_{pred}) of the model. Both of these costs are expressed in CPU time.

Tree-based methods

Tree-based methods work by partitioning the space of the inputs X into a set of rectangles. Afterwards, a simple model (e.g. a constant) is fit in each partition. The process starts by splitting the entire input space in two based on a variable of the input space and its value. The optimal variable and split point are chosen in order to achieve an accurate fit. Then, either or both of the resulting regions are split again in two, once again using the optimal input and split point. This procedure continues until a stopping criterion has been met. The occurring binary splits allow for model interpretability since the entire sample space can be described by a single tree. Tree-based methods can be used for both regression and classification purposes [198].

The prediction accuracy of a single tree is often not as high as that of other methods. Furthermore, a small change in the data can lead to an entirely different tree layout. These two issues and especially the predictive performance of the trees can be rectified by combining multiple trees through the implementation of ensemble methods such as bagging and boosting [195].

The concept behind ensemble methods is to build a prediction model by combining a number of simpler base methods, in two steps: First, a number of base learners must be created from the available data. The second step involves the combination of these learners into one ensemble predictor. The most common ensemble tree-based methods are random forests, bagged trees and gradient boosted trees. These methods, however, have some key differences between them.

Random forests and bagged trees, discussed here, operate similarly. They both build B regression trees and each tree is trained using bootstrap-sampled (i.e. sample a particular data-point and then reintroduce it to the dataset), versions of the original dataset. Bagging regression methods provide a prediction by averaging the outputs of the B trees that they consist of. If $\hat{y}_{i,b}$ is the prediction of each grown tree, then the final prediction of the bagging method $\hat{y}_{i,bag}$ is given by:

$$\hat{y}_{i,bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_{i,b} \tag{4.3}$$

Random forests and bagged trees differ only in the amount of input features N_{input} that are considered when building each tree. In bagged trees, all available features are considered. On the contrary, in random forests, a random subset of p input features is considered. This serves the purpose of de-correlating the individual trees, since the trees are not always built by selecting the global optimal features, but by selecting the optimal feature from a randomly sampled subset of the input features [195].

Gradient boosting and extreme gradient boosting are boosting methods. In the case of boosting methods, contrary to bagging methods, the B base trees are created sequentially. First, the first tree of the ensemble is created. Afterwards, each created tree is fitted to the difference between the value predicted by the previous tree and the real output. This way, each tree improves the shortcomings of the previous one. There is no averaging of the result of the B trees in this case [199].

Therefore, after building the b^{th} tree which outputs γ_{jb} and is trained on the residual of the output of the ensemble after the previous tree has been built, the output of the ensemble $f_b(x)$ can be written as:

$$f_b(x) = f_{b-1}(x) + \lambda \cdot \sum_{j=1}^{J} \gamma_{jb} I(x \in R_{jm}) \tag{4.4}$$

where I is the indicator function, and λ is the learning rate of the boosting procedure. λ serves the purpose of scaling the contribution of the output of each tree to the final prediction of the ensemble.

The result of the model is the output of the ensemble after the final tree has been built. Boosting methods are more prone to overfitting for large values of B than bagging methods. For this reason, B needs to be carefully selected through cross-validation.

Challenges

Applying data-driven methods to a real-world dataset presents several challenges. First and foremost, the dataset needs to be "cleaned": Given that the production dataset is derived from different production sites, different reactors, and different people, it is bound to contain some errors. These errors must be identified and corrected before any type of analysis. Then, there is the question of the format of the data. Even when the data is neatly organized in an SQL database, it still needs to be extracted and formatted (using the pandas python library [200], for example) so that it can be used to train models in a python framework. Afterwards comes the question of data type. In this particular application, there are both numeric and alphanumeric features (features that contain names instead of values). Since several of the implemented methods are not compatible with alphanumeric (categorical) features, those features need to be encoded in a way (i.e. binary encoding, one-hot encoding [105]) that allows them to be used in our models. Finally, once the data is ready, the task is to find the best performing model and to determine the hyperparameters that influence performance. Therefore, a hyperparameter optimization step must also be included. By following this step-by-step approach, we can establish a data pipeline specific to our data that allows us to overcome all the aforementioned challenges. This however requires experience, input from the process experts, along with a clear understanding of the data.

4.3.2 CFD modeling: Implementation and challenges

For this specific application, a digital "replica" of the process would have to be a 3D, time-dependent full reactor (40-50 disks) model which would include a complex reaction scheme. A complex reaction scheme, would lead to more degrees of freedom and an increased number of kinetic parameters that would need to be fitted. Apart from this, given the rotation of the inlet tube (and therefore the fact that the problem is not axisymmetric) a moving mesh would also need to be implemented. This would translate into a computationally intractable task. If we consider that the reactor interior geometry changes on a day-to-day basis, since the geometries of inserts and the disks on which they are placed change based on production quotas, a computationally expensive model is not a suitable method to study this industrial application. For this reason, aiming to drive the computational cost down, the problem was approached as follows:

- The problem is modeled in 2D.

- The boundary conditions for both the inlet and the outlet are selected in a way that is representative of their 3D characteristics.

- The model takes into account only 7-disk parts of the reactor in a divide and conquer approach.

- A simpler reaction scheme that still leads to accurate results is used.

To efficiently tackle the challenges of the process, a 2D, time-dependent model that accounts for the transport of mass, momentum, and species inside the reactor is proposed. The COMSOL Multiphysics® software was used for the CFD modeling. The interested reader can seek detailed information in the recent work of Papavasileiou *et al.* [32]; here the key points are summarized for completeness.

A reaction scheme consisting of a homogeneous reaction in the gas phase and a heterogeneous reaction for the deposition of α-Al_2O_3 is part of the model. The following assumptions are made: a) laminar and incompressible flow, b) constant temperature of in the entire reactor domain, c) ideal gas phase. The CFD model accounts for 7-disk "building blocks" of the reactor, in order to keep the computational cost low. To account for the rotation of the inlet tube, pulse velocity boundary conditions are applied at the inlets. To represent the placement of the holes on the inlet tube in the 2D computational geometry, a phase difference is included between the boundary conditions of each disk. A similar approach is taken for the outlet perforations. Since they are not aligned, pressure boundary conditions are applied at every other disk (1st open, 2nd closed and so forth). In order to model the deposition of α-Al_2O_3 under the $AlCl_3$–CO_2–HCl–H_2–H_2S chemical system, we implement a simple reaction scheme based on the work of Schierling *et al.* [156]. Implementing this simpler scheme results in a lower computational cost. The simulations account for two full rotations (or periods) of the feeding tube.

4.4 Combining equation-based and data-driven approaches using GappyPOD

In this work, the GappyPOD method is used for the reconstruction of several 7-disk reactor snapshots acquired using the aforementioned CFD model using limited - or "gappy" data. GappyPOD was first introduced by Everson and Sirovich [196] and then implemented, among others, to a CFD airfoil application by Willcox [35] and for non-linear fracture mechanics modeling [201]. Optimal sensor placement is another problem that can be solved using the GappyPOD method, as indicated in the works of Willcox [35] and Jo *et al.* [202]. This is achieved by finding the optimal way of filling the "gaps" in the data, or in other words, selecting the sensor positions that give the most information possible.

A concise overview of the method, along with the procedure followed for the acquisition of data and the metrics used for the evaluation of the method, are presented in the following paragraphs.

4.4.1 Overview

In this section, the GappyPOD method is summarized for completeness. Let's consider a dataset $\mathbf{X}$ of M vectors (represented as d-dimensional real vectors $x_1, \ldots, x_M$). A POD basis, $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$, of $\mathbf{X}$ is computed, such that $\mathbf{X}$ can be approximated as a linear combination of p vectors:

$$\widetilde{\mathbf{X}} = \sum_{j=1}^{p} c^j \mathbf{\Phi}^j \tag{4.5}$$

or in matrix-vector format:

$$\widetilde{\mathbf{X}} = \mathbf{\Phi} \cdot c \tag{4.6}$$

The size of the truncated POD basis $\mathbf{\Phi}$ is selected based on the error between the actual vector $\mathbf{X}$ and the reconstructed approximation $\tilde{\mathbf{X}}$:

$$\text{reconstruction error} = \|\mathbf{X} - \tilde{\mathbf{X}}\| \tag{4.7}$$

Another factor that can be taken into account when selecting the size of the truncated basis

is the total energy retained by the selected number of modes. For each basis vector j, the relative importance (E_j) is given by:

$$E_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \tag{4.8}$$

and therefore, the total energy retained for the k retained modes is given by:

$$E_{\text{total}} = \sum_{j=1}^{k} E_j \tag{4.9}$$

Let us consider a vector X' that is spanned by the same basis Φ and that only m values of this vector are known, such that the partial vector $X'_{partial}$ can be defined:

$$X'_{\text{partial}} = m \cdot X', m \in \mathbb{R}^{m \times N} \tag{4.10}$$

The goal is to find coefficients c', such that an approximation $\tilde{X}'$ of the vector X' can be defined as :

$$\tilde{X}' = X' \cdot c' \tag{4.11}$$

then:

$$X'_{\text{partial}} \approx m \cdot X' \cdot c' \tag{4.12}$$

Finding the values of c' that satisfy the above leads to an optimization problem, which results in the solution of the linear system:

$$M \cdot c' = (m \cdot \mathbf{\Phi})' \cdot X'_{\text{partial}} \tag{4.13}$$

with $M = (m \cdot \mathbf{\Phi})' \cdot (m \cdot \mathbf{\Phi})$

4.4.2 CFD data sampling

Snapshots, i.e. vectors containing information regarding the system's state at a specific time, of 12 different 7-disk reactor parts will be used for the implementation of the GappyPOD method. For each reactor part, there 31 available time-instances (each one with 1 second time difference from the previous). This way, the full dataset consists of 372 vectors.

At each time-instance, 4 quantities of interest are sampled along the lines connecting inlet-outlet at each disk level. The points of these lines are then interpolated at 250 specific query points using linear interpolation. In this manner, 250 evenly spaced points along each line are obtained. An example of the lines along which the quantities of interest are sampled is demonstrated in Fig. 4.3.

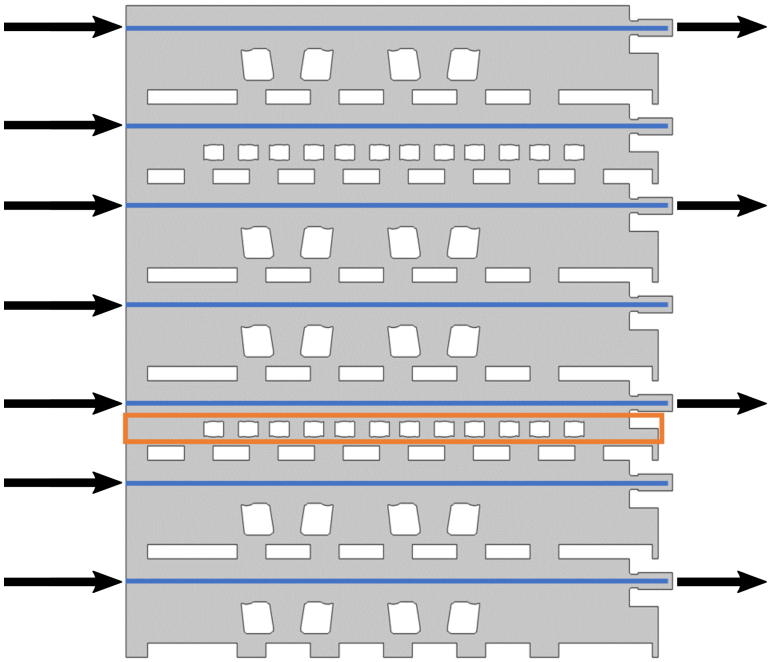


Figure 4.3: In blue: The seven lines along which the 4 quantities of interest $(U, p, C_{AlCl_3}, C_{H_2O})$ are sampled. In orange: The disk with available thickness measurements. The thickness measurements, as well as the α-Al_2O_3 deposition rates at the inserts of this disk, are also included for our implementation of GappyPOD. The arrows indicate the gas reactant inlets and outlets.

The quantities of interest at each point are:

1. The velocity magnitude (U).
2. The pressure (p).
3. The concentration of the precursor $AlCl_3$ (C_{AlCl_3}).
4. The concentration of water (C_{H_2O}).

Furthermore, the deposition rates as predicted by the CFD model along with the available thickness data for 3 positions (R_0, $R_{1/2}$, R) for each 7-disk reactor part, are included in each snapshot. An overview of the resulting dataset after sampling and organizing the vectors is presented in Fig. 4.4.

It is worth noting that a plethora of input parameters influences the final product, the most important of which include the configuration of the reactor's interior geometry and the production "recipe". The latter includes all the steps and chemical species involved in the production of a single coating layer. In this work, to make the simulations tractable, the focus lies on a single "recipe" for a single product and various geometries, without loss of generality.

4.4.3 Performance metrics

The performance of the GappyPOD approach will be evaluated using the Root Mean Squared Error (RMSE) between: a) the GappyPOD reconstruction and the POD reconstruction, b) the GappyPOD reconstruction and the snapshots of the reactor given by the CFD model. The RMSE between two values ($\hat{y}_i$ and y_i) for N observations can be written as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{4.14}$$

4.4.4 Mask selection

The effectiveness of GappyPOD depends on the condition number of matrix M, which is defined in Eq. (4.13). The matrix M is created from the inner products of the "gappy" POD vectors, which are the elements of the original POD vectors corresponding to the known elements of X'. Since these vectors are no longer orthogonal, the matrix M is fully populated. For orthogonality

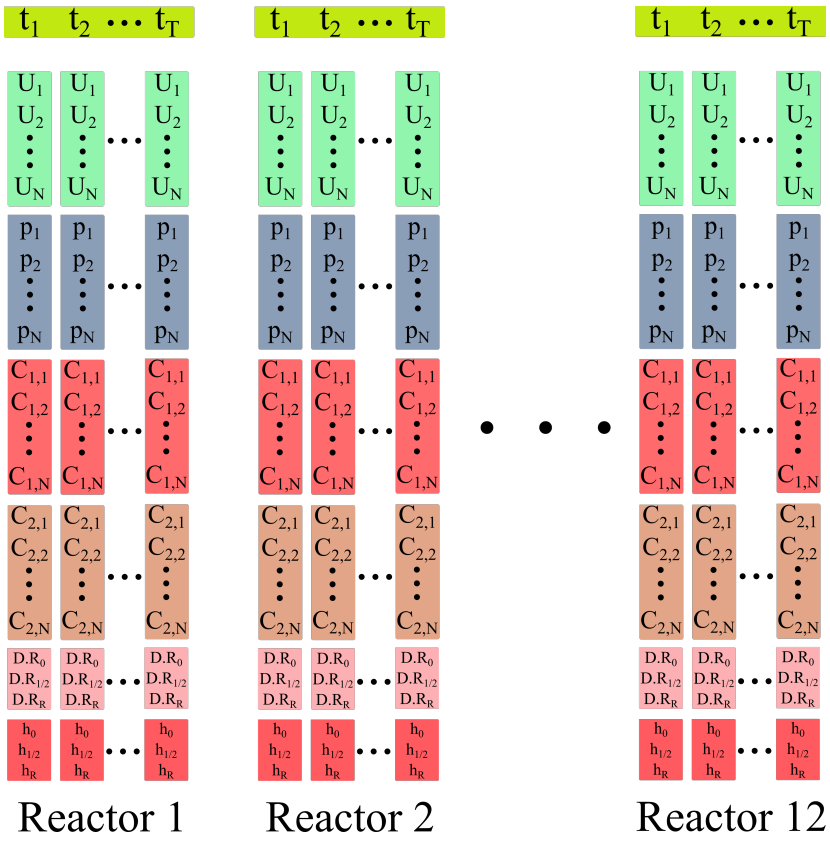


Figure 4.4: The final matrix considered for the GappyPOD method. A total of 31 time-instances for 12 different reactor geometries have been sampled. These contain all 4 quantities of interest (velocity magnitude, pressure, precursor concentration (C_1), water concentration (C_2) along with the calculated deposition rates (D.R) and the coating thickness measurements (h) taken from the production data. In our case, T $= 31$ (number of time-instances per reactor) and N $= 1750$ (total number of points: 7 lines containing 250 points each).

to be preserved, the known element positions and non-zero elements of M must be appropriately arranged. Additionally, the diagonal entries of M must not be too small, indicating that the POD basis element at that point should not be small. The condition number of the matrix M reflects these requirements, with a smaller condition number indicating greater satisfaction of these conditions. This analysis is detailed in [35], in the context of optimal sensor placement, and in [203], [204], which consider the angle between the measurement subspace and the low dimensional space that spans the data.

To determine the known values of the vector X' in a more systematic manner, a greedy

algorithm similar to the one proposed by Willcox [35] is implemented. However, in our case, the mask elements are selected in a way that reduces the reconstruction error. Considering m known values of each snapshot X', then the greedy algorithm implemented works as follows:

1. Initialize by randomly selecting m known values.
2. Starting with the first mask element, loop through all the possible positions for the known values and calculate the reconstruction error for each resulting mask.
3. Find the position of the element that minimizes the reconstruction error and place the first element there.
4. Repeat steps 2-3 for all remaining mask elements.

This way, we can efficiently find positions for the mask elements that yield an acceptable reconstruction error. It should be noted, however, that this does not always lead to the globally optimal positions.

4.5 Results

4.5.1 CFD model

CFD model parameters

To elaborate on the model summary made in Section 4.3.2, further information regarding the CFD model parameters is given in this section.

The prescribed inlet boundary conditions are inlet velocity conditions. For each disk, the gas feed velocity is a time-dependent pulse function that mirrors the inlet tube rotation, varying between 0 and V_{max}. There is a phase difference between the pulses of each disk. V_{max} and the aforementioned phase difference are determined based on the experimental conditions and geometry, taking into account: a) the 2 RPM rotational speed of the inlet tube, b) the total inlet gas flow rate, c) the number of disks per run, d) the two antipodal perforations per disk, e) the diameter of the perforations (0.002 m), and f) the 60° angle difference between the perforations of each disk.

Outlet pressure boundary conditions are applied at every other disk level. This way, we account for the real geometry where the outlet perforations are not aligned. This results in a model where only the first, the third, the fifth, and the seventh outlet from the top are considered open.

Seven different chemical species are considered, along with a simplified reaction scheme for the deposition of α-Al_2O_3. The molar fractions at the inlet are the following: CO_2 (0.0385), $AlCl_3$ (0.0169), HCl (0.0210), H_2O (10^{-6}), CO (10^{-6}), H_2 (0.9203), and H_2S (0.0033).

The process conditions for the alumina coating step are T=1005°C and p=80 mbar, as indicated in [1]. Further information can be found in the recent work of Papavasileiou *et al.* [32].

CFD model predictions

The CFD model has been tested for 4 different 7-disk reactor geometries. All four 7-disk geometries are building blocks of the test case reactor, whose 2D representation is shown in Fig. 4.2. It is possible to predict the α-Al_2O_3 coating thickness with a maximum relative error of 8% and within 5% mean absolute percentage error for each 7-disk geometry, when compared to the available production data. The maximum observed mean absolute percentage error for the α-Al_2O_3 coating thickness is 4.33%. Simulations for each geometry consist of about 10^6 degrees of freedom. The solution time for each geometry is approximately 3 core hours on an 11^{th} Gen Intel(R) Core(TM) i7-1185G7 processor. The results of the CFD simulations are summarized in Fig. 4.5.

4.5.2 Data-driven predictions

We implement the following tree-based methods: a) Regression Trees, b) Random Forests, c) Gradient Boosting Regression Trees (GBRT) and eXtreme Gradient Boosting Regression Trees (XGBoost). All the methods have comparable performance. Among them, the best performing is XGBoost and the results below focus on its predictions.

The dataset contains a total of 6114 observations and is split into a training set and a test set, using a ratio of 75/25. Each one of these observations contain thickness measurements at the R position for a particular disk (cf. Fig. 4.2), corresponding to a number of inputs, detailed in Section 4.2.1. The numerical features were standardized, and the categorical features were encoded using binary encoding.

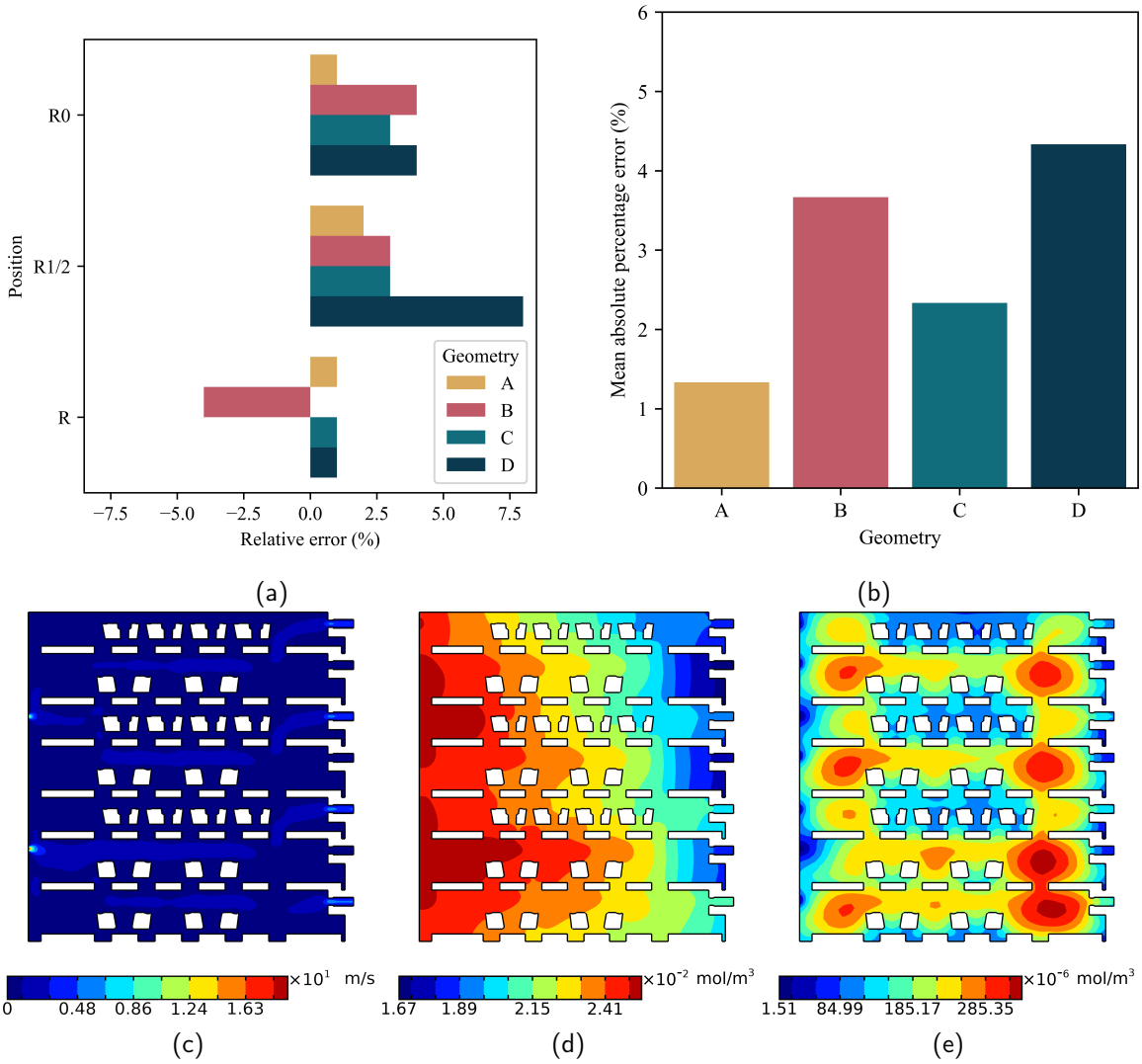


Figure 4.5: (a) Relative error for the CFD predictions for 3 different positions with available production data inside the reactor. Simulations are performed for four different 7-disk geometries in total. (b) Mean absolute percentage error (averaged over the 3 positions for which data are available) for the CFD simulations for the 4 different reactor geometries. (c) Velocity magnitude, (d) Precursor Concentration and (e) Water Concentration inside the reactor at a certain time during the deposition.

Hyperparameter selection

Optimal model performance, is influenced by the choice of hyperparameters for each method. The most important hyperparameters of the implemented tree-based ensemble methods are:

1. The maximum depth of the trees (d_{max}), i.e. the number of bifurcations of the main "branch" of the tree. Selecting too large a tree depth can lead to overfitting, which in essence means that the model fails to generalize accurately.

2. The number of trees (B). A large number of trees reduces the variance of bagging methods, however it can lead to overfitting in the case of boosting methods.

3. For boosting methods specifically, another important hyperparameter is the learning rate (λ). The choice of λ usually affects the optimal B. For example, a very small λ usually requires a large B to achieve satisfactory performance.

Searching for the optimal model hyperparameters in an exhaustive manner is a computationally expensive task. The time required for all 5 tree-based methods using an exhaustive grid search approach performing 10-fold cross-validation was 43 core hours on an 11^{th} Gen Intel(R) Core(TM) i7-1185G.

To demonstrate here the effect of d_{max}, results are shown for fixed values of B and λ (cf. Fig. 4.6). For a constant number of trees ($B = 10000$), boosting methods show better performance for low values of d_{max}. On the contrary, bagging methods indicate better performance for higher values of d_{max}.

Overall, for all the hyperparameters tested, boosting methods appear to outperform their bagging counterparts. Out of the two boosting methods, the XGBoost method displays higher training and predicting speed. Specifically, for the same training set and the same hyperparameters ($B = 10000$, $d_{\text{max}} = 5$ and $\lambda = 0.01$), the average training time over 10 cross-validation splits is 16.5s for the XGBoost model and 99.5s for the GBRT model. Moreover, the average prediction time is 20ms for the XGBoost model and 333ms for the GBRT model. Therefore, due to its lower computational cost, further hyperparameter tuning will take place for the XGBoost algorithm, in order to find the optimal hyperparameter combination.

After selecting the optimal value of maximum depth, we further investigate the effect of the number of trees B on the accuracy of the XGBoost model. As indicated in Table 4.2, the accuracy of the model drastically improves when $B \geq 500$, nevertheless, the trade-off is in the form of increased computational cost.

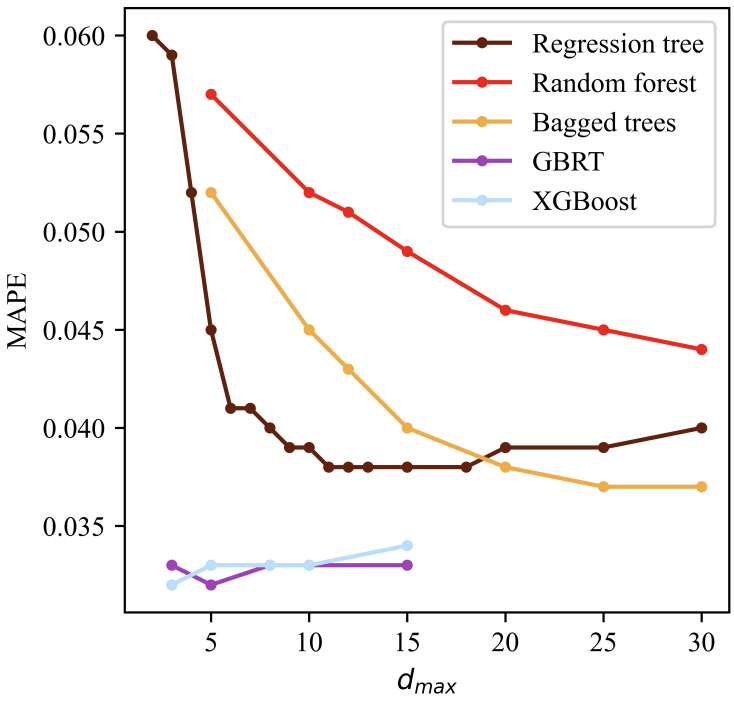


Figure 4.6: MAPE vs d_{max} for all methods after 10-fold cross-validation. $B = 10000$ for all ensemble methods. $\lambda = 0.01$ for the boosting methods. For the base method (regression tree) and bagging methods (Bagged Trees and Random Forests) increasing the maximum depth of the trees leads to a reduced MAPE. For the boosting methods (GBRT and XGBoost), the MAPE increases when increasing the maximum depth of the trees. Random forest regression performing worse than the simple regression tree can be attributed to the fact that it only considers a subset of available features when building each tree of the ensemble.

Table 4.2: XGBoost model results after cross-validation for various values of B, where $d_{\text{max}} = 5$ and $\lambda = 0.01$. As expected, an increased number of base predictors improves the performance of the ensemble boosting method. However, it also increases the training time and prediction time of the model. All metrics are averaged over 10 cross-validation splits.

Number of trees (B)	MAPE	$\bar{t}_{\text{train}}$ (s)	$\bar{t}_{\text{pred}}$ (ms)
10000	3.1%	16.3	20
5000	3.3%	8.0	14
2000	3.4%	3.3	9
1000	3.6%	1.7	9
500	3.9%	0.9	8
200	12.6%	0.4	8
100	33.8%	0.2	10

Following hyperparameter optimization and tuning, the final values selected for the XGBoost model are the following: $d_{\max} = 5$, $B = 10000$, $\lambda = 0.01$.

Machine learning outcomes

Two more accuracy metrics are introduced here, the mean square error (MSE) and the coefficient of determination (R^2). When the model is trained or tested on N observations and for each observation i the predicted value is $\hat{y}_i$ while the actual value is y_i and the average of the actual values is $\bar{y}$, MSE and R^2 can be written as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{4.15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4.16}$$

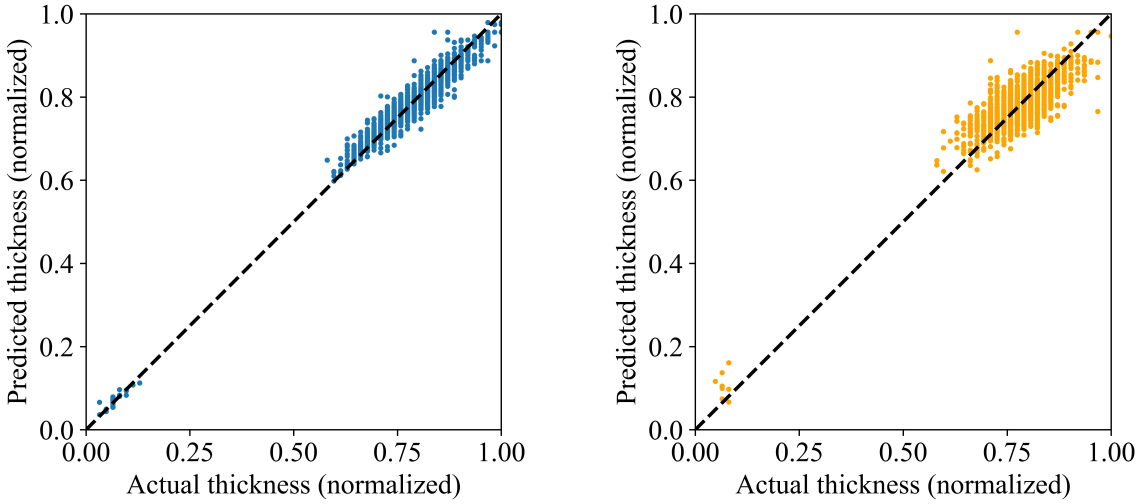
The prediction error of XGBoost regression model for the training set, reaches a MAPE of 0.9%, versus 3.1% for the test set. The prediction accuracy of the XGBoost model on the training set and on the test set can be summarized in Figs. 4.7a and 4.7b respectively. Due to the confidentiality of the production data, absolute α-Al_2O_3 thickness values cannot be presented. Therefore, only relative error values and normalized thickness values are presented.

4.5.3 CFD vs ML

Predictive accuracy

For the test-case reactor set-up presented in Fig. 4.2, the prediction results for the position closest to the outlet for both methods are given in Table 4.3. Disk position is counted from the bottom to the top of the reactor.

Despite the significant difference in the computational effort involved in the CFD model in comparison to the ML regression model, both methods have comparable accuracy on the test-case. CFD predictions for the test reactor have a mean absolute percentage error of 6%, while XGBoost makes predictions with a mean absolute percentage error of 4.4%. The high error in the



(a) (b)

Figure 4.7: (a) Training set performance: MSE:0.005 | MAE:0.051 | MAPE:0.9% | R^2:0.980. (b) Test set performance: MSE:0.059 | MAE:0.187 | MAPE:3.1% | R^2:0.753.

Table 4.3: XGBoost prediction accuracy vs CFD prediction accuracy for the coating thickness of inserts closest to the reactor outlet (R position). Errors relative to the available production data are presented. The high error in the prediction of the CFD model for the 6th reactor disk can be attributed to the fact that it is the bottom-most disk of the simulated 7-disk geometry, and therefore the effect of the inlets and outlets that are below it is not taken into account.

Disk position	CFD prediction	XGBoost prediction
39	3.2%	3.5%
35	1.0%	-3.1%
23	-4.0%	-7.0%
10	1.0%	-5.5%
6	20.6%	-2.8%
MAPE	6.0%	4.4%
Total prediction time (s)	43200	0.1

prediction of the CFD model for the 6th reactor disk (20.6%) can be attributed to the fact that it is the bottom-most disk of the simulated 7-disk geometry and therefore the effect of the inlets and outlets that are below it is not taken into account. This can be solved by an extra 7-disk simulation, where the disk of interest won't be in the bottom-most position. This would of course further increase the computational cost of the CFD approach. The maximum observed absolute

relative error for the predictions of the XGBoost model on the test-case reactor is 7%.

Computational performance

Although the predictive accuracy of the two approaches is similar, they demonstrate a very noticeable contrast when it comes to their computational performance. Specifically, in the case of CFD, making predictions for an entire production run would require 4 or 5 7-disk simulations. This corresponds to a computational cost of 12 to 15 core hours. On the other hand, using the XGBoost model to make predictions for an entire production run comes with a computational cost of less than 1 core second. This translates to a reduction of more than 99.99% in required resources.

4.5.4 GappyPOD

Results of our GappyPOD implementation will be presented for two different cases:

1. The case of the full dataset.

2. The case of a single reactor.

In each case, the dataset consists of time-instances of the state vector, over a period of 30 secs. Therefore, the full dataset eventually consists of 372 snapshots, whereas in the single reactor dataset, it consists of 31 vectors.

In both cases, 87.5% of the available snapshots are used to derive the POD basis of the training set. The rest of the snapshots (12.5%) are kept and used for the validation of the method. For both cases, the data are scaled in the range of $[0, 1]$ using min-max normalization.

The number of modes used for the POD basis are selected after checking the energy retained by the modes and the resulting reconstruction error. The total retained energy for the full dataset and the single reactor dataset, is shown in Fig. 4.8a and Fig. 4.8c respectively, whereas the reconstruction error as a function of the basis size is shown in Fig. 4.8b and Fig. 4.8d respectively.

The full reactor dataset requires at least 50 POD modes to capture more than 95% of the energy of the system, with a corresponding reconstruction error (RMSE) of 0.0059. The single reactor dataset, is accurately represented by 15 POD modes that reflect more than 98% of the

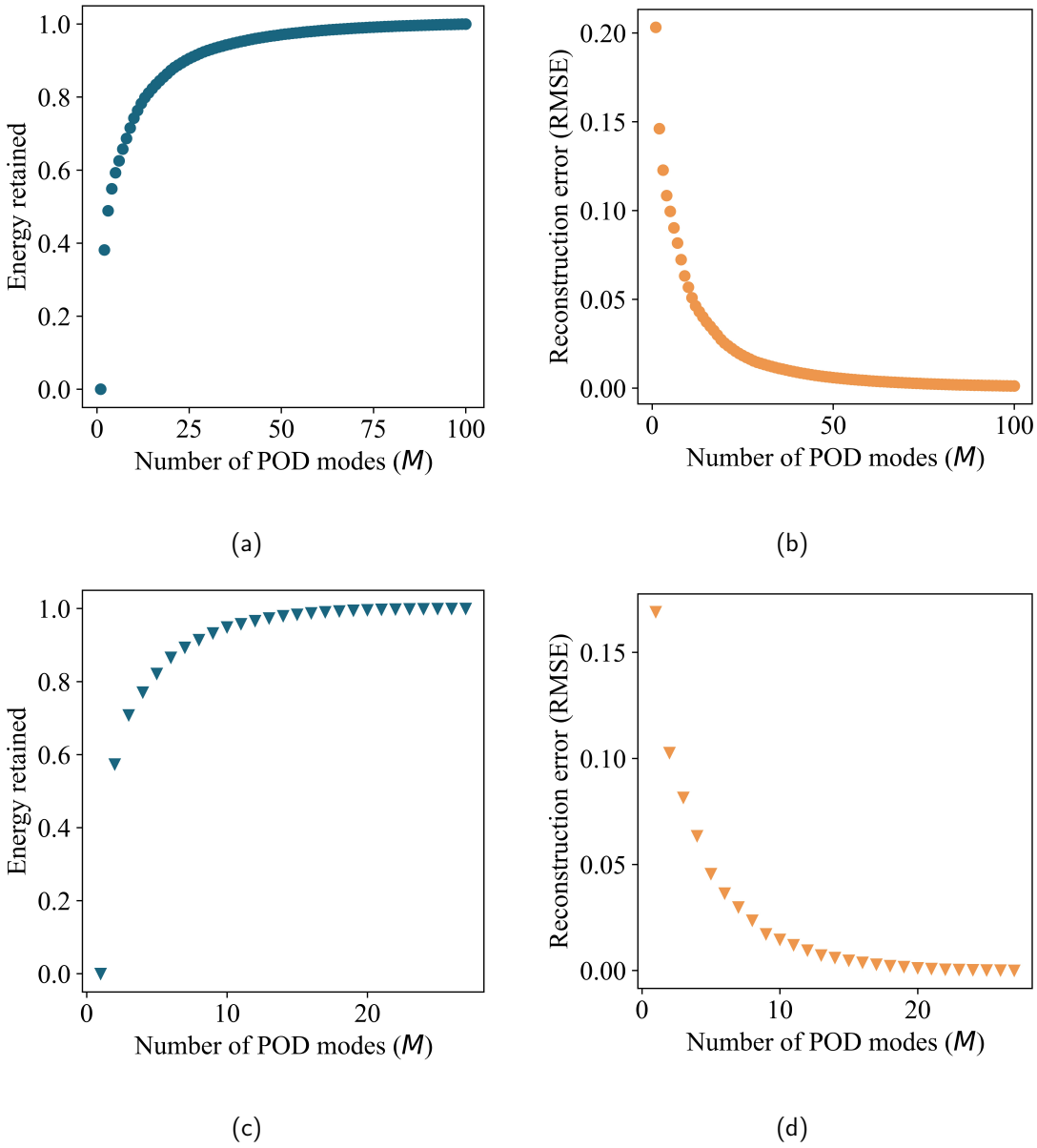


Figure 4.8: The energy retained (in blue) and the reconstruction error (in orange) of the POD approximation using M modes. (a), (b): Energy and reconstruction error for the full dataset. Only the first 100 modes are shown. (c), (d): Energy and reconstruction error for the single reactor case.

energy with a reconstruction error (RMSE) of 0.004. Eventually, for the immediate comparison of the results, the same basis size is considered, equal to 15 POD modes. The corresponding retained

energy and error are shown in Table 4.4.

Table 4.4: Number of POD modes selected for each case, along with the corresponding retained energy and reconstruction error.

Case	# POD modes	Energy retained	Recon. error (RMSE)
Full dataset	15	81.69%	0.0373
Single reactor	15	98.70%	0.0040
Single reactor	5	82.74%	0.0456

After selecting the size of the POD basis for each case, the mask elements for GappyPOD are obtained using the greedy algorithm described in Section 4.4.4. It should be noted that the mask length should be greater or equal to the size of the POD basis. For all three cases, we allow one mask element more than the size of the POD basis. It should be noted that in all cases the mask elements acquired consist of all the quantities of interest (velocity magnitude, pressure, precursor concentration, water concentration) discussed in Section 4.4.2.

After acquiring the mask elements, the RMSE between the GappyPOD approximation and the test set, along with the RMSE between the GappyPOD approximation and the POD reconstruction, can be calculated. Specifically, for the case of the full dataset, the RMSE between the GappyPOD approximation and the test set is 0.0648 while the RMSE between the GappyPOD approximation and the POD reconstruction is 0.0512 (cf. Fig. 4.9). For the case of the single reactor, the RMSE between the GappyPOD approximation and the test set is 0.0099 while the RMSE between the GappyPOD approximation and the POD reconstruction is 0.0064. If we choose to make a comparison using the number of POD modes with the same retained energy and reconstruction error, we choose 5 POD modes (82.74% retained energy and 0.046 reconstruction error) and 6 mask elements for the single reactor case. Then, the RMSE between the GappyPOD approximation and the test set is 0.0474 while the RMSE between the GappyPOD approximation and the POD reconstruction is 0.0143.

The performance of the method, is linked to how well the dataset is spanned by the selected POD vectors, generally implying that a larger POD basis is beneficial for the results. Nevertheless, since the ambition of this approach is to select only a few measurements as mask elements, it is more beneficial to work with the smallest possible number of POD vectors.

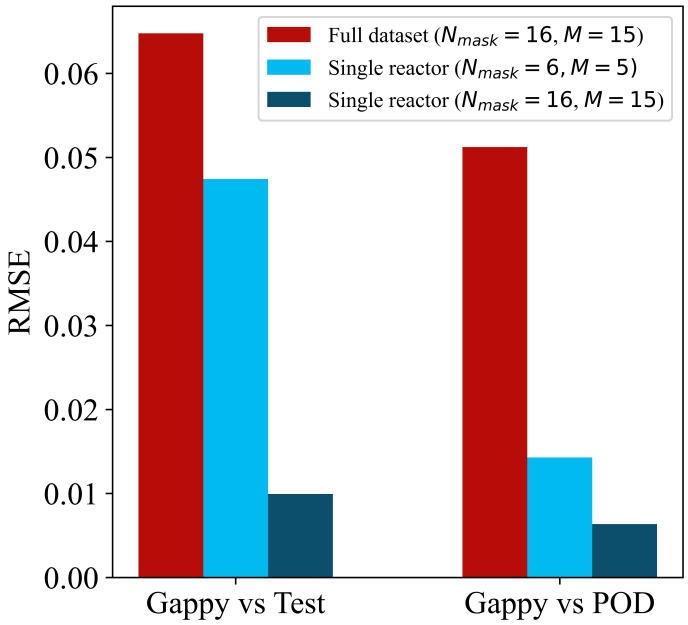


Figure 4.9: On the left: Error between the GappyPOD approximation and the snapshots of the test set for all cases. On the right: Error between the GappyPOD approximation and the POD approximation for all cases. It is evident that the single reactor case shows the lowest errors. This is probably due to the lower variance observed in the dataset of the single reactor when compared with the full dataset. For the case of the single reactor, using a smaller POD basis (5 modes instead of 15) leads to an increase in both errors.

4.6 Conclusions

This work presents an overview of the implementation of equation-based and machine-learning methods in industrial-scale deposition applications. The challenges associated with the complexity of the process and the characteristics of real production data are discussed and the methods to overcome them are presented.

In the equation-based approach, a reduced model is presented and validated with production measurements of the coating thickness. The simplifications introduced and the pertinent assumptions upon which they are based are discussed, along with the results. The trade-off between the computational cost associated with the CFD model and the physical insight obtained, is discussed and compared to the ML approach. Coating thickness predictions are possible with an average error of 6%. In addition, the CFD model, predicts the distributions of velocity, and reactive species,

illuminating thus, the mechanisms that contribute to the final product. Furthermore, it can be used to predict the thickness achieved in parts of the reactor where there are no measurements. Moreover, the CFD approach also allows extrapolating for different process conditions and different inlet reactant concentrations. For the 7-disk CFD approach, the results of Table 4.3, show that appropriate selection of the 7-disk "building blocks" for the simulations is of high importance for the accuracy of the prediction.

The ML approach is discussed in detail, as far as the possible specific methods are concerned. The suitability of each is assessed, based on the data available. Eventually the best performing ML method, XGBoost, is able to deliver accurate and time-efficient coating thickness predictions, but cannot provide insight into the transport of species that determines the coating thickness.

The implementation of GappyPOD for this specific application, shows how data-driven methods and CFD results can be intertwined to provide further insight on the important quantities of interest inside the reactor. By further analysis of the resulting mask elements, we can explore the hypothetical scenario of sensor placement inside such reactors. Furthermore, we can reconstruct entire snapshots from a few measurements inside the reactor, reducing in this way the computational cost of the problem.

It should be noted that the strategy employed here is not exclusive to CFD modeling. The same workflow could still be implemented in other applications, regardless of the equation-based modeling approach used. The only limiting factor would be the amount and type of available data for the application.

Another important observation is that specific *combinations* of inputs can lead to the same outputs. This merits further investigation, due to its importance in the actual production process, which is the topic of future work.

To conclude, it is clear that each individual approach is a valuable tool in studying a complex process offering different advantages: physical insight and extrapolation abilities in CFD and time-efficient, accurate predictions in ML. It is therefore worth investing the effort in each one of them, and ultimately, in merging them in a hybrid approach with additional benefits. Reduced representations, or the full state space of the CFD simulations could be used as features for predictive machine learning algorithms, potentially improving predictive performance. However,

this would come with the high computational cost associated with conducting the CFD simulations for all available production runs. Ideally, the resulting model could combine high accuracy, time-efficient predictions, and excellent extrapolation ability, moving in this way toward a digital twin of the process.

Chapter 5

Integrating supervised and unsupervised learning approaches to unveil critical process inputs

This chapter is reproduced from P. Papavasileiou, D. G. Giovanis, G. Pozzetti, M. Kathrein, C. Czettl, I. G. Kevrekidis, A. G. Boudouvis, S. P. A. Bordas, and E. D. Koronaki, "Integrating supervised and unsupervised learning approaches to unveil critical process inputs," *Computers & Chemical Engineering*, vol. 192, p. 108 857, 2025. DOI: 10.1016/j.compchemeng.2024.108857. As the first author of the article, I participated in the development of the proposed ML framework, the analysis of the resulting data, the validation and visualization of the results, and the writing of the manuscript. The CRediT contribution statement for the publication is presented below.

CRediT authorship contribution statement

Paris Papavasileiou: Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Methodology. **Dimitrios G. Giovanis**: Writing – review & editing, Validation, Supervision, Funding acquisition, Formal analysis, Conceptualization. **Gabriele Pozzetti**: Writing – review & editing, Resources, Data curation, Conceptualization. **Martin Kathrein**: Writing – review & editing, Validation, Resources, Project administration, Conceptualization. **Christoph Czettl**: Writing – review & editing, Resources, Data curation, Conceptualization. **Ioannis G. Kevrekidis**: Writing – review & editing, Methodology, Formal analysis,

Conceptualization. **Andreas G. Boudouvis**: Writing – review & editing, Supervision, Formal analysis. **Stéphane P.A. Bordas**: Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Eleni D. Koronaki**: Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Abstract

This study introduces a machine learning framework tailored to large-scale industrial processes characterized by a plethora of numerical and categorical inputs. The framework aims to (i) discern critical parameters that influence the output and (ii) generate accurate out-of-sample qualitative and quantitative predictions of production outcomes. Specifically, we address the pivotal question of the significance of each input in shaping the process outcome, using an industrial Chemical Vapor Deposition (CVD) process as an example. The initial objective involves merging subject matter expertise and clustering techniques exclusively on the process output, here, coating thickness measurements at various positions in the reactor. This approach identifies groups of production runs that share similar qualitative characteristics, such as film mean thickness and standard deviation. In particular, the differences of the outcomes represented by the different clusters can be attributed to differences in specific inputs, indicating that these inputs are potentially critical to the production outcome. Shapley value analysis corroborates the formed hypotheses. Leveraging this insight, we subsequently implement supervised classification and regression methods using the identified critical process inputs. The proposed methodology proves to be valuable in scenarios with a multitude of inputs and insufficient data for the direct application of deep learning techniques, providing meaningful insights into the underlying processes.

5.1 Introduction

Chemical vapor deposition (CVD) is a widely used chemical process for producing thin films with various properties, applied in semiconductor manufacturing [47], [48], membranes [205], [206], protective [54], [207] and wear-resistant [55], [208] coatings. Although Computational Fluid Dynamics (CFD) models traditionally explore CVD complexity [4], [5], [146], [150], [151], [161], [164], [209], [210] their efficiency and adequacy are challenged in cases involving unknown chemical reactions or intricate reactor geometries. The computational cost of large-scale industrial process models and the nonlinear nature of competing physical and chemical mechanisms further limit the utility of CFD as a viable "digital twin". It is also possible that there are different process outputs arise

for the same inputs, which is also linked to non-linearity [7], [143].

Recently, Machine Learning (ML) has emerged as a promising alternative in the era of Industry 4.0 with abundant process data. ML applications range from maintenance management [168]–[170] and production planning [21], [176] to outcome prediction [17], [36], process control [23] and optimization [22]. ML models can also be developed based on preexisting physics-based models, in order to further investigate the modeled process [8], [9], [211].

Despite recent advances in explainable AI (XAI), challenges persist in addressing the "black box" nature of ML models. However, tools such as SHAP (SHapley Additive exPlanations) offer improved explainability using a game theory approach [37], [212]–[214].

This study utilizes production data from an industrial CVD reactor for the production of wear-resistant cutting tool coatings. The data encompass details about the reactor setup and the process inputs; thickness measurements of the Ti(C,N)/α-Al_2O_3 coating in 15 positions within the reactor are considered process outputs.

Implementing state-of-the-art (SotA) methods faces challenges that include:

- Process complexity, namely, multiscale interacting phenomena in intricate geometries.
- A multitude of numerical and categorical inputs, with little insight of their impact on the process outcome.
- Noisy and heterogeneous data, collected over months or years with varying instrumentation and calibration, which cannot be categorized as "big".

Several different options are available in the literature related to the discovery of important process parameters and the facilitation of subsequent modeling attempts. Variable Importance in Projection (VIP) parameters [215], [216], a byproduct of Partial Least Squares (PLS) models, have traditionally been used to determine the impact of process inputs on the output [216]–[218].

Variable selection tools have been shown to enable improved performance and subsequently lead to a greater understanding of the importance of input variables on model output [219]. To this end, several powerful dimensionality reduction techniques based on Principal Component Analysis (PCA) or Diffusion Maps (DMaps) [220], [221] can lead to the discovery of effective process parameters [222], [223].

Despite the effectiveness of existing methods for strictly numerical data, challenges arise when dealing with datasets rich in categorical features, as seen in this application. This work aims to propose an ML workflow for the identification of critical process inputs without labeled data, an essential contribution to control, optimization, and experimental design.

Our approach involves an unsupervised analysis of process outputs to identify clusters of similar production runs. Subsequently, we analyze relevant process input data to discern distinguishing characteristics within these clusters. Our findings are supported by subject matter expertise. Shifting to supervised learning, we use cluster labels to train a classifier for predicting these labels given specific process inputs. Furthermore, we attempt to create a regression model for predicting thickness measurements. Finally, we employ SHAP and Shapley values to interpret the model output.

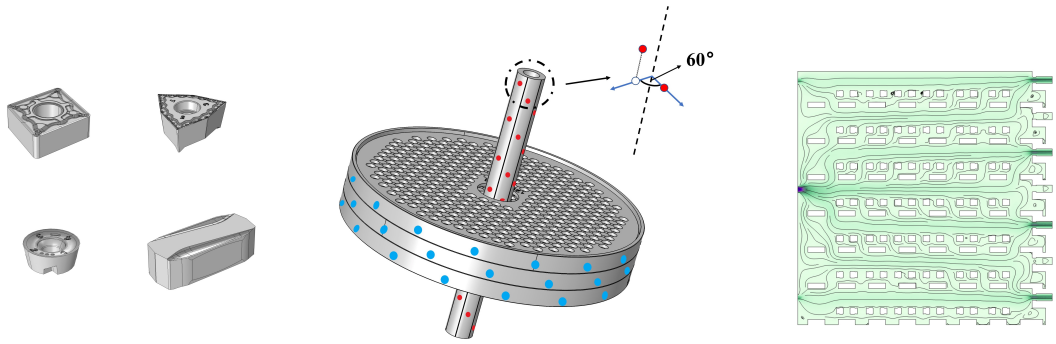
The manuscript is structured as follows. A brief overview of the process and the available production data is given in Section 5.2. The various machine learning methods implemented (supervised, unsupervised) are presented in Section 5.3. The results are discussed in Section 5.4, followed by concluding remarks in Section 5.5.

5.2 Process overview

The studied process involves two coating steps carried out inside a commercial, industrial-scale Sucotec SCT600TH CVD reactor. To start with, a Ti(C,N) base layer of approximately 9 µm is deposited under a chemical system consisting of $TiCl_4$–CH_3CN–H_2–N_2–CO at a temperature of 900°C and a pressure of 100 mbar [70], [71] on cemented carbide cutting tool inserts, shown in Fig. 5.1a. $TiCl_4$ is used as the Ti source and CH_3CN is used as the source of C and N. The second step involves the deposition of an alumina layer under specific conditions: T=1005°C and p=80 mbar, from a mixture of gas reactants that includes $AlCl_3$–CO_2–HCl–H_2–H_2S. This step takes around 3 hours to complete [1]. For a more detailed review of the α-Al_2O_3 coating step, the interested reader is referred to previous work [32].

The CVD reactor consists of 40-50 perforated disks, stacked one on top of the other. The inserts to be coated are placed on each disk. For illustrative purposes, a schematic of three such disks is shown in Fig. 5.1b. Specially designed perforations in the cylindrical feeding tube, which is

placed in the center of the reactor, ensure the uniform distribution of gas reactants over and around the inserts: the perforations are placed antipodally and there is a 60° angle difference between the axis connecting the inlets at each disk level. The feeding tube rotates at a fixed rotational speed of 2 RPM, something that also ensures the uniform distribution of the reactants over and around the inserts. Signs of deposition can be observed on all surfaces within the reactor, be it inserts, disks, or walls.



(a) (b) (c)

Figure 5.1: (a) Examples of the coated cutting tool inserts. (b) A 3D representation of a 3-disk part of the reactor. The inserts are placed on each of these disks. In red: inlet perforations on the rotating inlet tube. In blue: outlet perforations for each disk. The perforations and the rotation of the inlet tube allow for the even flow of the gas reactants inside the reactor. The deposition can take place in everywhere within the reactor (disks, inserts, walls etc.). (c) Streamlines for a snapshot of a 2D time-dependent CFD model for a 7-disk part of the reactor [32]. The wide white rectangular areas represent the disks, while the smaller white areas of various shapes represent the inserts.

It is worth noting that each insert has a dedicated disk design which ultimately suggests that the interior geometry of the CVD reactor changes every time that it is set up.

The desired process outcome is uniform coating thickness distribution for the same insert and also uniform mean thickness across all production runs, all reactors, and all production sites, as this ensures consistent product life (quality) [74]. In practice, the desired uniformity is not always achieved, and therefore a systematic way of identifying the influential aspects of coating uniformity becomes necessary.

5.2.1 Available data

After each production run, thickness measurements are taken at three positions on five disks of interest, schematically shown in a representative geometry in Fig. 5.2. The thickness of the Ti(C,N) and α-Al_2O_3 coating layers is measured using the Calotest method [197]. In this method, a small spherical cavity is created on the coated inserts by using a rotating ball with a known geometry, producing a tapered cross section of the film. When observed under an optical microscope, this allows the measurement of the thickness of both layers. These measurements have been used in previous work, both for the development of a CFD model of the process [32], and for the implementation of ML approaches for the prediction of coating thickness [36].

Coating thickness is a vital measure of product quality for CVD applications. The long-term experience of the practitioners led to the selection of these 15 measurements for testing the quality of each production run. It should be noted that in case additional quality-related data (i.e. roughness of the coating) become available, they can be easily incorporated into the framework presented in this paper, in conjunction with thickness.

Additionally, the available dataset contains information about a) the process input parameters and b) the reactor geometry and setup. Some examples of these features include, but are not limited to:

- The components of the reactor setup that determine the overall interior geometry, i.e. the sequence according to which the disk/inserts are stacked to form the overall reactor.
- The surface area of the inserts on each disk.
- The production "recipe", a feature that encodes several process parameters and steps. We should note that there can be several versions of one recipe. There are a total of four base recipes present in the dataset with five versions for each (marked $V21$, $V20$, and older variants). This makes up a total of 20 recipes.
- The serial number of the reactor used for the production run.

An important contribution of this work emerged in the context of data exploration and pre-processing. It became necessary to engineer additional features based on our intuition (subject

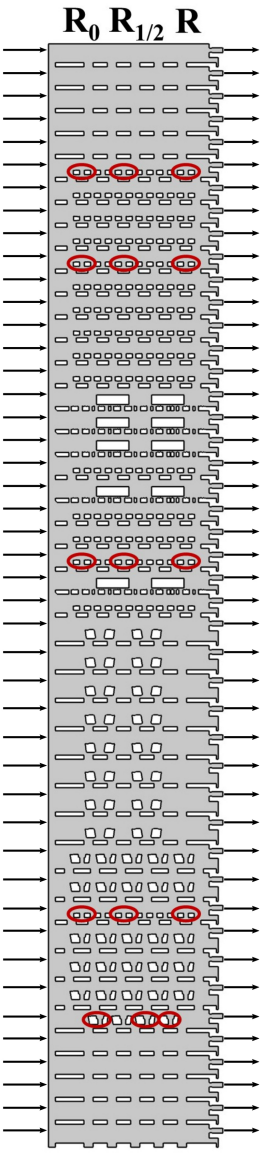


Figure 5.2: A 2D representation of the CVD reactor. The inlet of the reactants is found on the left side, and the outlet is located on the right side. The gray area represents the domain where the gas reactants flow. The white areas represent the stacked disks and the inserts that are coated. Circled in red, we highlight the 15 locations where inserts are most commonly measured (R_0: locations closest to the reactor's inlet, $R_{1/2}$: locations around the middle of the disk radius, R: locations closest to the reactor's outlet). These thickness measurements can be used for several tasks, such as the development of CFD or ML approaches for the prediction of the process outcome. The arrows indicate the gas reactant inlets and outlets.

matter expertise) regarding the existing inputs. These engineered features include the total surface area per reactor, the standard deviation of the surface area within the reactor, and the difference between the nominal and actual surface area within the reactor. The nominal surface area is the surface area considered by the production recipe and does not always coincide with the actual surface area. For more information on the available data, its type and its characteristics, the interested reader is referred to Table 5.1 and to previous work by Papavasileiou *et al.* [36]. A comparison of our approach with systematic methods for feature combinations, such as polynomial combination or even symbolic regression, is underway and outside the scope of this work.

Table 5.1: Available data for each production run. Asterisks denote inputs deemed potentially important by empirical knowledge.

Feature	**Origin**
Thickness - 15 disks of interest	Raw data
Number of inserts per disk	Raw data
Surface area of inserts per disk	Raw data
Disk setup sequence	Raw data
Insert geometry per disk	Raw data
*Production "recipe"	Raw data
Reactor used	Raw data
Year of production	Raw data
*Total surface area of inserts inside the reactor	Engineered
*|Nominal "recipe" surface area - actual surface area|	Engineered
*Surface area standard deviation	Engineered

The availability of data on the outcome of the process in the form of thickness measurements (mentioned in this section and visually presented in Fig. 5.2), along with information on the reactor setup, motivates the use of several machine learning methods. Following the data cleanup and feature engineering steps briefly presented in Fig. 5.3, the data can be easily used for the implementation of a plethora of ML methods. A detailed overview of the methods implemented is presented in Section 5.3.

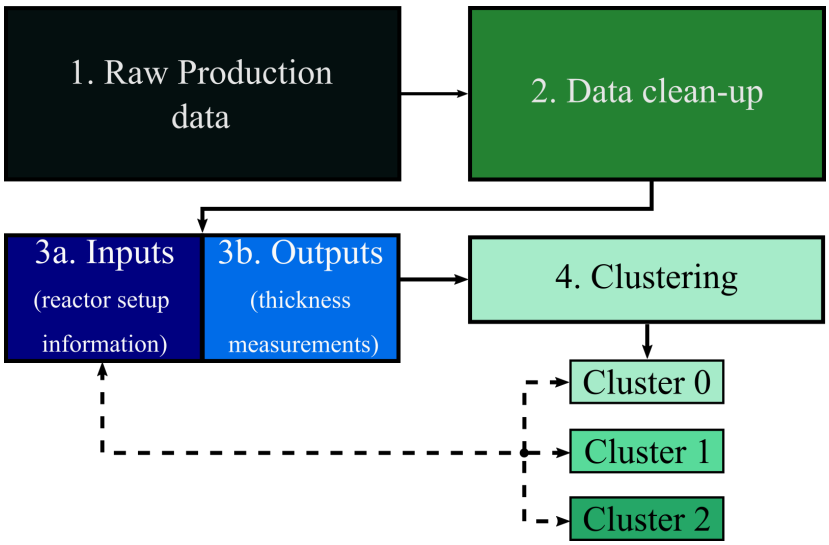


Figure 5.3: A flowchart indicating the steps leading to the clustering part of the analysis. First, the raw production data are extracted. Then, the observations with missing values are disregarded and typographical errors in the entries are corrected. In step 3, the data are divided in inputs and outputs. This step also includes the engineering of new potentially useful features by using the existing ones. Subsequently, the production runs are clustered based solely on the outputs (15 thickness measurements). Finally, we attempt to correlate the resulting clusters to differences in the process inputs.

5.3 Machine learning methods

5.3.1 Unsupervised learning

Unsupervised learning algorithms take unlabeled data as inputs to discover interesting patterns in the data (e.g. association rule analysis) or try to create subgroups - or clusters - of similar observations within the dataset [82]. Dimensionality reduction techniques such as the widely used Principal Component Analysis (PCA), autoencoders [224], and diffusion maps [220], [221] also fall under unsupervised learning, as they provide a reduced data representation without requiring the corresponding response. The clustering and dimensionality reduction techniques implemented are briefly discussed in the following sections.

Clustering

Clustering algorithms are based on the concept of dissimilarity (or similarity) between observations, which determines their grouping. Typically, these algorithms use a similarity matrix, where pairwise similarities between observations are represented. For quantitative variables, the commonly employed metric is the Euclidean distance, while alternative distance metrics can also be used [80], [82].

Clustering algorithms are categorized into various categories. Partitional approaches, such as the k-means algorithm, involve assigning observations to clusters based on distances to centroids iteratively, requiring an *a priori* choice of the number of clusters and sensitive to initial centroid positions [83]. Density-based algorithms, such as OPTICS and DBSCAN, identify clusters by considering areas of high density separated by low-density regions. Certain algorithm parameters, such as the minimum points that form a cluster and the minimum distance between the core points require specification [84]–[86]. Hierarchical clustering methods link data points according to criteria, progressively creating clusters until a single cluster is achieved in the case of agglomerative clustering, or progressively splitting clusters starting until each observation is its own cluster in the case of divisive clustering. The results depend on the distance metric and the linkage criteria selected [87], [88]. Additional methods include model-based and spectral methods [225], [226].

Here, we focus on *agglomerative* hierarchical clustering, implementing a Ward linkage criterion for merging the clusters. This is an established variance minimization approach [227] that works by minimizing the sum of squared differences within all clusters. Agglomerative hierarchical clustering is selected because it provides insight on how the data merges depending on the number of clusters chosen. This information is readily available in the form of a dendrogram, such as the one presented in Fig. 5.4a.

For this specific problem, the 15 available thickness measurements of 603 production runs are used as inputs (cf. Section 5.2.1). The clustering results are then interpreted based on the characteristics of the resulting clusters. Our goal is to identify production runs that are similar to each other and to try to uncover the discerning features of these clusters.

5.3.2 Supervised learning

Supervised learning algorithms, unlike unsupervised ones, require labeled data, associating features (x_i) with responses (y_i). Supervised learning tasks include regression for continuous variables and classification for binary or ordinal responses [81].

The methods evaluated for this work include: (a) linear methods: for regression, lasso [89], and ridge [90] regression, and logistic regression for classification tasks. (b) Support vector machines (SVMs) [91] that can be categorized as linear or nonlinear methods based on the kernel used for classification tasks. (c) Tree-based methods: involving classification and regression trees [92] and their ensemble counterparts such as random forests [93], gradient-boosted trees [94], extra trees [95], and XGBoost [34], which combine numerous trees to enhance performance [96]. (d) Artificial neural networks (ANN), whose diverse architectures [97] can provide valuable options for both classification and regression tasks.

In this work, logistic regression, random forests, SVM, extra trees, gradient-boosted trees, XGBoost, and ANNs are implemented for supervised learning tasks. However, only the methods that demonstrate the best performance for our dataset are presented in Section 5.4.

5.3.3 Shapley values

Shapley values, originally introduced by Shapley [212] and proposed as a tool to analyze machine learning models in [37], [228] assess the average contribution of each feature's value to predictions, providing an understanding of how alterations to a variable might influence the ultimate model output. The fundamental idea of Shapley value-based explanations in machine learning is to fairly distribute credit for a model's output among its input features, using principles from cooperative game theory. To bring machine learning models into a game theory context, each of the inputs of the model can be considered a player that either joins or does not join the game. Joining the game means that the input value is known, while not joining the game means that the input value is unknown. Shapley values have an additive nature. In the context of explaining machine learning models, this means that the SHAP values for all input features will always total the difference between the baseline (expected) model output and the actual model output for the prediction being explained.

In the context of this work, a SHAP (SHapley Additive exPlanations) analysis is conducted on the proposed regression models (cf. Sections 5.3.2 and 5.4.5) using the *shap* Python library.

5.4 Results

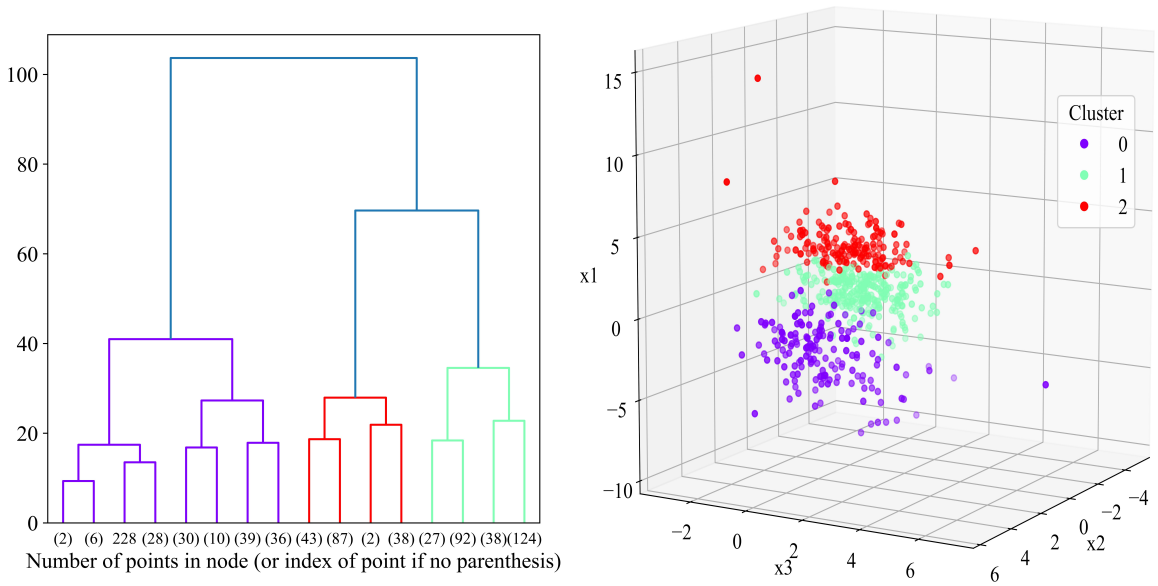
5.4.1 Clustering

As mentioned in Section 5.3.1, the agglomerative hierarchical clustering algorithm with a Ward linkage criterion is implemented for clustering the 603 production runs.

The clustering algorithm utilizes the 15 thickness measurements for each of the 603 production runs, forming a 603×15 matrix. Clusters are then created solely based on the process outputs. Subsequently, the distinctive features are identified by analyzing the process inputs for each production run.

The hierarchical clustering algorithm generates a dendrogram that illustrates cluster levels, member counts, and dissimilarities. The clusters are depicted as branches of a tree, culminating in the "trunk", representing the final cluster (by agglomerating smaller ones). In our case, the resulting dendrogram is shown in Fig. 5.4a. By selecting a dissimilarity threshold, we can discern one, two, three, or more clusters. In Fig. 5.4a, the three clusters are colored purple, red, and green. A higher dissimilarity threshold merges the red and green clusters into a single blue cluster (as shown in Fig. 5.4a). The resulting clusters are visualized in a reduced three-dimensional space (through projection on three principal components) in Fig. 5.4b.

As mentioned above, the thickness and its uniformity throughout production runs is a very effective process performance indicator and product quality metric. Thus, production runs with a higher average thickness and a lower standard deviation can be considered superior to those with a lower average thickness and higher standard deviation. We observe that the thickness within the clusters follows a normal distribution, and therefore we can calculate the first and second statistical moments (that is, the mean (μ_{thick}), and standard deviation (σ_{thick}) and visualize the thickness distributions as shown in Fig. 5.5. These distributions and their qualitative significance for the process are the basis for the following sections, where we first try to identify parameters that potentially cause these qualitative differences (cf. Section 5.4.2) and then try to exploit them



(a) (b)

Figure 5.4: (a) Resulting dendrogram of the clusters output by the implemented agglomerative hierarchical clustering algorithm using a Ward linkage criterion. The three main clusters of interest are colored purple, red, and green. We note that by selecting a slightly higher dissimilarity threshold, the red and green clusters can be merged and viewed as a larger cluster (shown in blue). (b) The three resulting clusters, visualized in a reduced 3D space. The three clusters appear to be well-formed. PCA was used for finding the 3D reduced space.

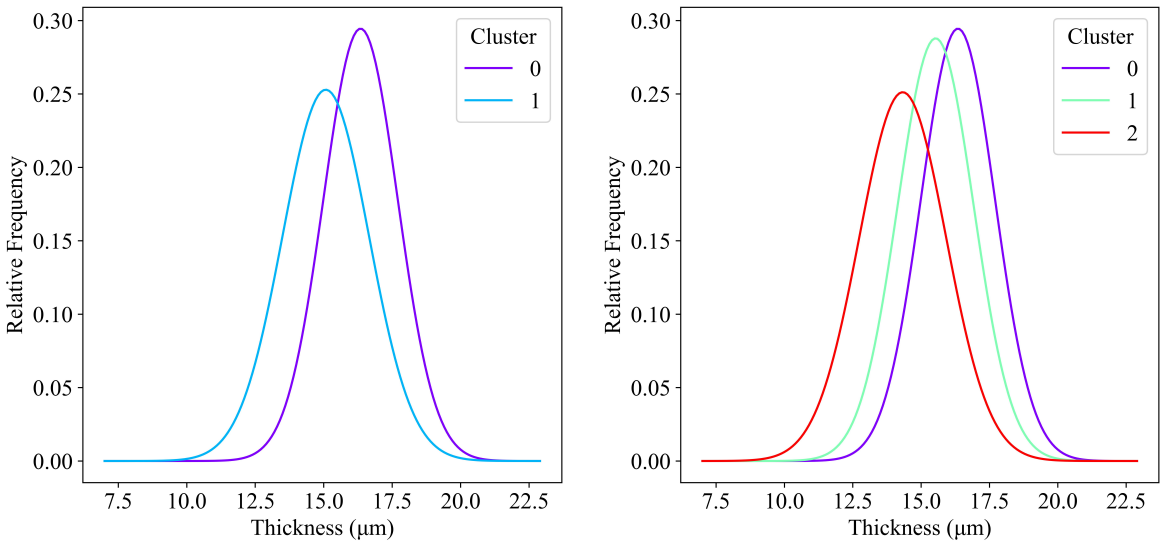
to make predictions for future production runs in Sections 5.4.3 and 5.4.4.

5.4.2 Critical input identification

In this section, the focus shifts to the process inputs whose variation is critical for each cluster. We propose three different ways for assessing the relative importance of process inputs.

1. Intuition-based approach: By finding characteristics that are predominantly different in each cluster, we can assess their importance on the process outcome (cf. Section 5.4.2).

2. Supervised learning approach: Classification algorithms are trained using the cluster labels of the clustering step as outputs and various inputs: some process inputs lead to higher accuracy, which is an indication of their importance. Conversely, less important inputs have an adverse effect on the accuracy of the classifier (cf. Section 5.4.3).



(a) (b)

Figure 5.5: Thickness distribution in the case of: a) 2 clusters and b) 3 clusters. High average thickness and low standard deviation is a measure of process efficiency and product quality. The production runs in the "purple" cluster demonstrate superior quality characteristics.

3. Shapley value approach: The importance of input features for classification or regression can be assessed using Shapley values (cf. Section 5.4.5).

Combining clustering and subject matter expertise

When two clusters are considered (Fig. 5.5a), cluster 0 demonstrates superior characteristics, with the highest average thickness and the lowest standard deviation (Table 5.2). Further examination reveals that cluster 0 is characterized by production runs predominantly using recipe version $V21$, while cluster 1 comprises runs using version $V20$ and older versions, indicating recipe version as the main distinguishing feature.

Table 5.2: Characteristics of each cluster in the case of two clusters. The recipe version used for production is the discerning feature of the two clusters.

Cluster	μ_{thick} (µm)	σ_{thick} (µm)	Predominant recipe versions
0	16.35	1.355	$V21$
1	15.08	1.578	$V20$ & older

When three clusters (Fig. 5.5b), are identified by the clustering algorithm, cluster 0 exhibits superior characteristics, with the highest average thickness and the lowest standard deviation (Table 5.3). In particular, cluster 0 comprises production runs using recipe version $V21$, and is practically the same as cluster 0 in the two-cluster case mentioned in the previous paragraph. Clusters 1 and 2 predominantly use $V20$ and older versions and are the result of the splitting of cluster 1 identified in the two-cluster case. This cluster splitting, in essence, means that even among production runs using recipe version $V20$ and older, there are certain cases where favorable quality characteristics are achieved. This raises the question: which is the critical input that led to this difference in quality?

Further assessment drew our attention to an engineered feature, the absolute value of the difference between the nominal and actual total surface area to be coated. The nominal surface area is the predetermined production setting, specified for increments of $1\mathrm{m}^2$. It is selected to be as close as possible to the actual total surface area of the to-be-coated inserts within the reactor. In practice, this nominal surface area rarely matches the actual value of the total surface area and this discrepancy is evident when comparing the distributions between clusters 1 and 2, as shown in Fig. 5.6; On average, for the members of cluster 2, the difference between the nominal and actual total surface area is greater than $0.5\mathrm{m}^2$, while in cluster 1 it is less than $0.5\mathrm{m}^2$. This analysis suggests that when the value of this difference is less than $0.5\mathrm{m}^2$, the qualitative characteristics of the products are superior, thus leading to a clear and cost-free improvement suggestion: define preset production parameters for increments of $0.5\mathrm{m}^2$ (instead of $1\mathrm{m}^2$) of the total surface area.

Table 5.3: Characteristics of three clusters: Discerning features include the recipe version used for production and the absolute difference between nominal and actual surface area.

Cluster	μ_{thick} (µm)	σ_{thick} (µm)	Predominant recipe versions	|Nominal recipe surface area - actual surface area| (cm^2)
0	16.35	1.354	$V21$	4892
1	15.53	1.386	$V20$ & older	4628
2	14.32	1.588	$V20$ & older	5526

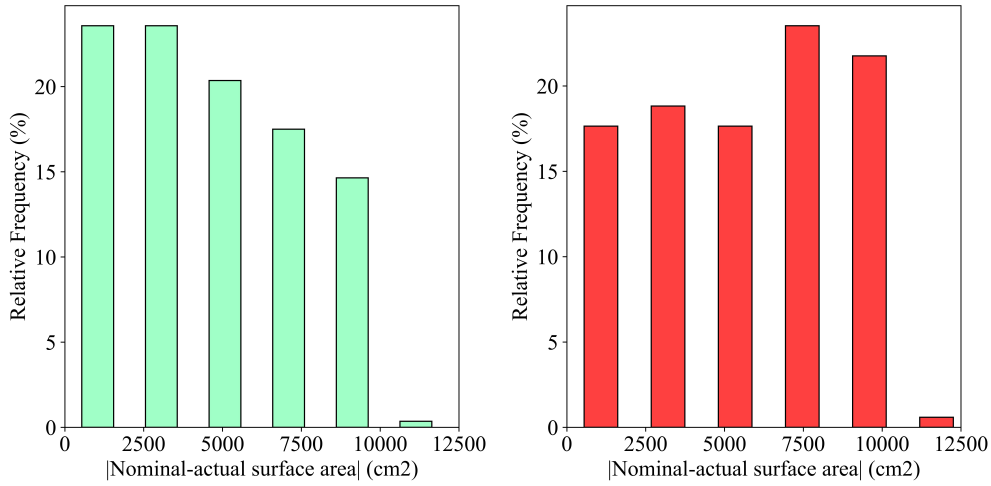


Figure 5.6: Distributions of |Nominal recipe surface area - actual surface area| for clusters 1 (in green) and 2 (in red). Cluster 2 includes relatively more observations with values larger than 5000 cm^2 when compared with cluster 1.

5.4.3 Classification

We train a classifier to predict cluster labels that resulted from the clustering analysis, using as inputs the dominant features identified in the previous section (clustering). This is useful in practice to predict the overall quality characteristics of the production run, as these cluster labels correspond to distinct thickness distributions.

The results for a binary (two-cluster case) and a multi-label classification (three-cluster case) task are presented. For these tasks, we divide the 603 observations into a training set and a test set using an 80/20 ratio.

Initially, classification models take as input the two important features identified through clustering. However, these are not the only discernible differences between clusters; other features, such as the year of production, the reactor used, and the standard deviation of the surface area in the reactor (i.e., the variation of the insert surface area of each disk inside the reactor), also have marked differences between clusters. Therefore, these inputs are also considered when training the classifier.

The initial step involves training a random forest classification model (*n_estimators=1000,*

max_depth=6) to predict whether a production run belongs to cluster 0 or 1 in Fig. 5.5a, treating it as a binary classification problem. The classifier, as shown in the confusion matrices in Figs. 5.7a and 5.7b, accurately distinguishes between clusters 0 and 1 production runs both for the training (accuracy $= 0.954$) and test set (accuracy $= 0.958$). The calculated accuracy, f1 score, precision, and recall metrics are presented in detail in Table 5.4.

Subsequently, a random forest classification model (*n_estimators=1000, max_depth=6*) is developed to determine if a production run belongs to cluster 0, 1, or 2 in Fig. 5.5b, making it a multi-label classification problem. As demonstrated in the confusion matrices in Figs. 5.7c and 5.7d, the classifier identifies cluster 0 members very accurately, for both training and test datasets. However, it sometimes struggles to distinguish between members of cluster 1 and cluster 2, often misclassifying them as members of the other cluster. The accuracy of the classifier on the test set is 0.793. As in the two-cluster case, all metrics are presented in Table 5.4. Since this is not a binary classification problem, the f1 score, precision and recall metrics are macro-averaged [229].

Table 5.4: Classification metrics for the two-cluster and three-cluster cases. The metrics for the three-cluster case have been macro-averaged.

	Accuracy	f1	Precision	Recall
2-cluster case				
Training Set	0.968	0.958	0.954	0.962
Test Set	0.967	0.958	0.958	0.958
3-cluster case (macro-averaged metrics)				
Training Set	0.840	0.848	0.844	0.852
Test Set	0.793	0.792	0.795	0.790

5.4.4 Regression

In the present work, regression is used as a tool that allows for the prediction of the average coating thickness for each production run, using fewer measurements than the 15 currently used. Specifically, we use the features identified through clustering and five thickness measurements (the closest to the reactor's inlet (R_0)) as inputs. This leads to accurate prediction of the mean coating thickness (average of $R_{1/2}$ and R) for both training ($R^2 = 0.914$) and the test set ($R^2 = 0.722$)

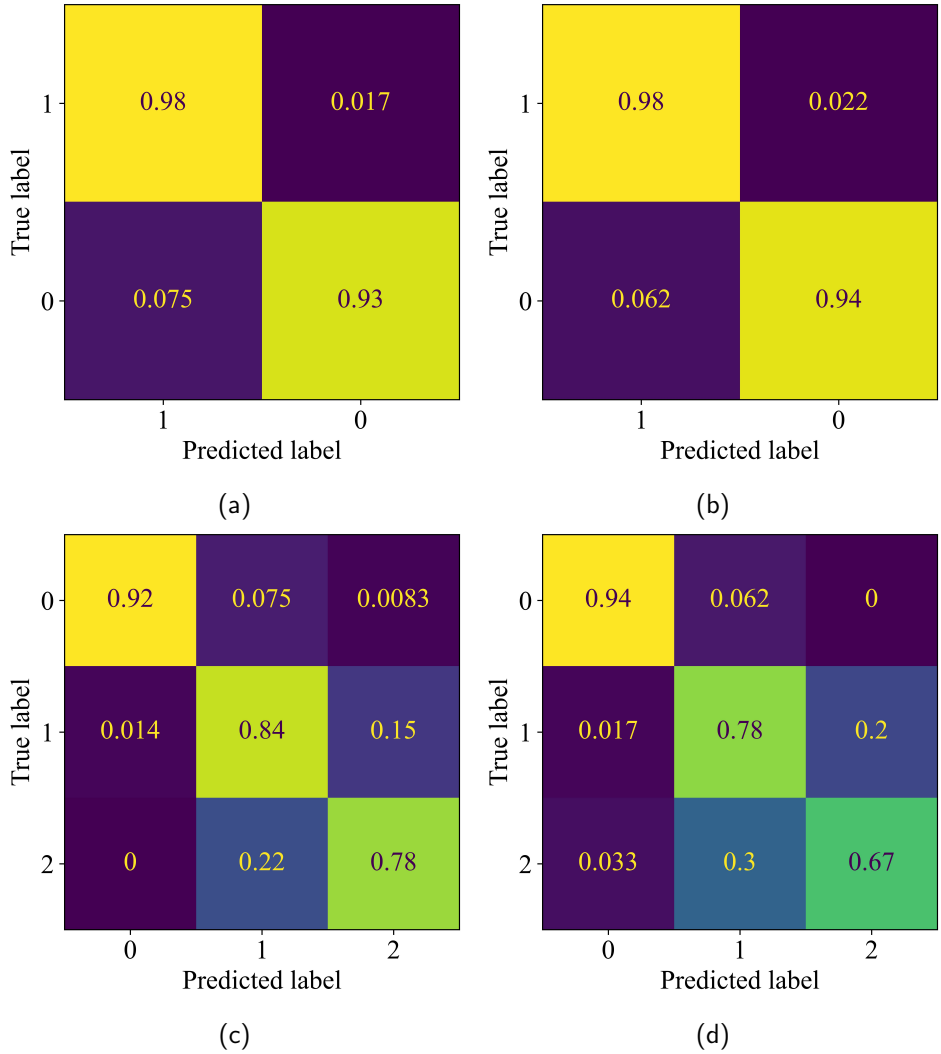


Figure 5.7: Confusion matrices for (a),(c) the training set and (b),(d) the test set of the two-cluster and three-cluster classification cases, respectively.

(cf. Fig. 5.8). This method proves valuable for streamlined post-production quality control as it allows for precise quality assessment with only one third of the previously required measurements.

5.4.5 Shapley value analysis

The most influential features that affect the predicted average coating thickness are identified by computing the SHAP values for the developed regression model. To create a measure of "global" feature importance, we calculate the mean absolute SHAP values for each input. These values are

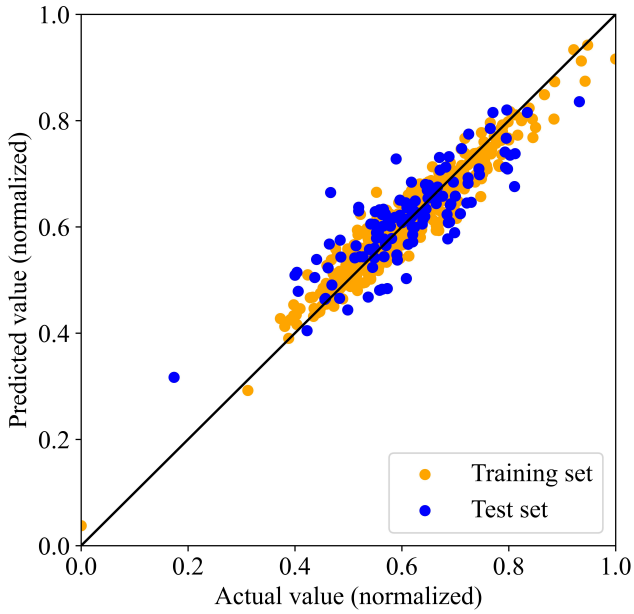


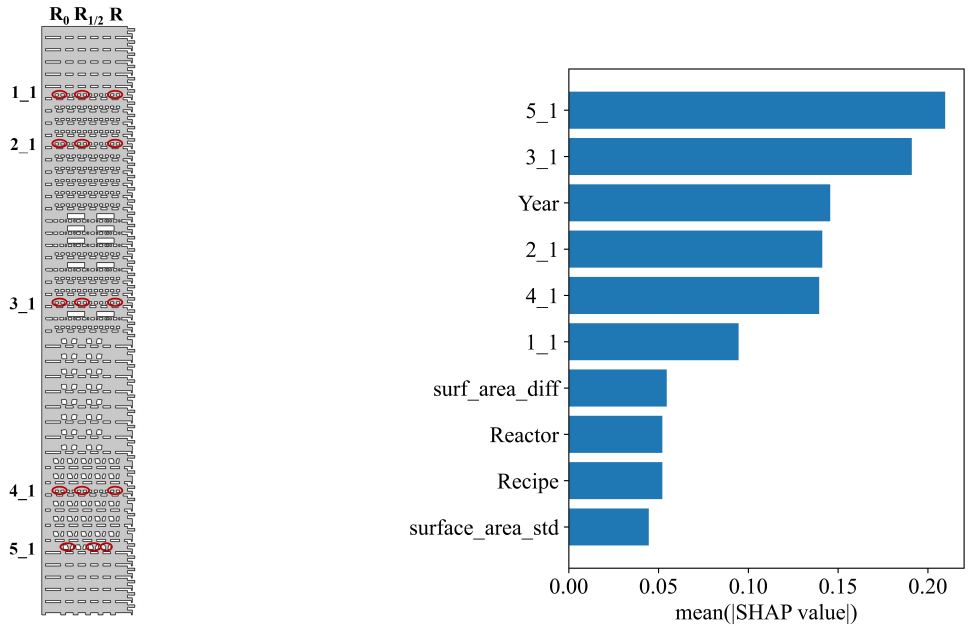
Figure 5.8: Training set performance metrics: MSE: 0.067, MAE: 0.198, R^2: 0.914, MAPE: 1.26%. Test set performance metrics: MSE: 0.264, MAE: 0.409, R^2: 0.722, MAPE: 2.62%.

shown in Fig. 5.9b. The five thickness measurements provided along with the year of production emerge as the most crucial features. Of the five thickness measurements provided, the lowest contribution comes from the measurement on the first disk from the top of the reactor. They are followed by the four remaining features, i.e., recipe, difference between the nominal and actual substrate surface area within the reactor (surf_area_diff), standard deviation of the surface area (surface_area_std) and the reactor used for production. These four features demonstrate a similar contribution to the model's predictions.

5.5 Conclusions

This study introduces a data-driven approach for uncovering patterns and influencing process inputs in an industrial Chemical Vapor Deposition (CVD) process, addressing challenges associated with process complexity and dataset characteristics.

Our analysis relies on subject matter expertise, combined with supervised and unsupervised learning methods. The main premise is that the performance of data-driven algorithms, given



(a) (b)

Figure 5.9: (a) 2D representation of the reactor, indicating the positions of the thickness measurements used as inputs for the regression problem. (b) Calculated mean absolute SHAP values for each of the inputs to the regression model. The five provided thickness measurements along with the year of production appear to be the dominant features, followed by surf_area_diff, the reactor and the recipe used for production and the standard deviation of surface area within the reactor.

a specific data set, is influenced by and indicative of the importance of the inputs used during training. This is supported here by intuition about critical process inputs and *some* knowledge about the important quality characteristics.

We use unsupervised learning to obtain meaningful data labels that correspond to groups of production runs of similar quality. We then use these labels, in the context of supervised learning, to predict the outcome for a new set of inputs, thus providing a cost-efficient shortcut for quality control.

The importance of features in investigated using Shapley values, which corroborates both subject matter expertise and also the conclusions drawn from the accuracy of classification methods. The results of this study offer opportunities to streamline post-production quality control and contribute to the ongoing refinement of the manufacturing process.

It is worth noting that this framework is adaptable to other processes, contingent on data availability. Even in cases with limited data, this approach unveils potential process-determining inputs, corroborating the insights of process experts in a purely data-driven manner.

Furthermore, consistent and improved data collection in the coming years will not only aid in validating and enhancing the developed predictive models, but also contribute to the continuous optimization of the overall process.

Chapter 6

A machine learning framework for analyzing the impact of reaction center configurations on astrocyte metabolic states

This chapter is reproduced from an article currently under review, reproduced from the preprint available here: P. Papavasileiou, S. Farina, E. D. Koronaki, A. G. Boudouvis, S. P. A. Bordas, and A. Skupin, *Machine Learning-based Predictions of Spatial Metabolic Profiles Demonstrate the Impact of Morphology on Astrocytic Energy Metabolism*, 2024. DOI: 10.1101/2024.09.18.613725.

The work presented in this article was an interdisciplinary collaboration with Dr. Sofia Farina and Prof. Alexander Skupin from the Integrative Cell Signaling Department of the Luxembourg Centre for Systems Biomedicine (LCSB). As the first author of the article, I participated in the development of the proposed ML models, the analysis of the data provided by Dr. Farina, the validation and visualization of the results, and the writing of the manuscript.

According to the Contributor Role Taxonomy (CRediT), this corresponds to the following roles: Investigation, Methodology, Formal analysis, Validation, Visualization, and Writing - original draft.

Abstract

This work introduces a machine learning framework that allows the investigation of the influence of reaction centers on the metabolic state of astrocyte cells. The proposed ML framework takes advantage of spatial astrocyte metabolic data stemming from numerical simulations for different reaction center configurations and allows for the following: (i) Discovery of cell groups of similar metabolic states and investigation of the reaction center configuration within each group. This approach allows for an analysis of the importance of the specific location of the reaction centers for a potentially critical metabolic state of the cell. (ii) Qualitative prediction of the energetic state of the cell (based on [ATP]: [ADP]) and quantitative prediction of the metabolic state of the cell by predicting the spatial average concentration of the metabolites or the complete spatial metabolic profile within the cell. (iii) Finally, the framework allows for the post hoc analysis of the developed quantitative predictive models using a SHAP approach to investigate the influence of the reaction center positions for further support of the insights drawn in steps (i)-(iii). Following the implementation of the framework, we observe that a uniform mitochondrial distribution within the cell results in the most robust energetic cell state. On the contrary, realizations of polarized mitochondrial distributions exhibit the worst overall cell health. Furthermore, we can make accurate qualitative predictions regarding cell health (accuracy $= 0.9515$, recall $= 0.9753$) and satisfactory predictions for the spatial average concentration and spatial concentration profiles of most of the metabolites involved. The techniques proposed in this study are not restricted to the dataset used. They can be easily used in other datasets that include findings from various metabolic computational models.

6.1 Introduction

Understanding the complex interplay of molecules within cells is crucial for advancing fields such as medicine, biotechnology, and pharmacology. In the intricate landscape of cellular biology, metabolism is a complex series of interconnected pathways occurring in living cells. It operates through specific biochemical reactions and produces energy and other essential biochemical compounds. Energy in the form of ATP is the fuel of all living systems, and metabolism is designed

to optimally regulate it. Metabolism is thus a prerequisite for the optimal function and survival of cells and, in extension, for the survival of organisms. An example of particularly important cells are astrocytes (cf. Fig. 6.1), the most abundant glial cells and crucial energetic supporters of the energy-intensive brain [230]. In general, the study of cellular metabolism has evolved significantly, with researchers now employing a multidisciplinary approach that integrates both biological experiments and computational modeling. Given the intricate nature of metabolism, employing mathematical models is essential for a comprehensive investigation [231].

Metabolic processes are not uniformly distributed throughout the cell. Subcellular compartments such as mitochondria, endoplasmic reticulum, and cytoplasm exhibit distinct metabolic activities. In astrocytes, the enzymatic distribution of hexokinase seems to be fundamental for glucose uptake [232], while the location of mitochondria appears to be crucial for calcium activity [233]. Thus, the current research direction aims to include spatial cellular information [234] to spatially quantify metabolites and their dynamics over time. Recent advances in analytical techniques have contributed to obtaining a snapshot of the cellular status. For example, spatial metabolomics [235] aims to identify and analyze metabolites directly within their -usually- geometrically complex spatial surroundings. Imaging and image analysis techniques have been proven to be useful in investigating spatio-temporal intracellular ATP and cellular morphological changes [236], [237]. In addition, these spatially resolved data contribute to a more comprehensive understanding of how metabolic processes are compartmentalized and coordinated within the cell.

Complementing the analytical techniques mentioned above, computational approaches have become a valuable tool for unraveling the complexity of cellular metabolism. Classical metabolic modeling approaches range from stoichiometric models [238] to kinetic simulations [239]–[241]. These models are capable of predicting and simulating the dynamics of the metabolic system. In addition, they can help guide experimental design and generate hypotheses. The main limitation of these models is the assumption of a well-mixed cellular environment, which neglects the spatial heterogeneity present in the biological systems that we discussed above. Several recent computational models have proposed spatially resolved kinetic models and agent-based simulations [242], [243] exploring how metabolite concentrations can vary in spatial dimension in different cellular morphologies of varying geometric complexity. This modeling approach is particularly valuable as

it approaches biological reality and is well suited for the study of phenomena such as organelle crosstalk and the impact of spatial constraints on metabolic fluxes [40], [244]–[247]. Biological snapshots of metabolite concentrations obtained from *in vivo* and *in vitro* cells offer valuable glimpses into cellular states [248] and can be used as starting points for spatially resolved models. Although these data can characterize the cellular state at the moment they are collected, they are unable to capture the dynamic nature of cellular metabolism. Moreover, there is a limit to the data that can be collected from a cellular sample: staining a cell to gain information on one metabolite can prevent the investigation of another. Lastly, the lack of comprehensive data on the temporal aspects of metabolic processes hinders the ability of computational models to accurately simulate and predict the real-time behavior of cellular metabolism. Bridging these gaps in both experimental snapshots and computational modeling data is essential for understanding the intricate dynamics that govern cellular metabolic networks.

Addressing the limitations in our current understanding of cellular metabolism, machine learning techniques [249], [250] could be applied to bridge the gap between static biological snapshots and dynamic models. For this purpose, we implement a machine learning approach on a dataset consisting of the results of a spatially resolved computational metabolic model of an astrocyte [40]. This computational model provides us with spatial information of the metabolites in a simplified two-dimensional rectangular cellular domain, given different configurations of reaction centers in the form of coordinates on the x- and y-axes.

The proposed approach aims to discover reaction centers (inputs) that are potentially critical to the metabolic state of the cell (output). To this end, the following steps are necessary: a) Discovery of groups of similar metabolic profiles, using *only* the output of the computational model (spatial metabolic concentrations at steady state). This is achieved through the use of clustering algorithms. By analyzing the inputs corresponding to the resulting clusters, we can draw insights into the relationship between the reaction center position and the metabolic state of the cell. b) Qualitative prediction of cell health status using *only* the coordinates of the reaction centers as input. This is made possible through the use of classification algorithms. The input-output relationship insights derived from the previous step are expected to greatly influence the predictions of the classification algorithm used. c) Quantitative prediction of the metabolic state of the cell

using the coordinates of the reaction centers as inputs. This is enabled by the use of regression algorithms, in our case, Artificial Neural Networks (ANNs). We are able to predict both the average concentration of the metabolites in the domain and the spatial profile of the metabolites in the domain at steady state. Last but not least, d) since the explainability of the developed "black-box" ANN models is very important for our application, a SHAP analysis [37] can be performed for the developed regression models. The obtained values can shed more light on the effect of each input (reaction center coordinates) on model output (spatial metabolite concentration) and further indicate whether the insights derived in the previous steps are meaningful.

These techniques have the potential to decode the complexity inherent in cellular metabolism, offering a means to generate more comprehensive and accurate representations of metabolic processes. Hopefully, they can also be applied to spatially resolved computational models of higher metabolic or geometric complexity, which are predominant when it comes to cells.

6.2 Methods

6.2.1 Computational Model

Biochemical Reaction Model

We consider a spatially resolved metabolic model, proposed in [39], [40], which prioritizes the arrangement of the reaction sites in the domain and the geometries of the domain at the expense of a more elementary chemical model.

In its simplicity, the model captures the main fundamental metabolic energy pathways in five chemical reactions: glycolysis, mitochondrial activity, and lactate dehydrogenase. Glycolysis is described by two chemical reactions named HXK and PYRK. The first one accounts for the enzymes: hexokinase, phosphoglucose isomerase, phosphofructose kinase, and fructose bisphosphate aldolase. HXK consumes glucose (GLC) and adenosine triphosphate (ATP) producing adenosine diphosphate (ADP) and glyceraldehyde (GLY). The second reaction, PYRK, uses the product of the first reaction to produce ATP and pyruvate PYR. Now, PYR can either be used by the lactate dehydrogenase enzyme (LDH) to produce lactate (LAC) or enter the mitochondria and contribute to mitochondrial activity. Mitochondrial activity accounts for the Krebs cycle and oxida-

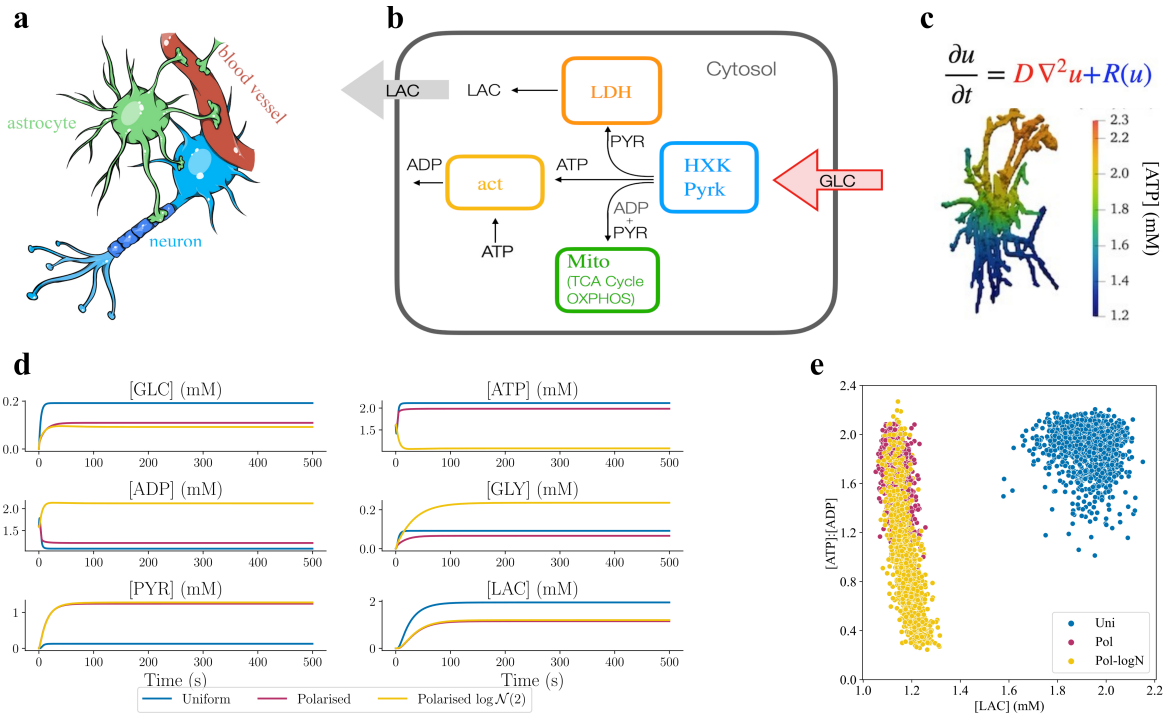


Figure 6.1: (a) A sketch illustrating the crucial location of an astrocyte between a neuron and a blood vessel, relevant for their metabolism. (b) A concise overview of the metabolic model used to describe the main pathways in astrocytes. (c) An example of $[\mathrm{ATP}]$ concentration profile obtained solving the metabolic model in a 3D human astrocyte obtained from a confocal microscopy image. (d) Dynamic evolution of the six considered metabolites averaged inside the 2D domain for three sampled realizations, one for each distribution of reaction centers. (e) Scatter plot of the spatial average $[\mathrm{ATP}] : [\mathrm{ADP}]$ vs spatial average $[\mathrm{LAC}]$ for all available cell configurations at steady state - Clear distinctions between uniform configurations and polarized and polarized log-normal configurations.

tive phosphorylation, producing ATP. Finally, the energetic production within the cell is balanced by considering the cellular activity that consumes ATP.

The chemical model is then described as follows:

$$\mathrm{HXK} := \mathrm{GLC} + 2\,\mathrm{ATP} \rightarrow 2\,\mathrm{ADP} + 2\mathrm{GLY} \tag{6.1}$$

$$\mathrm{PYRK} := \mathrm{GLY} + 2\,\mathrm{ADP} \rightarrow 2\,\mathrm{ATP} + \mathrm{PYR} \tag{6.2}$$

$$\mathrm{LDH} := \mathrm{PYR} \rightarrow \mathrm{LAC} \tag{6.3}$$

$$\mathrm{Mito} := \mathrm{PYR} + 28\,\mathrm{ADP} \rightarrow 28\,\mathrm{ATP} \tag{6.4}$$

$$\mathrm{act} := \mathrm{ATP} \rightarrow \mathrm{ADP}\,, \tag{6.5}$$

An overview of the model is presented in Fig. 6.1.

Mathematical model

Mathematically, the model is then translated into a reaction-diffusion system [251] through a set of partial differential equations (PDE), which allows us to a) solve the metabolic model in a geometrical bounded domain; b) account for the molecules' diffusivity; c) distribute spatially the chemical reaction sites inside the domain.

In a bounded 2-dimensional domain, Ω, we consider a fixed number, $M \in \mathbb{R}^+$, of reaction sites for chemical reactions: HXK, PYRK, LDH and Mito, which are spatially distributed using a spatial reaction rate density, $\mathcal{K}_j$. Spatial reaction rates are defined as the product between classical reaction rates K_j, and Gaussian functions defined with a center $\{\mathbf{x}_i\}_{i=1}^{M} \in \Omega$ and variance $\sigma_i \in \mathbb{R}^+$. The cellular activity, act, operates homogeneously in the domain Ω with reaction rate K_{act}.

The reaction-diffusion system is defined as follows:

$$
\begin{cases}
\frac{\partial[\mathrm{GLC}]}{\partial t} = & D_{[\mathrm{GLC}]}\nabla^2[\mathrm{GLC}] - \mathcal{K}_{\mathrm{HXK}}[\mathrm{GLC}][\mathrm{ATP}]^2 + J_{\mathrm{in}} \\
\frac{\partial[\mathrm{ATP}]}{\partial t} = & D_{[\mathrm{ATP}]}\nabla^2[\mathrm{ATP}] - 2\mathcal{K}_{\mathrm{HXK}}[\mathrm{GLC}][\mathrm{ATP}]^2 + 2\mathcal{K}_{\mathrm{PYRK}}[\mathrm{ADP}]^2[\mathrm{GLY}] \\
& +28\mathcal{K}_{\mathrm{Mito}}[\mathrm{PYR}][\mathrm{ADP}]^{28} - K_{\mathrm{act}}[\mathrm{ATP}] \\
\frac{\partial[\mathrm{ADP}]}{\partial t} = & D_{[\mathrm{ADP}]}\nabla^2[\mathrm{ADP}] + 2\mathcal{K}_{\mathrm{HXK}}[\mathrm{GLC}][\mathrm{ATP}]^2 - 2\mathcal{K}_{\mathrm{PYRK}}[\mathrm{ADP}]^2[\mathrm{GLY}] \\
& +K_{\mathrm{act}}[\mathrm{ATP}] - 28\mathcal{K}_{\mathrm{Mito}}[\mathrm{PYR}][\mathrm{ADP}]^{28} \\
\frac{\partial[\mathrm{GLY}]}{\partial t} = & D_{[\mathrm{GLY}]}\nabla^2[\mathrm{GLY}] + 2\mathcal{K}_{\mathrm{HXK}}[\mathrm{GLC}][\mathrm{ATP}]^2 - \mathcal{K}_{\mathrm{PYRK}}[\mathrm{ADP}]^2[\mathrm{GLY}] \\
\frac{\partial[\mathrm{PYR}]}{\partial t} = & D_{[\mathrm{PYR}]}\nabla^2[\mathrm{PYR}] + \mathcal{K}_{\mathrm{PYRK}}[\mathrm{ADP}]^2[\mathrm{GLY}] - \mathcal{K}_{\mathrm{LDH}}[\mathrm{PYR}] \\
& -\mathcal{K}_{\mathrm{Mito}}[\mathrm{PYR}][\mathrm{ADP}]^{28} \\
\frac{\partial[\mathrm{LAC}]}{\partial t} = & D_{[\mathrm{LAC}]}\nabla^2[\mathrm{LAC}] + \mathcal{K}_{\mathrm{LDH}}[\mathrm{PYR}] - \eta_{\mathrm{LAC}}[\mathrm{LAC}]
\end{cases}
\tag{6.6}
$$

where:

- The source of GLC is described through a function $J_{\mathrm{in}} : \Omega \times [0,T] \to \mathbb{R}$:

$$
J_{\mathrm{in}}(x,t) = \begin{cases} \alpha \in \mathbb{R} \quad \text{if} \quad (x,t) \in \Omega_{\mathrm{in}} \times [0,T], \quad \text{where} \quad \Omega_{\mathrm{in}} \subset \Omega \\ 0 \quad \text{otherwise.} \end{cases} . \tag{6.7}
$$

- The degradation of LAC, which is proportional to the amount of LAC in region $\Omega_{\mathrm{out}} \subset \Omega$ is described by the function $\eta_{\mathrm{LAC}} : \Omega \times [0,T] \to \mathbb{R}$

$$
\eta_{\mathrm{LAC}}(x,t) = \begin{cases} \eta \in \mathbb{R} \quad \text{if} \quad (x,t) \in \Omega_{\mathrm{out}} \times [0,T], \quad \text{where} \quad \Omega_{\mathrm{out}} \subset \Omega \\ 0 \quad \text{otherwise.} \end{cases} . \tag{6.8}
$$

For more details on the mathematical model and parameters, we refer the readers to [40].

6.2.2 Data acquisition

Data were acquired by numerically solving the reaction-diffusion system that arose from the metabolic model presented in the previous section. We used standard finite element methods

[252] using the FEniCS software [253]. First, the reaction-diffusion system was converted to its corresponding weak form. Then, we spatially discretize the two-dimensional rectangular domain by finite elements using the package *mshr* (number of finite elements 25298 and number of dofs 13207). We temporally discretize the time derivative using a backward Euler scheme with a time step of 0.15 (s) [254]. The solution of the weak form is defined on the space of piecewise Lagrangian finite elements of degree one.

We consider a 2-dimensional rectangular domain ($[0, l] \times [0, L]$, with width $l = 4$ µm and length $L = 140$ µm) where we place 10 reaction sites per chemical reaction with a spatial extent of $\sigma = 1.0$ µm. The input/inlet of the system is the entrance of GLC in the bottom left corner, while the output/outlet is the outflux of LAC in the opposite corner. To investigate the crucial role of spatial arrangement in cellular domains, we consider three possible distributions of the reaction sites: uniform, polarized, and polarized log-normal. The uniform distribution considers the 10 reaction sites per chemical reaction to be sorted from uniform distributions. The polarized consider an extreme reaction site configuration supposing that glycolysis is located at the bottom of the rectangular domain close to the GLC influx while the 10 reaction sites for LDH are sorted at the top of the rectangular domain. The main difference between polarized and polarized log-normal lies in the distribution of the mitochondria. In the first case, six reaction sites for Mito are placed where glycolysis is located, and four reaction sites are at the top of a part of the rectangular domain, to ensure that some mitochondria can be found throughout the domain. The polarized log-normal setting uses a log-normal distribution to sort the 10 Mito reaction sites, causing mitochondria to be located mainly in the lower part of the domain and almost none co-located with LDH. Examples of the three distributions can be seen in Figure 6.2.

The dataset used in this study is composed of 1,428 uniform, 1,336 polarized, and 1,314 polarized log-normal realizations. The information for each realization are the x and y location of the reaction center sites, the average concentration of the six metabolites in the domain at steady state and the spatial concentration at each grid point inside the discretized domain for the six metabolites at the steady state.

6.2.3 Unsupervised learning

Unsupervised learning algorithms process unlabeled data to discover interesting patterns within the data. For instance, they might perform an association rule analysis or create clusters of similar observations in a dataset [82]. Furthermore, dimensionality reduction techniques such as the widely used Principal Component Analysis (PCA), autoencoders [224], and diffusion maps [220], [221] —also fall under the umbrella of unsupervised learning since they provide a reduced data representation without considering the corresponding response variable (or label) of the data.

In the upcoming sections, we will provide a concise overview of the clustering and dimensionality reduction techniques that have been implemented.

Clustering

For our clustering study, which aims to discover groups of cells that demonstrate similar metabolic profiles, we implement an *agglomerative* hierarchical clustering algorithm [87], [88]. Agglomerative clustering starts with a number of clusters equal to the number of observations and progressively merges clusters until a single cluster remains. The way these clusters merge is based on the dissimilarity metric and the linkage criterion used. Here, the Euclidean distance is implemented as the dissimilarity metric and a Ward linkage criterion [227]. This criterion minimizes the total variance within the cluster by merging the clusters in a way that leads to the smallest increase in variance after each merge. Specifically, it aims to minimize the sum of squared differences within all clusters. The scikit-learn AgglomerativeClustering module is used for this task [255]. Agglomerative hierarchical clustering is selected because it provides insight on how the data merges as the number of clusters changes. This information is readily available in the form of a dendrogram, such as the one presented in Fig. 6.3a.

Dimensionality reduction

Each row spatial concentration data matrix realization has 79,242 columns, one for each of the six metabolite concentration values at each of the 13,207 grid points. Reducing the dimensionality of the spatial data will be very beneficial when manipulating the data and when training predictive models later on. For this task, Principal Component Analysis (PCA) is implemented [112]. PCA

linearly transforms the original data onto a new coordinate system where PCs can be easily identified. The amount of PC retained for subsequent analysis depends on user criteria. In this work, we aim to retain 99.9% of the variance and to keep the reconstruction Root Mean Square Error (RMSE) below 0.03. The scikit-learn PCA module is used for this task [255].

6.2.4 Supervised learning

Supervised learning algorithms, in contrast to unsupervised ones, are based on labeled data. In labeled data, the features (x_i) are associated with the corresponding responses (y_i). These models use available data to make predictions for future observations. Supervised learning encompasses regression for continuous variables and classification for binary or ordinal responses [81].

In the present work, the coordinates of the 40 metabolite reaction centers can be considered as inputs, with the response variables being the spatial metabolite concentrations or the spatial average metabolite concentrations. Based on these continuous variables, binary variables can also be engineered (healthy vs. non-energized cell) for classification purposes.

The methods evaluated for this work include: (a) linear methods: for regression, lasso [89], and ridge [90] regression, and logistic regression for classification tasks. (b) Support vector machines (SVMs) [91] that can be categorized as linear or nonlinear methods based on the kernel used for classification tasks. (c) Tree-based methods: involving classification and regression trees [92] and their ensemble counterparts such as random forests [93], gradient-boosted trees [94], extra trees [95], and XGBoost [34], which combine numerous trees to improve performance [96]. (d) Artificial neural networks (ANN), whose diverse architectures [97] can provide valuable options for both classification and regression tasks.

In the present work, logistic regression, random forests, SVM, extra trees, gradient-boosted trees, XGBoost, and ANNs are implemented for supervised learning tasks. However, results are presented only for methods that demonstrate the best performance for our dataset, namely logistic regression for classification tasks and ANNs for regression tasks. An overview of the supervised learning approaches used in this work is presented in Fig. 6.2c.

Logistic regression is implemented for the classification of cells as healthy or non-energized, based solely on the coordinates of the metabolite reaction centers. The scikit-learn LogisticRe-

gressionCV module is used for this task [255]. This module also has the added benefit of including cross-validation and hyperparameter optimization in the training process, thus reducing overfitting.

In terms of the regression tasks of this work, Artificial Neural Networks (ANNs) are implemented for the prediction of spatial concentration profiles and the spatial average concentrations of the metabolites. The TensorFlow [98] and Keras Python libraries [99] are used for the development and training of the ANN models in this work.

As the performance of ANNs is significantly influenced by their architecture, optimizing the architecture during the training process is crucial. To achieve this, a Bayesian optimization approach, based on the work of [130] is employed for hyperparameter tuning in each ANN model. For this task, we use the keras-tuner Python library [256]. Similarly to other optimization methods, Bayesian optimization aims to find optimal values for bounded parameters (hyperparameters in our case), denoted $x_1, x_2, ..., x_n \in X$ that minimize an objective function $f(X)$ (equivalent to the loss function of the neural network). In Bayesian optimization, a probabilistic model is constructed for $f(X)$, which allows us to identify the best points in X for evaluating $f(X)$ in subsequent steps. Unlike local gradient-based methods, this approach considers all available information about $f(X)$ [130].

6.2.5 SHAP analysis

Shapley values, originally introduced by Shapley [212] in the field of game theory and proposed as a tool to analyze machine learning models in [37], intricately assess the average contribution of each feature's value to predictions. In this way, they provide an understanding of how perturbations of a variable can influence the output of the model, thus shedding light on models that have traditionally been considered "black boxes".

In this work, a SHAP analysis is performed on 3 of the ANN regression models developed for the prediction of spatial average metabolite concentrations. The final goal is to discover not only the inputs have the most influence on model output, but also the type of influence they have. However, the results of this SHAP analysis only provide information about the relationships between inputs and *model outputs*. These results shed light on previously "opaque" ML methods but should not be used to make causal claims about input/output relationships.

6.3 Results

As we are aiming to discover different groups of similar realizations, our analysis starts with the clustering of our realizations, focusing only on the outputs (concentration profiles). After clustering, we will explore the characteristics of each cluster and analyze the differences in the inputs corresponding to each cluster with the goal of finding differences between the clusters. These different inputs could be potentially important for the differences in the clusters and in extension to the differences in the outputs (concentration profiles) on which the clustering was based. As we will show, this initial clustering approach can enable a lot of opportunities for further data-driven approaches.

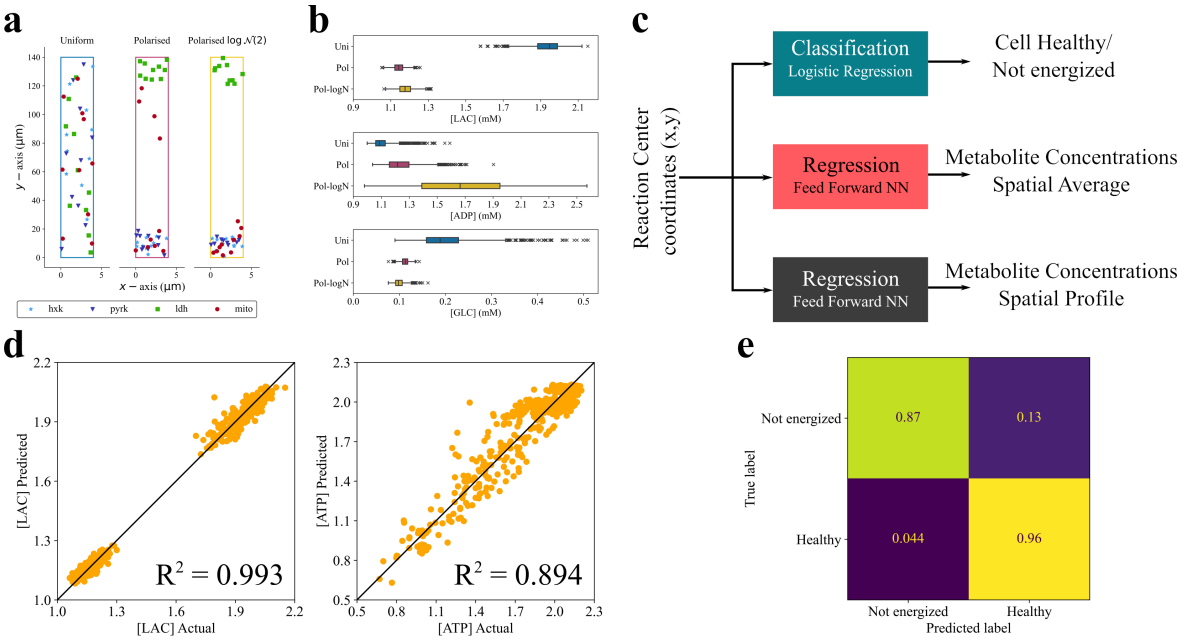


Figure 6.2: (a) Samples of the three different types of reaction center configurations: uniform, lognormal and polarized lognormal. Glucose enters the cellular domain from the origin of the axis, while lactate exits from the opposite vertex. (b) Distributions of three metabolites of interests for each type of reaction center configuration. (c) The three implemented supervised learning approaches. (d) Out of sample parity plots for the spatial average prediction of $[\mathrm{LAC}]$ and $[\mathrm{ATP}]$. The concentration units are mM. (e) Out of sample confusion matrix for the prediction of the energetic state of the cell (non-energized vs healthy).

6.3.1 Clustering

An agglomerative hierarchical clustering algorithm is used to discover observations with similar characteristics solely based on the results (outputs) of the computational model. The results consist of the concentration values of the six metabolites at the 13,207 grid points used for the computational simulations. A subset of the entire dataset is used for clustering. Subsequently, the input to the clustering algorithm is a matrix of dimension 1,767x79,242.

Following clustering, we attempt to analyze the resulting realization groups. We investigate a) the mean spatial averages of the six different metabolite concentrations, b) the mean $[\mathrm{ATP}] : [\mathrm{ADP}]$, and c) the distribution of the y coordinates of mitochondria in each cluster. The mean spatial averages of all six metabolites provide a great overview of the metabolic state of the cell. The mean $[\mathrm{ATP}] : [\mathrm{ADP}]$ is an excellent indicator of cell health, as cells that demonstrate a ratio lower than 1 can be considered non-energized and in a state of deterioration, while cells with a ratio greater than 1 are considered adequately energized and thus healthy. Finally, by investigating the distribution of the y-axis coordinates of the mitochondria for each cluster, we can uncover possible relationships between the metabolic state of the cell and the locations of the mitochondria. Our analysis reveals the following.

1. Based on the results of Table 6.1, it appears that cluster 1 contains healthier cells, given the fact that the average $[\mathrm{ATP}] : [\mathrm{ADP}]$ is the highest of the 5 clusters with a value of 1.910. It also appears that cluster 1 has the lowest values for $[\mathrm{GLY}]$ and $[\mathrm{PYR}]$, suggesting that this is the most efficient cluster that consumes these two substrates to maximize ATP production. Cluster 3 contains fewer energized cells, given its average $[\mathrm{ATP}] : [\mathrm{ADP}]$ of 0.435. It should also be noted that cluster 1 contains only realizations with a uniform reaction site distribution. This is also evident in Fig. 6.4d.

2. When comparing the clusters containing realizations of non-uniform reaction center distributions, we can see that for clusters 2 and 3 the mitochondria are located close to the cell inlet (and subsequently closer to the glucose entering the cell), whereas the cells for clusters 0 and 4, the mitochondria have better coverage of the spatial domain. This result is visualized in the histogram of Fig. 6.4d. This is a hint that mitochondrial distribution is a great driver

of cell health, as clusters 2 and 3 demonstrate a lower average $[\mathrm{ATP}] : [\mathrm{ADP}]$ (0.852 and 0.435, respectively) when compared to clusters 0 and 4 (1.610 and 1.388, respectively).

3. Last, solely based on the concentration values the algorithm can discern the three main groups (uniform, polarized and polarized log-normal) of cells available in the dataset. Cluster 1 contains solely uniform realizations, clusters 0 and 4 contain mostly polarized and some polarized log-normal realizations. Clusters 2 and 3 contain almost exclusively polarized log-normal realizations.

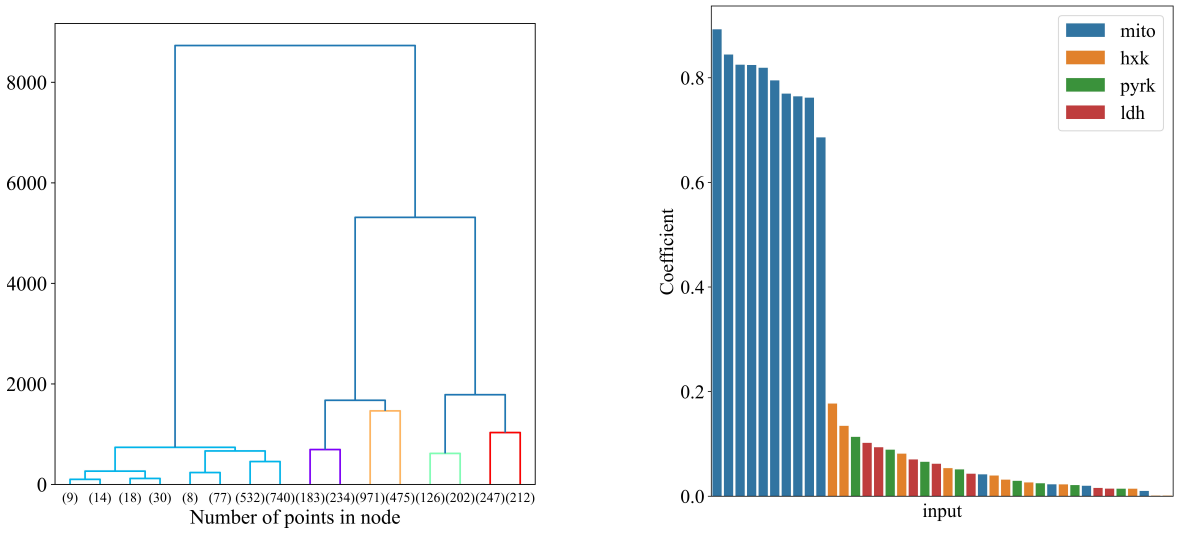


Figure 6.3: (a) Agglomerative hierarchical clustering dendrogram. (b) Top 40 logistic regression coefficients for the non-energized/ healthy cell classification problem. The y-axis coordinates of the mitochondria appear to be highly influential for the model. Essentially, the dominating coefficient values of the mitochondrial y-axis coordinates show that a better spread of mitochondria within the cell -and not close to the cell inlet- increases the probability of a healthy cell.

Following clustering, our objective is to take advantage of the information extracted in this step to create predictive models that allow for the prediction of the energy state of the cell, as well as for the prediction of the spatial metabolite concentrations inside the cell.

Table 6.1: Mean spatial metabolite concentration averages (in mM) in each cluster. Lowest and highest values for each metabolite are presented in bold.

Cluster	[GLC]	[ATP]	[ADP]	[GLY]	[PYR]	[LAC]	[ATP]:[ADP]
0	0.110	1.961	1.239	0.101	**1.260**	**1.143**	1.610
1	**0.201**	**2.097**	**1.103**	**0.079**	**0.150**	**1.935**	**1.910**
2	**0.098**	1.457	1.742	**0.159**	1.261	1.179	0.852
3	0.104	**0.962**	**2.238**	0.130	1.286	1.221	**0.435**
4	0.104	1.849	1.351	0.121	1.256	1.148	1.388

6.3.2 Discerning between Healthy and Non-energized cells

As already established, an important indicator of cell health is the spatial average of $[\mathrm{ATP}] : [\mathrm{ADP}]$ (ATP-to-ADP ratio) within the cell. When $[\mathrm{ATP}] : [\mathrm{ADP}] \geq 1$, the cell is considered adequately energized and healthy, whereas when $[\mathrm{ATP}] : [\mathrm{ADP}] < 1$ the cell is considered non-energized and unhealthy.

Given this threshold and the calculated spatial averages of $[\mathrm{ATP}] : [\mathrm{ADP}]$ for all available samples, we can convert the continuous output (ratio) to binary (health), where health $= 0$ when ratio < 1, and health $= 1$ when ratio ≥ 1. Using the available reaction center coordinates for each observation as inputs and the binary variable health as output, we can train a logistic regression model that can predict the health status of a cell, given only the coordinates of the reaction centers.

The inputs consist of the coordinates (x, y) for 40 different reaction centers (10 of each of the four types), totaling 80 inputs. Before training the logistic regression model, the inputs are centered to 0 and scaled by their standard deviation.

Of the 4,078 observations, 85% is used as a training set and 15% as a test set. Checking the performance of the model on both sets can allow us to avoid overfitting and make sure that the resulting model generalizes well in unseen data.

The resulting logistic regression model can discern between healthy and non-energized cells with high accuracy for both the training (0.9412) and the test set (0.9515). Further classification metrics, such as the f1 score, recall, and precision, are presented in Table 6.2. The confusion matrices for the performance of the logistic regression algorithm in both the training and the test set are presented in Fig. 6.2e.

Given the nature of the logistic regression algorithm, we can use the resulting coefficients to

try to make sense of the connection between the input variables and the output (cell health state). Observing Fig. 6.3b, it is evident that the y coordinates of the mitochondria highly influence the model output.

Table 6.2: Binary classification metrics. The trained logistic regression model shows great performance for both the test and training sets.

	Training Set	Test Set
f1	0.9638	0.9701
Accuracy	0.9412	0.9515
Precision	0.9561	0.9650
Recall	0.9716	0.9753

6.3.3 Predicting spatial metabolite concentrations

Although the prediction of cell health status is already quite valuable, it is often preferable to have models that can predict a continuous variable rather than a binary variable. Similarly to the classification problem, the coordinates of the 40 reaction centers can be used to predict continuous variables. These continuous values can be either the spatial averages of the six metabolite concentrations or, taking it one step further, the spatial profiles of the metabolite concentrations.

As in the binary classification problem, 85% of the 4,078 observations are used as a training set and 15% as a test set.

Spatial averages

ANN models are used for the prediction of spatial averages of the six metabolite concentrations. A different model is trained for each metabolite concentration. To ensure optimal performance, we include hyperparameter optimization in model training. The hyperparameters tested are:

1. The number of network layers (N_{layers}). Varied between 2 and 5.
2. The number of neurons per layer (N_{neurons}). Varied between 2 and 100.
3. The activation function used in all of the layers. The functions tested are: a) sigmoid, b) tanh, and c) Relu. It should be noted that the output layer always has a single neuron using a linear activation function.

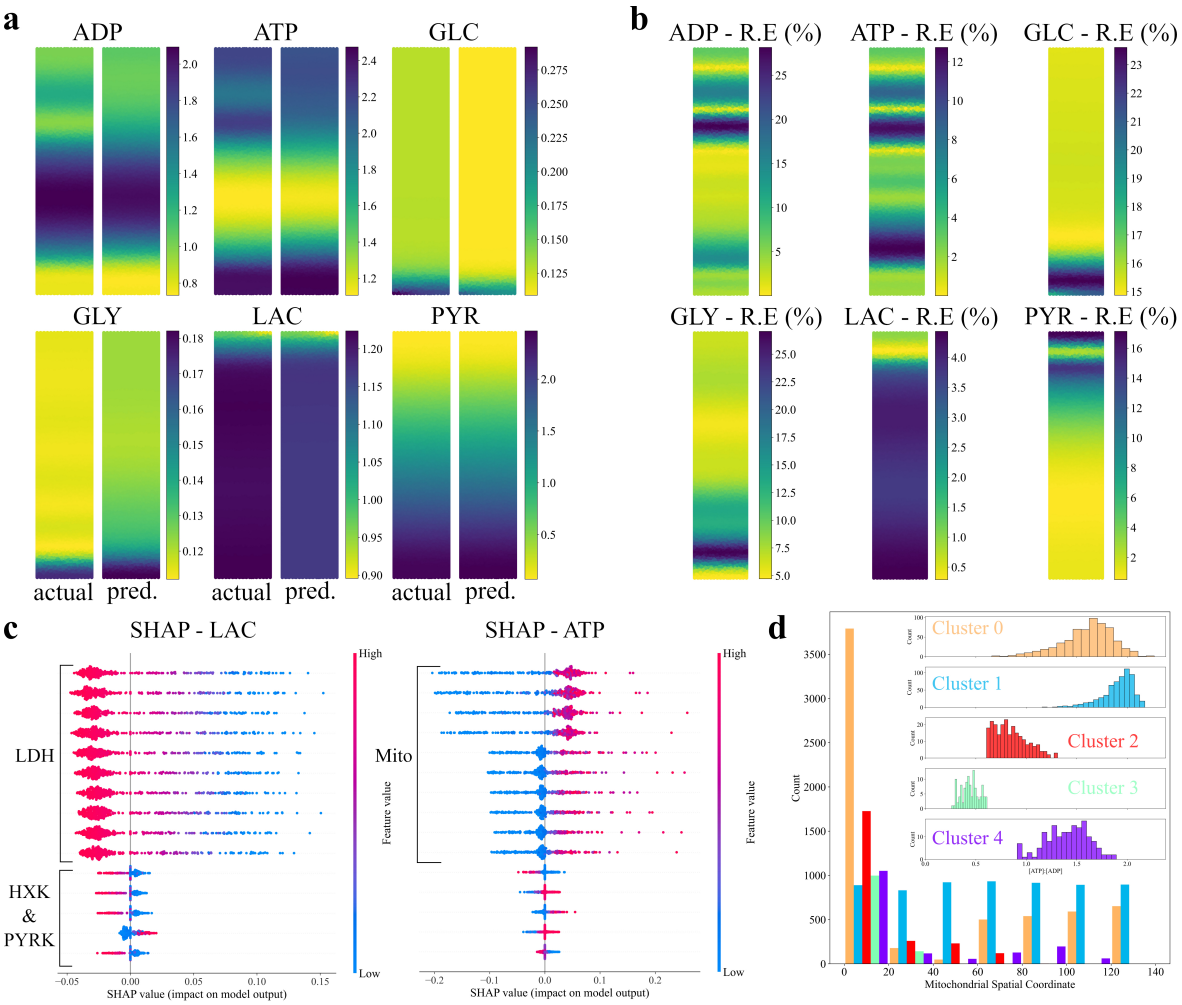


Figure 6.4: (a) Actual and predicted spatial metabolite concentration (in mM) profiles for randomly sampled, out of sample reaction center configurations. (b) Absolute relative error (%) between the actual and predicted metabolite concentrations presented in (a). (c) SHAP values for the spatial average $[\mathrm{LAC}]$ and $[\mathrm{ADP}]$ predictive models. For the $[\mathrm{LAC}]$ model, the location of the LDH reaction centers is the most important for the output of the model, followed by the locations of HXK and PYRK reaction centers. For the $[\mathrm{ATP}]$ model, the location of the mitochondrial reaction centers is the most important for the model output. (d) Histogram showing the distribution of the mitochondrial reaction center for each discovered cluster. On the top right, histograms of the $[\mathrm{ATP}] : [\mathrm{ADP}]$ for each cluster. Clusters 0, 1 and 4 appear to have more observations with $[\mathrm{ATP}] : [\mathrm{ADP}]>1$, whereas clusters 2 and 3 appear to be more problematic. When investigating the y coordinates of the mitochondria for each cluster, it is clear that for clusters 2 and 3, the mitochondria are located primarily close to the input of the cell ($y_{\mathrm{Mito}} = 0$), while for clusters 0, 1 and 4 the mitochondria have better coverage of the domain.

50% of the test set is used as a validation set during training to avoid overfitting through early stopping. This means that if the validation error does not drop after a certain number of training epochs (10 in our case), the training stops.

The resulting neural networks, along with their performance in the test set, are presented in Table 6.3. The models developed for the prediction of $[\mathrm{PYR}]$ and $[\mathrm{LAC}]$ demonstrate excellent performance with a $R^2 = 0.998$ and $R^2 = 0.993$ respectively. Satisfactory performance is observed for the predictions of $[\mathrm{ADP}]$ ($R^2 = 0.914$), $[\mathrm{ATP}]$ ($R^2 = 0.894$) and $[\mathrm{GLC}]$ ($R^2 = 0.862$). The developed $[\mathrm{GLY}]$ regression model shows a rather modest performance ($R^2 = 0.595$).

Table 6.3: Optimal hyperparameters and spatial average metabolite concentration regression test set performance metrics for the developed ANN models. Results for $[\mathrm{PYR}]$ and $[\mathrm{LAC}]$ are excellent. For $[\mathrm{GLC}]$, $[\mathrm{ATP}]$ and $[\mathrm{ADP}]$, the results are satisfactory. However, the $[\mathrm{GLY}]$ model demonstrates a slightly worse performance.

	N_{layers}	N_{neurons}	activation	R^2	MAPE	MAE	MSE
$[\mathrm{GLC}]$	5	82	sigmoid	0.862	9.21%	1.33E-02	3.77E-04
$[\mathrm{ATP}]$	4	67	sigmoid	0.894	4.63%	7.78E-02	1.14E-02
$[\mathrm{ADP}]$	4	77	sigmoid	0.914	5.11%	6.94E-02	9.24E-03
$[\mathrm{GLY}]$	3	97	sigmoid	0.595	33.89%	3.58E-02	3.66E-03
$[\mathrm{PYR}]$	5	97	sigmoid	0.998	3.62%	1.64E-02	4.59E-04
$[\mathrm{LAC}]$	5	97	sigmoid	0.993	1.49%	2.12E-02	8.60E-04

Spatial profiles

Given the satisfactory performance of the predictive models for spatial average concentrations, we can go a step further, trying to predict the spatial concentration profiles of the six metabolites in the cells given only the positions of the 40 reaction centers. Given the nature of the output vector (length of 79,242 - six metabolite concentration values for each of the 13,207 grid points), it is clear that dimensionality reduction methods can be useful for reducing the dimensions of the output space.

A model is developed for the spatial concentration profile of each metabolite. This translates to an output vector of length 13,207 (1 concentration for each grid point) for each developed model.

Principal component analysis (PCA) is used for this task, following the centering of the output

data using their mean and their scaling using their standard deviation. The number of principal components retained for each output is chosen based on the reconstruction root mean squared error (RMSE). The number of principal components retained is chosen so that the reconstruction root mean squared error (RMSE) is lower than 0.03. This leads to a different number of principal components (PCs) for each metabolite. Specifically, 17 PCs are retained for the spatial $[\mathrm{ATP}]$ and spatial $[\mathrm{ADP}]$ models, 5 PCs are kept for the spatial $[\mathrm{PYR}]$ and $[\mathrm{LAC}]$ models, 9 PCs are retained for the spatial $[\mathrm{GLY}]$ model, and finally, 11 PCs are used for the spatial $[\mathrm{GLC}]$ model.

As with the spatial average concentration predictive models, we include hyperparameter optimization when training the spatial profile models. The hyperparameters tested are:

1. The number of network layers (N_{layers}). Varied between 2 and 5.
2. The number of neurons per layer (N_{neurons}). Varied between 10 and 650.
3. The activation function used in all of the layers. The functions tested are a) sigmoid, b) tanh, c) relu, and d) elu. Note that the output layer size is equal to the number of PCs retained for the model. Furthermore, output neurons always use a linear activation function.

Once again, 50% of the test set is used as a validation set and early stopping is used during training to avoid overfitting.

The resulting neural networks, along with their performance in the test set, are presented in Table 6.4. The models developed for the spatial concentration prediction of $[\mathrm{PYR}]$ and $[\mathrm{LAC}]$ demonstrate excellent performance with $\mathrm{R}^2 = 0.956$ and $\mathrm{R}^2 = 0.972$ respectively. Reasonable performance is observed for the predictions of $[\mathrm{ADP}]$ ($\mathrm{R}^2 = 0.698$), $[\mathrm{ATP}]$ ($\mathrm{R}^2 = 0.700$) and $[\mathrm{GLC}]$ $\mathrm{R}^2 = 0.775$). The developed $[\mathrm{GLY}]$ regression model shows unsatisfactory performance ($\mathrm{R}^2 = 0.474$). It is evident that the performance of the models follows the trend of the models trained for simpler spatial average concentration predictions.

Some examples of predicted spatial metabolite concentrations are presented in Fig. 6.4, along with the actual metabolite concentration profile and the absolute relative reconstruction error (%). It is evident that, while most of the trends are retained, there are certain zones of higher relative errors for certain models. However, it should be noted that for certain metabolite concentrations,

Table 6.4: Optimal hyperparameters and spatial profile metabolite concentration regression test set performance metrics for the developed ANN models. Performance for [PYR] and [LAC] are excellent. For [GLC], [ATP] and [ADP], the results are reasonable. However, the [GLY] model demonstrates a non-satisfactory performance.

	N_{layers}	N_{neurons}	activation	R^2	MAPE	MAE	MSE
[GLC]	2	285	sigmoid	0.775	9.24%	1.34E-02	6.16E-04
[ATP]	3	500	sigmoid	0.700	12.93%	1.69E-01	6.31E-02
[ADP]	3	650	sigmoid	0.698	13.10%	1.69E-01	6.24E-02
[GLY]	5	340	sigmoid	0.474	64.43%	4.40E-02	5.78E-03
[PYR]	5	10	elu	0.956	9.33%	2.42E-02	1.47E-03
[LAC]	5	10	elu	0.972	1.40%	2.17E-02	1.43E-03

the actual values can be quite small, which means that small absolute errors in the prediction can lead to very high relative errors.

6.3.4 Effect of inputs on model output (SHAP analysis)

Following the training of individual models for both spatial metabolite concentrations and spatial average concentrations, we would like to investigate the influence of model inputs on model outputs.

For this task, we will focus on three spatial average concentration models. Specifically, we will perform a SHAP analysis for the spatial average predictive models of [LAC], [ATP], and [ADP]. The results of the SHAP analysis for [LAC] and [ATP] are presented in Fig. 6.4c.

For the [LAC] model, it appears that the positions of the LDH reaction centers on the y-axis are the most influential variables for the model output. Furthermore, it appears that when the reaction centers are located high on the y-axis, the model output is negatively influenced.

For the [ATP] model, it appears that the positions of the mitochondria on the y-axis are the most influential variables for the model output. Furthermore, we can observe that when the mitochondria are located close to the inlet of the cell, the predicted spatial average [ATP] is negatively affected.

For the [ADP] model, it appears that the positions of the mitochondria on the y-axis are the most influential variables for the model output. Furthermore, we can observe that when the mitochondria are located close to the outlet of the cell, the predicted spatial average [ADP] is negatively affected.

The results of the SHAP analysis for the $[\mathrm{ATP}]$ and $[\mathrm{ADP}]$ models appear to be mirror images of each other. It looks like when the mitochondria are located close to the cell inlet, the average $[\mathrm{ATP}]$ of the cell is negatively affected, while the average $[\mathrm{ADP}]$ is positively affected. However, when the mitochondria are located close to the cell outlet, the average $[\mathrm{ATP}]$ appears to be positively affected, while the average $[\mathrm{ADP}]$ is negatively affected.

It is worth pointing out that although SHAP analysis can shed light on the relationship between model inputs and outputs, it cannot be used to make causal claims regarding these relationships.

6.4 Conclusions & Perspectives

Several machine learning methods are implemented to gain insight into metabolism within astrocyte cells by exploring a dataset consisting of static metabolic images of astrocytes generated from a reaction-diffusion computational model solved in a two-dimensional geometrical domain.

Hierarchical clustering of the dataset, based solely on the outputs of the computational experiments, results in 5 groups. Each group has its distinctive metabolic characteristics. The discovered clusters reveal differences not only regarding the average metabolite concentrations of their members but also on the side of the inputs (reaction centers). By analyzing the reaction center configurations corresponding to each cluster, we observe that a uniform mitochondrial distribution results in the best overall cell health. On the contrary, realizations of polarized log-normal mitochondrial distributions exhibit the worst overall cell health. It is also shown that a high concentration of mitochondria close to the cell inlet is not beneficial to the cell, especially compared to a better coverage of the cell. Our results are in agreement with the discussion presented on the original model [40].

Following the encouraging results of the clustering analysis, we developed a classification model able to accurately predict cell health (healthy/non-energized) taking only the reaction center positions as input. The nature of the logistic regression model used also allows us to confirm the importance of mitochondrial distribution on cell health.

Moving one step forward and past the binary nature of healthy/non-energized output, we constructed artificial neural network (ANN) regression models that can satisfactorily predict the average spatial concentration of the six metabolites of interest. We show that it is also possible to

recreate the spatial profiles for the six metabolites with satisfactory accuracy. For both regression problems, the LAC and PYR predictive models perform best, followed by those for ATP, ADP, and GLC. The GLY models exhibit only a modest performance. In general, regression models that predict spatial average metabolite concentrations perform better than their counterparts that predict the *full* spatial profiles.

Following regression, SHAP analysis allows us to see into the black boxes that are ANNs, highlighting the importance and influence of different reaction center positions on model output. Specifically, the positions of the lactate dehydrogenase (LDH) reaction centers greatly influence the spatial average concentration of [LAC], and the positions of the mitochondria influence the spatial average concentration of [ATP] and [ADP], consistent with the formulation of the metabolic model.

Based on these encouraging results, several future steps are possible. The first step would involve applying a similar approach to more complex computational domains that better represent actual astrocytic cells. Going further, we could use existing experimental images that provide mitochondrial location data [257] as initial inputs for our approach, with a suitable choice of computational model. Extending this methodology to different or more detailed metabolic pathways could also provide insights into a broader range of species that influence metabolism.

While in this paper we have considered a simplified two-dimensional domain and a basic metabolic model, our machine learning approach can serve as a bridge linking the spatial information from *in vitro* or *in vivo* images with computational models, thereby gaining deeper insights into metabolic dynamics and cellular states. It should be highlighted that the methods used in this work are not limited to the dataset presented; they can be readily applied to other datasets derived from various metabolic computational models.

Chapter 7

Conclusions

7.1 Summary and conclusions

By proposing a blend of equation-based and data-driven methods, this thesis attempts to answer the core question presented in Section 1.1:

> "How should we approach processes that lack the necessary data to be optimally modeled by data-driven methods, but are also too complex to be optimally modeled solely by equation based methods?"

In Chapter 3 we presented an efficient chemistry-enhanced CFD model of the CVD process used for the α-Al_2O_3 coating of cemented carbide cutting tools. This CFD model was the first step taken towards answering the above question, as it allowed for insights of a previously black-box process, while also allowing for the accurate and prediction of alumina thickness. The model also allowed for several computational experiments, which were conducted with the aim of better understanding the rate-determining mechanism of the alumina deposition. The results suggest that the process is probably in the reaction-limited regime.

In Chapter 4, we presented a comparison between the CFD model developed in Chapter 3 and data-driven approaches that can also provide accurate predictions of alumina thickness. The developed ensemble tree-based models -the best performer being an XGBoost model- were trained using *only* production data, and were able to demonstrate better performance than their CFD counterparts (MAPE of 4.4% vs. 6.0%). Furthermore, a reduction of around 99.99% was observed

when comparing the time required for thickness predictions for an entire reactor using the XGBoost model with the time required for prediction via CFD. However, the benefits of the data-driven model were not without drawbacks, as the regression models developed are oblivious to the phenomena occurring within the reactor and therefore unable to provide relative insights. This chapter was finalized with the implementation of another data-driven approach, namely GappyPOD [35], which allowed for the reconstruction of the full state-space of the CFD simulations using only limited data with satisfactory accuracy.

These two chapters aimed to propose a way of modeling the process in a way that allows for quick predictions (in applications where it is required, e.g. optimization) while retaining the option of investigating the physical and chemical complexity of the process. This way, we believe that question 1 presented in Section 1.2:

> "Using this CVD process as a case study, what is the best computational workflow that allows us to make accurate predictions in a resource-efficient and low-cost way?"

was adequately answered.

With Chapter 5, we proposed a combination of unsupervised and supervised learning techniques for the analysis of production data and the discovery of potential critical process parameters. By performing agglomerative hierarchical clustering on the process outputs (thickness measurements), we were able to identify clusters of production runs with evident qualitative differences. This was of course due to the qualitative significance that the process outputs (coating thickness measurements) have for the process. Using the first two statistical moments of thickness (mean and standard deviation) of the discovered clusters, we discovered significant differences between the groups. By analyzing the inputs corresponding to each group, we were able to identify inputs that differed significantly between the groups and can be potentially critical to the process. The two prominent ones were the production "recipe" and a feature that demonstrates the difference between the actual substrate surface area within the reactor and the substrate surface area for which the process operators planned. These inputs were then used, alongside others, to train models (random forests for classification and XGBoost for regression) for qualitative and quantitative predictions of future production runs. A SHAP analysis was also performed to quantify the effect

of each input on the output of the model. The acquired insights were corroborated by subject matter expertise.

Subsequently, in Chapter 6 we demonstrated the applicability of the framework proposed in Chapter 5 on an entirely different dataset. Using a dataset consisting of computational experiments for astrocyte cell metabolism and through the use of agglomerative hierarchical clustering on the metabolite concentrations output by the computational model, we were able to identify reaction centers (mitochondria) whose locations are potentially critical for the health (or energy state) of the cell. Leveraging this insight, we developed models (logistic regression for classification, ANNs for regression) capable of qualitatively predicting cell health and quantitatively predicting cell metabolite concentrations. As in the previous chapter, a SHAP analysis was performed, which, along with subject matter expertise, corroborated our findings.

With these two chapters, we aimed to propose a way of identifying potentially critical parameters using solely data, and to further test the framework on an entirely different application. Thus, we believe that question 2 posed in Section 1.2:

"Can we determine critical process parameters from the available production data?"

was also adequately answered.

Thus, by answering these two research questions, we believe that the proposed approach can effectively tackle processes that *"are too complex to be optimally modeled by equation-based methods, but also lack the necessary data to be optimally modeled by data-driven methods."*

7.2 Future directions

There is plethora of avenues for the evolution and expansion of the work and the approach presented in the work. Although the results for the two different applications are encouraging, it would be worthwhile to try to apply the approach in more real-world applications, with the aim of testing its versatility even more rigorously. This could be quite straightforward to achieve given the several examples of complex processes with limited available data, both in industrial and research settings.

Focusing entirely on the data-driven part, it would be interesting to explore how the approach can be tweaked to handle spatially resolved data from dynamical systems and whether it will be

able to extract similar insights for the process given such data. Furthermore, for the predictive part of the approach for spatially resolved dynamical systems, the implementation of more sophisticated ANN-based models, such as latent neural operators [258] or attention-based architectures (which are capable of capturing long-range dependencies in sequential data), such as the Perceiver IO [259] can be investigated.

As causal insights are vital for our understanding of every process, the implementation of double/debiased machine learning (DML) approaches [260] alongside other causal inference methods [261] within the framework would be of great interest and importance.

We believe that this work can assist in modeling complex processes characterized by limited data and sincerely hope that the proposed approach will prove valuable in a range of distinct applications.

Bibliography

[1] D. Hochauer, C. Mitterer, M. Penoy, S. Puchner, C. Michotte, H. Martinz, H. Hutter, and M. Kathrein, "Carbon doped α-Al2O3 coatings grown by chemical vapor deposition," *Surface and Coatings Technology*, vol. 206, no. 23, pp. 4771–4777, 2012. DOI: 10.1016/j.surfcoat.2012.03.059.

[2] H. Endo, K. Kuwana, K. Saito, D. Qian, R. Andrews, and E. A. Grulke, "CFD prediction of carbon nanotube production rate in a CVD reactor," *Chemical Physics Letters*, vol. 387, no. 4, pp. 307–311, 2004. DOI: 10.1016/j.cplett.2004.01.124.

[3] B. Mitrovic, A. Gurary, and L. Kadinski, "On the flow stability in vertical rotating disc MOCVD reactors under a wide range of process parameters," *Journal of Crystal Growth*, vol. 287, no. 2, pp. 656–663, 2006. DOI: 10.1016/j.jcrysgro.2005.10.131.

[4] G. Gakis, E. Koronaki, and A. Boudouvis, "Numerical investigation of multiple stationary and time-periodic flow regimes in vertical rotating disc CVD reactors," *Journal of Crystal Growth*, vol. 432, pp. 152–159, 2015. DOI: 10.1016/j.jcrysgro.2015.09.026.

[5] I. G. Aviziotis, T. Duguet, C. Vahlas, and A. G. Boudouvis, "Combined Macro/Nanoscale Investigation of the Chemical Vapor Deposition of Fe from Fe(CO)5," *Advanced Materials Interfaces*, vol. 4, no. 18, p. 1 601 185, 2017. DOI: 10.1002/admi.201601185.

[6] K. Jeon, S. Yang, D. Kang, J. Na, and W. B. Lee, "Development of surrogate model using CFD and deep neural networks to optimize gas detector layout," *Korean Journal of Chemical Engineering*, vol. 36, no. 3, pp. 325–332, 2019. DOI: 10.1007/s11814-018-0204-8.

[7] E. Koronaki, P. Gkinis, L. Beex, S. Bordas, C. Theodoropoulos, and A. Boudouvis, "Classification of states and model order reduction of large scale Chemical Vapor Deposition processes with solution multiplicity," *Computers & Chemical Engineering*, vol. 121, pp. 148–157, 2019. DOI: 10.1016/j.compchemeng.2018.08.023.

[8] P. Gkinis, E. Koronaki, A. Skouteris, I. Aviziotis, and A. Boudouvis, "Building a data-driven reduced order model of a chemical vapor deposition process from low-fidelity CFD simulations," *Chemical Engineering Science*, vol. 199, pp. 371–380, 2019. DOI: 10.1016/j.ces.2019.01.009.

[9] R. Spencer, P. Gkinis, E. Koronaki, D. Gerogiorgis, S. Bordas, and A. Boudouvis, "Investigation of the chemical vapor deposition of Cu from copper amidinate through data driven efficient CFD modelling," *Computers & Chemical Engineering*, vol. 149, p. 107 289, 2021. DOI: 10.1016/j.compchemeng.2021.107289.

[10] T. P. Raptis, A. Passarella, and M. Conti, "Data Management in Industry 4.0: State of the Art and Open Challenges," *IEEE Access*, vol. 7, pp. 97 052–97 093, 2019. DOI: 10.1109/ACCESS.2019.2929296.

[11] M. Bertolini, D. Mezzogori, M. Neroni, and F. Zammori, "Machine Learning for industrial applications: A comprehensive literature review," *Expert Systems with Applications*, vol. 175, p. 114 820, 2021. DOI: 10.1016/j.eswa.2021.114820.

[12] J. Wang, J. Yan, C. Li, R. X. Gao, and R. Zhao, "Deep heterogeneous GRU model for predictive analytics in smart manufacturing: Application to tool wear prediction," *Computers in Industry*, vol. 111, pp. 1–14, 2019. DOI: 10.1016/j.compind.2019.06.001.

[13] Z. Li, R. Liu, and D. Wu, "Data-driven smart manufacturing: Tool wear monitoring with audio signals and machine learning," *Journal of Manufacturing Processes*, vol. 48, pp. 66–76, 2019. DOI: 10.1016/j.jmapro.2019.10.020.

[14] K. Kammerer, B. Hoppenstedt, R. Pryss, S. Stökler, J. Allgaier, and M. Reichert, "Anomaly Detections for Manufacturing Systems Based on Sensor Data—Insights into Two Challenging Real-World Production Settings," *Sensors*, vol. 19, no. 24, p. 5370, 2019. DOI: 10.3390/s19245370.

[15] R. Iqbal, T. Maniak, F. Doctor, and C. Karyotis, "Fault Detection and Isolation in Industrial Processes Using Deep Learning Approaches," *IEEE Trans. Ind. Inf.*, vol. 15, no. 5, pp. 3077–3084, 2019. DOI: 10.1109/tii.2019.2902274.

[16] D. Kim and S. Kang, "Effect of Irrelevant Variables on Faulty Wafer Detection in Semiconductor Manufacturing," *Energies*, vol. 12, no. 13, p. 2530, 2019. DOI: 10.3390/en12132530.

[17] M. Papananias, T. E. McLeay, M. Mahfouf, and V. Kadirkamanathan, "A Bayesian framework to estimate part quality and associated uncertainties in multistage manufacturing," *Computers in Industry*, vol. 105, pp. 35–47, 2019. DOI: 10.1016/j.compind.2018.10.008.

[18] M. Saqlain, B. Jargalsaikhan, and J. Y. Lee, "A Voting Ensemble Classifier for Wafer Map Defect Patterns Identification in Semiconductor Manufacturing," *IEEE Trans. Semicond. Manufact.*, vol. 32, no. 2, pp. 171–182, 2019. DOI: 10.1109/tsm.2019.2904306.

[19] D. P. Penumuru, S. Muthuswamy, and P. Karumbu, "Identification and classification of materials using machine vision and machine learning in the context of industry 4.0," *J Intell Manuf*, vol. 31, no. 5, pp. 1229–1241, 2020. DOI: 10.1007/s10845-019-01508-6.

[20] J. Heger, J. Branke, T. Hildebrandt, and B. Scholz-Reiter, "Dynamic adjustment of dispatching rule parameters in flow shops with sequence-dependent set-up times," *International Journal of Production Research*, vol. 54, no. 22, pp. 6812–6824, 2016. DOI: 10.1080/00207543.2016.1178406.

[21] P. Agarwal, M. Tamer, M. H. Sahraei, and H. Budman, "Deep Learning for Classification of Profit-Based Operating Regions in Industrial Processes," *Ind. Eng. Chem. Res.*, vol. 59, no. 6, pp. 2378–2395, 2020. DOI: 10.1021/acs.iecr.9b04737.

[22] K. D. Humfeld, D. Gu, G. A. Butler, K. Nelson, and N. Zobeiry, "A machine learning framework for real-time inverse modeling and multi-objective process optimization of composites for active manufacturing control," *Composites Part B: Engineering*, vol. 223, p. 109 150, 2021. DOI: 10.1016/j.compositesb.2021.109150.

[23] Y. Ma, W. Zhu, M. G. Benton, and J. Romagnoli, "Continuous control of a polymerization system with deep reinforcement learning," *Journal of Process Control*, vol. 75, pp. 40–47, 2019. DOI: 10.1016/j.jprocont.2018.11.004.

[24] J. Dornheim, N. Link, and P. Gumbsch, "Model-Free Adaptive Optimal Control of Episodic Fixed-Horizon Manufacturing Processes using Reinforcement Learning," *Int. J. Control Autom. Syst.*, vol. 18, no. 6, pp. 1593–1604, 2020. DOI: 10.1007/s12555-019-0120-7.

[25] I. M. Cavalcante, E. M. Frazzon, F. A. Forcellini, and D. Ivanov, "A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing," *International Journal of Information Management*, vol. 49, pp. 86–97, 2019. DOI: 10.1016/j.ijinfomgt.2019.03.004.

[26] H. Du and Y. Jiang, "Backup or Reliability Improvement Strategy for a Manufacturer Facing Heterogeneous Consumers in a Dynamic Supply Chain," *IEEE Access*, vol. 7, pp. 50 419–50 430, 2019. DOI: 10.1109/access.2019.2911620.

[27] A. Kara and I. Dogan, "Reinforcement learning approaches for specifying ordering policies of perishable inventory systems," *Expert Systems with Applications*, vol. 91, pp. 150–158, 2018. DOI: 10.1016/j.eswa.2017.08.046.

[28] P. Priore, B. Ponte, R. Rosillo, and D. de la Fuente, "Applying machine learning to the dynamic selection of replenishment policies in fast-changing supply chain environments," *International Journal of Production Research*, vol. 57, no. 11, pp. 3663–3677, 2019. DOI: 10.1080/00207543.2018.1552369.

[29] S. M. Moosavi, K. M. Jablonka, and B. Smit, "The Role of Machine Learning in the Understanding and Design of Materials," *Journal of the American Chemical Society*, vol. 142, no. 48, pp. 20 273–20 287, 2020. DOI: 10.1021/jacs.0c09105.

[30] A. M. Schweidtmann, E. Esche, A. Fischer, M. Kloft, J.-U. Repke, S. Sager, and A. Mitsos, "Machine Learning in Chemical Engineering: A Perspective," *Chemie Ingenieur Technik*, vol. 93, no. 12, pp. 2029–2039, 2021. DOI: 10.1002/cite.202100083.

[31] G. Miragliotta, A. Sianesi, E. Convertini, and R. Distante, "Data driven management in Industry 4.0: A method to measure Data Productivity," *IFAC-PapersOnLine*, 16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018, vol. 51, no. 11, pp. 19–24, 2018. DOI: 10.1016/j.ifacol.2018.08.228.

[32] P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, T. J. Mountziaris, and S. P. A. Bordas, "An efficient chemistry-enhanced CFD model for the investigation of the rate-limiting mechanisms in industrial Chemical Vapor Deposition reactors," *Chemical Engineering Research and Design*, vol. 186, pp. 314–325, 2022. DOI: 10.1016/j.cherd.2022.08.005.

[33] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression," *Energy*, vol. 164, pp. 465–474, 2018. DOI: 10.1016/j.energy.2018.08.207.

[34] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[35] K. Willcox, "Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition," *Computers & Fluids*, vol. 35, no. 2, pp. 208–226, 2006. DOI: 10.1016/j.compfluid.2004.11.006.

[36] P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "Equation-based and data-driven modeling strategies for industrial coating processes," *Computers in Industry*, vol. 149, p. 103 938, 2023. DOI: 10.1016/j.compind.2023.103938.

[37] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[38] P. Papavasileiou, D. G. Giovanis, G. Pozzetti, M. Kathrein, C. Czettl, I. G. Kevrekidis, A. G. Boudouvis, S. P. A. Bordas, and E. D. Koronaki, "Integrating supervised and unsupervised learning approaches to unveil critical process inputs," *Computers & Chemical Engineering*, vol. 192, p. 108 857, 2025. DOI: 10.1016/j.compchemeng.2024.108857.

[39] S. Farina, S. Claus, J. S. Hale, A. Skupin, and S. P. A. Bordas, "A cut finite element method for spatially resolved energy metabolism models in complex neuro-cell morphologies with minimal remeshing," *Adv. Model. and Simul. in Eng. Sci.*, vol. 8, no. 1, p. 5, 2021. DOI: 10.1186/s40323-021-00191-8.

[40] S. Farina, V. Voorsluijs, S. Fixemer, D. Bouvier, S. Claus, M. Ellisman, S. P. Bordas, and A. Skupin, "Mechanistic multiscale modelling of energy metabolism in human astrocytes reveals the impact of morphology changes in Alzheimer's Disease," *PLOS Computational Biology*, vol. 19, no. 9, e1011464, 2023. DOI: 10.1371/journal.pcbi.1011464.

[41] P. Papavasileiou, S. Farina, E. D. Koronaki, A. G. Boudouvis, S. P. A. Bordas, and A. Skupin, *Machine Learning-based Predictions of Spatial Metabolic Profiles Demonstrate the Impact of Morphology on Astrocytic Energy Metabolism*, 2024. DOI: 10.1101/2024.09.18.613725.

[42] P. Papavasileiou, E. D. Koronaki, A. G. Boudouvis, S. P. A. Bordas, G. Pozzetti, M. Kathrein, and C. Czettl, "Development of an efficient CFD model for an industrial scale CVD reactor with rotating gas feeding system," presented at the 10th GRACM International Congress on Computational Mechanics, Virtual conference, 2021.

[43] P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, S. P. A. Bordas, and A. G. Boudouvis, "Assessment of CFD and ML modelling strategies for industrial chemical vapor deposition reactors," presented at the 13th PESXM (Panhellenic Scientific Conference on Chemical Engineering), Patras, Greece, 2022.

[44] P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "A comparison of equation-based and machine learning models of industrial scale deposition processes," presented at the ECCOMAS2022 International Congress on Computational Mechanics, Oslo, Norway, 2022.

[45] P. Papavasileiou, E. D. Koronaki, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "An efficient CFD model of an industrial scale CVD reactor allowing accurate coating thickness predictions," presented at the WCCM-APCOM 15th World Congress on Computational Mechanics and 8th Asian Pacific Congress on Computational Mechanics, Yokohama, Japan (virtual congress), 2022.

[46] P. Papavasileiou, E. D. Koronaki, D. G. Giovanis, G. Pozzetti, M. Kathrein, C. Czettl, A. G. Boudouvis, and S. P. A. Bordas, "Investigation of defining process inputs using unsupervised machine learning," presented at the 14th European Congress of Chemical Engineering and 7th European Congress of Applied Biotechnology, Berlin, Germany, 2023.

[47] R. M. Biefeld, "The metal-organic chemical vapor deposition and properties of III–V antimony-based semiconductor materials," *Materials Science and Engineering: R: Reports*, vol. 36, no. 4, pp. 105–142, 2002. DOI: 10.1016/S0927-796X(02)00002-5.

[48] D. R. Cote, S. V. Nguyen, A. K. Stamper, D. S. Armbrust, D. Tobben, R. A. Conti, and G. Y. Lee, "Plasma-assisted chemical vapor deposition of dielectric thin films for ULSI semiconductor circuits," *IBM Journal of Research and Development*, vol. 43, no. 1.2, pp. 5–38, 1999. DOI: 10.1147/rd.431.0005.

[49] A. de Graaf, J. van Deelen, P. Poodt, T. van Mol, K. Spee, F. Grob, and A. Kuypers, "Development of atmospheric pressure CVD processes for highquality transparent conductive oxides," *Energy Procedia*, Proceedings of Inorganic and Nanostructured Photovoltaics (E-MRS 2009 Symposium B), vol. 2, no. 1, pp. 41–48, 2010. DOI: 10.1016/j.egypro.2010.07.008.

[50] K. C. Topka, B. Diallo, M. Puyo, P. Papavasileiou, C. Lebesgue, C. Genevois, Y. Tison, C. Charvillat, D. Samelor, *et al.*, "Critical Level of Nitrogen Incorporation in Silicon Oxynitride Films: Transition of Structure and Properties, toward Enhanced Anticorrosion Performance," *ACS Applied Electronic Materials*, vol. 4, no. 4, pp. 1741–1755, 2022. DOI: 10.1021/acsaelm.2c00018.

[51] K. C. Topka, B. Diallo, M. Puyo, E. Chesneau, F. Inoubli, S. Ponton, C. Genevois, D. Samelor, R. Laloo, *et al.*, "Silicon Oxynitride Coatings Are Very Promising for Inert and

Durable Pharmaceutical Glass Vials," *ACS Applied Engineering Materials*, vol. 1, no. 12, pp. 3268–3283, 2023. DOI: 10.1021/acsaenm.3c00584.

[52] L. F. Cheng, Y. Xu, L. Zhang, and X. Yin, "Preparation of an oxidation protection coating for c/c composites by low pressure chemical vapor deposition," *Carbon*, vol. 38, no. 10, pp. 1493–1498, 2000. DOI: 10.1016/S0008-6223(00)00086-5.

[53] T. Goto, "Thermal barrier coatings deposited by laser CVD," *Surface and Coatings Technology*, vol. 198, no. 1, pp. 367–371, 2005. DOI: 10.1016/j.surfcoat.2004.10.084.

[54] S. Jia, W. Chen, J. Zhang, C. .-. Lin, H. Guo, G. Lu, K. Li, T. Zhai, Q. Ai, *et al.*, "CVD growth of high-quality and large-area continuous h-BN thin films directly on stainless-steel as protective coatings," *Materials Today Nano*, vol. 16, p. 100 135, 2021. DOI: 10.1016/j.mtnano.2021.100135.

[55] M. Kathrein, W. Schintlmeister, W. Wallgram, and U. Schleinkofer, "Doped CVD Al2O3 coatings for high performance cutting tools," *Surface and Coatings Technology*, Proceedings of the 29th International Conference on Metallurgical Coatings and Thin Films, vol. 163–164, pp. 181–188, 2003. DOI: 10.1016/s0257-8972(02)00483-8.

[56] M. Gassner, N. Schalk, M. Tkadletz, C. Czettl, and C. Mitterer, "Thermal crack network on CVD TiCN/α-Al2O3 coated cemented carbide cutting tools," *International Journal of Refractory Metals and Hard Materials*, vol. 81, pp. 1–6, 2019. DOI: 10.1016/j.ijrmhm.2019.02.006.

[57] J. R. Creighton and J. E. Parmeter, "Metal CVD for microelectronic applications: An examination of surface chemistry and kinetics," *Critical Reviews in Solid State and Materials Sciences*, vol. 18, no. 2, pp. 175–237, 1993. DOI: 10.1080/10408439308242560.

[58] J. Hwang, V. B. Shields, C. I. Thomas, S. Shivaraman, D. Hao, M. Kim, A. R. Woll, G. S. Tompa, and M. G. Spencer, "Epitaxial growth of graphitic carbon on C-face SiC and sapphire by chemical vapor deposition (CVD)," *Journal of Crystal Growth*, vol. 312, no. 21, pp. 3219–3224, 2010. DOI: 10.1016/j.jcrysgro.2010.07.046.

[59] H. Li, Y. Li, A. Aljarb, Y. Shi, and L.-J. Li, "Epitaxial Growth of Two-Dimensional Layered Transition-Metal Dichalcogenides: Growth Mechanism, Controllability, and Scalability," *Chemical Reviews*, vol. 118, no. 13, pp. 6134–6150, 2018. DOI: 10.1021/acs.chemrev.7b00212.

[60] M. Sabzi, S. H. Mousavi Anijdan, M. Shamsodin, M. Farzam, A. Hojjati-Najafabadi, P. Feng, N. Park, and U. Lee, "A Review on Sustainable Manufacturing of Ceramic-Based Thin Films by Chemical Vapor Deposition (CVD): Reactions Kinetics and the Deposition Mechanisms," *Coatings*, vol. 13, no. 1, p. 188, 2023. DOI: 10.3390/coatings13010188.

[61] Z. Chen, Y.-Y. Lian, X. Liu, F. Feng, B.-Y. Yan, J.-B. Wang, Y.-W. Lv, J.-P. Song, C.-J. Liu, *et al.*, "Recent research and development of thick CVD tungsten coatings for fusion application," *Tungsten*, vol. 2, no. 1, pp. 83–93, 2020. DOI: 10.1007/s42864-020-00041-x.

[62] R. Groenen, J. Löffler, P. M. Sommeling, J. L. Linden, E. A. G. Hamers, R. E. I. Schropp, and M. C. M. van de Sanden, "Surface textured ZnO films for thin film solar cell applications by expanding thermal plasma CVD," *Thin Solid Films*, 3rd International Conference on Coatings on Glass (ICCG), vol. 392, no. 2, pp. 226–230, 2001. DOI: 10.1016/S0040-6090(01)01032-X.

[63] H. Kim, S.-H. Bae, T.-H. Han, K.-G. Lim, J.-H. Ahn, and T.-W. Lee, "Organic solar cells using CVD-grown graphene electrodes," *Nanotechnology*, vol. 25, no. 1, p. 014 012, 2013. DOI: 10.1088/0957-4484/25/1/014012.

[64] R. E. Tressler, "Recent developments in fibers and interphases for high temperature ceramic matrix composites," *Composites Part A: Applied Science and Manufacturing*, vol. 30, no. 4, pp. 429–437, 1999. DOI: 10.1016/S1359-835X(98)00131-6.

[65] W. Jin, Z. Si, Y. Lu, S. Bei-zhi, W. Yi, L. Guang-de, X. Zhong-fang, C. Jie, H. Heng-ping, *et al.*, "Oxidation behavior and high-temperature flexural property of CVD-SiC-coated PIP-C/SiC composites," *Ceramics International*, vol. 44, no. 14, pp. 16 583–16 588, 2018. DOI: 10.1016/j.ceramint.2018.06.082.

[66] B. Liu and S. Ma, "Precise synthesis of graphene by chemical vapor deposition," *Nanoscale*, vol. 16, no. 9, pp. 4407–4433, 2024. DOI: 10.1039/D3NR06041A.

[67] D. Quinto, "Technology perspective on CVD and PVD coated metal-cutting tools," *International Journal of Refractory Metals and Hard Materials*, vol. 14, no. 1-3, pp. 7–20, 1996. DOI: 10.1016/0263-4368(96)83413-5.

[68] H. Prengel, W. Pfouts, and A. Santhanam, "State of the art in hard coatings for carbide cutting tools," *Surface and Coatings Technology*, vol. 102, no. 3, pp. 183–190, 1998. DOI: 10.1016/S0257-8972(96)03061-7.

[69] A. Paseuth, H. Fukui, and K. Yamagata, "Improvement of mechanical properties and cutting performance of modified MT-TiCxN1-x coating by moderate temperature chemical vapor deposition," *Surface and Coatings Technology*, vol. 291, pp. 54–61, 2016. DOI: 10.1016/j.surfcoat.2016.02.023.

[70] C. Czettl, C. Mitterer, U. Mühle, D. Rafaja, S. Puchner, H. Hutter, M. Penoy, C. Michotte, and M. Kathrein, "CO addition in low-pressure chemical vapour deposition of medium-temperature TiCxN1-x based hard coatings," *Surface and Coatings Technology*, Proceedings of the 38th International Conference on Metallurgical Coatings and Thin Films (ICMCTF), vol. 206, no. 7, pp. 1691–1697, 2011. DOI: 10.1016/j.surfcoat.2011.07.086.

[71] M. Tkadletz, J. Keckes, N. Schalk, I. Krajinovic, M. Burghammer, C. Czettl, and C. Mitterer, "Residual stress gradients in α-Al2O3 hard coatings determined by pencil-beam X-ray nanodiffraction: The influence of blasting media," *Surface and Coatings Technology*, vol. 262, pp. 134–140, 2015. DOI: 10.1016/j.surfcoat.2014.12.028.

[72] G. L. Tingey, "Kinetics of the Water—Gas Equilibrium Reaction. I. The Reaction of Carbon Dioxide with Hydrogen," *J. Phys. Chem.*, vol. 70, no. 5, pp. 1406–1412, 1966. DOI: 10.1021/j100877a011.

[73] F. Bustamante, R. M. Enick, A. Cugini, R. P. Killmeyer, B. H. Howard, K. S. Rothenberger, M. V. Ciocco, B. D. Morreale, S. Chattopadhyay, *et al.*, "High-temperature kinetics of the homogeneous reverse water-gas shift reaction," *AIChE J.*, vol. 50, no. 5, pp. 1028–1041, 2004. DOI: 10.1002/aic.10099.

[74] M. Bar-Hen and I. Etsion, "Experimental study of the effect of coating thickness and substrate roughness on tool wear during turning," *Tribology International*, vol. 110, pp. 341–347, 2017. DOI: 10.1016/j.triboint.2016.11.011.

[75] C. K. Harris, D. Roekaerts, F. J. J. Rosendal, F. G. J. Buitendijk, Ph. Daskopoulos, A. J. N. Vreenegoor, and H. Wang, "Computational fluid dynamics for chemical reactor engineering," *Chemical Engineering Science*, Chemical Reaction Engineering: From Fundamentals to Commercial Plants and Products, vol. 51, no. 10, pp. 1569–1594, 1996. DOI: 10.1016/0009-2509(96)00021-8.

[76] R. O. Fox, "CFD Models for Analysis and Design of Chemical Reactors," in *Advances in Chemical Engineering*, ser. Computational Fluid Dynamics, G. B. Marin, Ed., vol. 31, Academic Press, 2006, pp. 231–305. DOI: 10.1016/S0065-2377(06)31004-6.

[77] D. F. Fletcher, "The future of computational fluid dynamics (CFD) simulation in the chemical process industries," *Chemical Engineering Research and Design*, vol. 187, pp. 299–305, 2022. DOI: 10.1016/j.cherd.2022.09.021.

[78] T. M. Mitchell, *Machine Learning* (McGraw-Hill Series in Computer Science). New York: McGraw-Hill, 1997.

[79] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. D. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PLoS ONE*, vol. 14, no. 1, H. A. Kestler, Ed., e0210236, 2019. DOI: 10.1371/journal.pone.0210236.

[80] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Unsupervised Learning," in *An Introduction to Statistical Learning: With Applications in R*, ser. Springer Texts in Statistics, G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds., New York, NY: Springer US, 2021, pp. 497–552. DOI: 10.1007/978-1-0716-1418-1_12.

[81] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Statistical Learning," in *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer US, 2021, pp. 15–57. DOI: 10.1007/978-1-0716-1418-1_2.

[82] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised Learning," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer Series in Statistics, T. Hastie, R. Tibshirani, and J. Friedman, Eds., New York, NY: Springer, 2009, pp. 485–585. DOI: 10.1007/978-0-387-84858-7_14.

[83] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Oakland, CA, USA, 1967, pp. 281–297.

[84] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[85] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, 1999. DOI: 10.1145/304181.304187.

[86] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, 19:1–19:21, 2017. DOI: 10.1145/3068335.

[87] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012. DOI: 10.1002/widm.53.

[88] Vijaya, S. Sharma, and N. Batra, "Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India: IEEE, 2019, pp. 568–573. DOI: 10.1109/COMITCon.2019.8862232.

[89] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[90] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. DOI: 10.1080/00401706.1970.10488634.

[91] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, 1995. DOI: 10.1007/BF00994018.

[92] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York: Chapman and Hall/CRC, 1984. DOI: 10.1201/9781315139470.

[93] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[94] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[95] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, pp. 3–42, 2006. DOI: 10.1007/s10994-006-6226-1.

[96] T. Hastie, R. Tibshirani, and J. Friedman, "Ensemble Learning," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, T. Hastie, R. Tibshirani, and J. Friedman, Eds., New York, NY: Springer, 2009, pp. 605–624. DOI: 10.1007/978-0-387-84858-7_16.

[97] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*. Cham: Springer International Publishing, 2018. DOI: 10.1007/978-3-319-94463-0.

[98] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.

[99] F. Chollet *et al.*, *Keras*, 2015.

[100] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[101] J. W. Tukey, "Exploratory data analysis," *Reading/Addison-Wesley*, 1977.

[102] S. Morgenthaler, "Exploratory data analysis," *WIREs Computational Statistics*, vol. 1, no. 1, pp. 33–44, 2009. DOI: 10.1002/wics.2.

[103] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," in *SoutheastCon 2016*, 2016, pp. 1–6. DOI: 10.1109/SECON.2016.7506650.

[104] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. New York: Chapman and Hall/CRC, 2019, 310 pp. DOI: 10.1201/9781315108230.

[105] K. Potdar, T. S., and C. D., "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *IJCA*, vol. 175, no. 4, pp. 7–9, 2017. DOI: 10.5120/ijca2017915495.

[106] D. M. Hawkins, "The Problem of Overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004. DOI: 10.1021/ci0342472.

[107] X. Ying, "An Overview of Overfitting and its Solutions," *Journal of Physics: Conference Series*, vol. 1168, p. 022 022, 2019. DOI: 10.1088/1742-6596/1168/2/022022.

[108] A. Blum, A. Kalai, and J. Langford, "Beating the hold-out: Bounds for K-fold and progressive cross-validation," in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, Santa Cruz California USA: ACM, 1999, pp. 203–208. DOI: 10.1145/307400.307439.

[109] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?" *Journal of the American Statistical Association*, vol. 119, no. 546, pp. 1434–1445, 2024. DOI: 10.1080/01621459.2023.2197686.

[110] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). New York, NY: Springer New York, 2009. DOI: 10.1007/978-0-387-84858-7.

[111] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature Selection: A Data Perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 94:1–94:45, 2017. DOI: 10.1145/3136625.

[112] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, 2016. DOI: 10.1098/rsta.2015.0202.

[113] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear Discriminant Analysis," in *Robust Data Mining*, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, Eds., New York, NY: Springer, 2013, pp. 27–33. DOI: 10.1007/978-1-4419-9878-1_4.

[114] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 1200–1205. DOI: 10.1109/MIPRO.2015.7160458.

[115] N. Abd-Alsabour, "A Review on Evolutionary Feature Selection," in *2014 European Modelling Symposium*, 2014, pp. 20–26. DOI: 10.1109/EMS.2014.28.

[116] C. Khosla and B. S. Saini, "Enhancing Performance of Deep Learning Models with different Data Augmentation Techniques: A Survey," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 2020, pp. 79–85. DOI: 10.1109/ICIEM48762.2020.9160048.

[117] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data Augmentation Can Improve Robustness," in *Advances in Neural Information Processing Systems*, vol. 34, Curran Associates, Inc., 2021, pp. 29 935–29 948.

[118] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 1 2019. DOI: 10.1186/s40537-019-0197-0.

[119] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[120] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.

[121] N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, "An Investigation of SMOTE Based Methods for Imbalanced Datasets With Data Complexity Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6651–6672, 2023. DOI: 10.1109/TKDE.2022.3179381.

[122] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, 2022. DOI: 10.48550/arXiv.1312.6114, pre-published.

[123] J. Fang, C. Tang, Q. Cui, F. Zhu, L. Li, J. Zhou, and W. Zhu, "Semi-Supervised Learning with Data Augmentation for Tabular Data," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, ser. CIKM '22, New York, NY, USA: Association for Computing Machinery, 2022, pp. 3928–3932. DOI: 10.1145/3511808.3557699.

[124] L. Prechelt, "Early Stopping - But When?" In *Neural Networks: Tricks of the Trade*, G. B. Orr and K.-R. Müller, Eds., Berlin, Heidelberg: Springer, 1998, pp. 55–69. DOI: 10.1007/3-540-49430-8_3.

[125] Y. Tian and Y. Zhang, "A comprehensive survey on regularization strategies in machine learning," *Information Fusion*, vol. 80, pp. 146–166, 2022. DOI: 10.1016/j.inffus.2021.11.005.

[126] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[127] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, e1484, 2023. DOI: 10.1002/widm.1484.

[128] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020. DOI: 10.1016/j.neucom.2020.07.061.

[129] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges* (The Springer Series on Challenges in Machine Learning). Cham: Springer International Publishing, 2019. DOI: 10.1007/978-3-030-05318-5.

[130] J. Snoek, H. Larochelle, and R. P. Adams, *Practical bayesian optimization of machine learning algorithms*, 2012.

[131] C. Williams and C. Rasmussen, "Gaussian Processes for Regression," in *Advances in Neural Information Processing Systems*, vol. 8, MIT Press, 1995.

[132] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential Model-Based Optimization for General Algorithm Configuration," in *Learning and Intelligent Optimization*, C. A. C. Coello, Ed., Berlin, Heidelberg: Springer, 2011, pp. 507–523. DOI: 10.1007/978-3-642-25566-3_40.

[133] N. DeCastro-García, Á. L. Muñoz Castañeda, D. Escudero García, and M. V. Carriegos, "Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm," *Complexity*, vol. 2019, no. 1, p. 6 278 908, 2019. DOI: 10.1155/2019/6278908.

[134] G. Ozaydin-Ince, A. M. Coclite, and K. K. Gleason, "CVD of polymeric thin films: Applications in sensors, biotechnology, microelectronics/organic electronics, microfluidics, MEMS, composites and membranes," *Rep. Prog. Phys.*, vol. 75, no. 1, p. 016 501, 2011. DOI: 10.1088/0034-4885/75/1/016501.

[135] K. F. Jensen, E. O. Einset, and D. I. Fotiadis, "Flow phenomena in chemical vapor deposition of thin films," *Annual Review of Fluid Mechanics*, vol. 23, no. 1, pp. 197–232, 1991. DOI: 10.1146/annurev.fl.23.010191.001213.

[136] C. Theodoropoulos, N. Ingle, and T. Mountziaris, "Computational studies of the transient behavior of horizontal MOVPE reactors," *Journal of Crystal Growth*, vol. 170, no. 1-4, pp. 72–76, 1997. DOI: 10.1016/s0022-0248(96)00637-9.

[137] H. van Santen, C. R. Kleijn, and H. E. A. van den Akker, "On multiple stability of mixed-convection flows in a chemical vapor deposition reactor," *International Journal of Heat and Mass Transfer*, vol. 44, no. 3, pp. 659–672, 2001. DOI: 10.1016/s0017-9310(00)00121-6.

[138] J. Cho and T. J. Mountziaris, "Onset of flow recirculation in vertical rotating-disc chemical vapor deposition reactors," *AIChE Journal*, vol. 59, no. 9, pp. 3530–3538, 2013. DOI: 10.1002/aic.14179.

[139] C. Theodoropoulos, T. Mountziaris, H. Moffat, and J. Han, "Design of gas inlets for the growth of gallium nitride by metalorganic vapor phase epitaxy," *Journal of Crystal Growth*, vol. 217, no. 1-2, pp. 65–81, 2000. DOI: 10.1016/s0022-0248(00)00402-4.

[140] P. Yousefian and S. Pimputkar, "Computational fluid dynamics modeling of a new high-pressure chemical vapor deposition reactor design," *Journal of Crystal Growth*, vol. 566–567, p. 126 155, 2021. DOI: 10.1016/j.jcrysgro.2021.126155.

[141] H. Barua and A. Povitsky, "Numerical model of carbon chemical vapor deposition at internal surfaces," *Vacuum*, vol. 175, p. 109 234, 2020. DOI: 10.1016/j.vacuum.2020.109234.

[142] S. Zou, J. Xiao, V. Wu, and X. D. Chen, "Analyzing industrial CVD reactors using a porous media approach," *Chemical Engineering Journal*, vol. 415, p. 129 038, 2021. DOI: 10.1016/j.cej.2021.129038.

[143] P. Gkinis, I. Aviziotis, E. Koronaki, G. Gakis, and A. Boudouvis, "The effects of flow multiplicity on GaN deposition in a rotating disk CVD reactor," *Journal of Crystal Growth*, vol. 458, pp. 140–148, 2017. DOI: 10.1016/j.jcrysgro.2016.10.065.

[144] N. Cheimarios, M. Kavousanakis, G. Kokkoris, and A. Boudouvis, "Beware of symmetry breaking and periodic flow regimes in axisymmetric CVD reactor setups," *Computers & Chemical Engineering*, vol. 124, pp. 124–132, 2019. DOI: 10.1016/j.compchemeng.2019.02.005.

[145] B. Mitrovic, A. Parekh, J. Ramer, V. Merai, E. Armour, L. Kadinski, and A. Gurary, "Reactor design optimization based on 3D modeling of nitrides deposition in MOCVD vertical rotating disc reactors," *Journal of Crystal Growth*, vol. 289, no. 2, pp. 708–714, 2006. DOI: 10.1016/j.jcrysgro.2005.12.107.

[146] B. Mitrovic, A. Gurary, and W. Quinn, "Process conditions optimization for the maximum deposition rate and uniformity in vertical rotating disc MOCVD reactors based on CFD modeling," *Journal of Crystal Growth*, vol. 303, no. 1, pp. 323–329, 2007. DOI: 10.1016/j.jcrysgro.2006.11.247.

[147] T. Oshika, A. Nishiyama, K. Nakaso, M. Shimada, and K. Okuyama, "Unveiling the magic of H_2S on the CVD-Al_2O_3 coating," *J. Phys. IV France*, vol. 09, no. PR8, Pr8-877-Pr8–883, 1999. DOI: 10.1051/jp4:19998110.

[148] A. Blomqvist, C. Århammar, H. Pedersen, F. Silvearv, S. Norgren, and R. Ahuja, "Understanding the catalytic effects of H2S on CVD-growth of α-alumina: Thermodynamic gas-phase simulations and density functional theory," *Surface and Coatings Technology*, vol. 206, no. 7, pp. 1771–1779, 2011. DOI: 10.1016/j.surfcoat.2011.09.018.

[149] S. Ruppi, "Influence of Process Conditions on the Growth and Texture of CVD Alpha-Alumina," *Coatings*, vol. 10, no. 2, p. 158, 2020. DOI: 10.3390/coatings10020158.

[150] N. Cheimarios, E. D. Koronaki, and A. G. Boudouvis, "Illuminating nonlinear dependence of film deposition rate in a CVD reactor on operating conditions," *Chemical Engineering Journal*, vol. 181–182, pp. 516–523, 2012. DOI: 10.1016/j.cej.2011.11.008.

[151] E. D. Koronaki, G. P. Gakis, N. Cheimarios, and A. G. Boudouvis, "Efficient tracing and stability analysis of multiple stationary and periodic states with exploitation of commercial CFD software," *Chemical Engineering Science*, vol. 150, pp. 26–34, 2016. DOI: 10.1016/j.ces.2016.04.043.

[152] P. Etchepare, H. Vergnes, D. Samélor, D. Sadowski, C. Brasme, B. Caussat, and C. Vahlas, "Amorphous Alumina Coatings on Glass Bottles Using Direct Liquid Injection MOCVD for Packaging Applications," in *13th International Ceramics Congress, Part of CIMTEC 2014*, 2014, pp. 117–122. DOI: 10.4028/www.scientific.net/AST.91.117.

[153] W. Ban, S. Kwon, J. Nam, J. Yang, S. Jang, and D. Jung, "Al2O3 thin films prepared by plasma-enhanced chemical vapor deposition of dimethylaluminum isopropoxide," *Thin Solid Films*, vol. 641, pp. 47–52, 2017. DOI: 10.1016/j.tsf.2017.02.007.

[154] M. Gassner, N. Schalk, M. Tkadletz, M. Pohler, C. Czettl, and C. Mitterer, "Influence of cutting speed and workpiece material on the wear mechanisms of CVD TiCN/α-Al2O3 coated cutting inserts during turning," *Wear*, vol. 398–399, pp. 90–98, 2018. DOI: 10.1016/j.wear.2017.11.019.

[155] R. Stylianou, M. Tkadletz, N. Schalk, M. Penoy, C. Czettl, and C. Mitterer, "Effects of reference materials on texture coefficients determined for a CVD α-Al2O3 coating," *Surface and Coatings Technology*, vol. 359, pp. 314–322, 2019. DOI: 10.1016/j.surfcoat.2018.12.095.

[156] M. Schierling, E. Zimmermann, and D. Neuschütz, "Deposition kinetics of Al $_2$ O $_3$ from AlCl $_3$ -CO $_2$ -H $_2$ -HCl gas mixtures by thermal CVD in a hot-wall reactor," *J. Phys. IV France*, vol. 09, no. PR8, Pr8-85-Pr8–91, 1999. DOI: 10.1051/jp4:1999811.

[157] L. Catoire and M. T. Swihart, "High-Temperature Kinetics of AlCl3 Decomposition in the Presence of Additives for Chemical Vapor Deposition," *J. Electrochem. Soc.*, vol. 149, no. 5, p. C261, 2002. DOI: 10.1149/1.1467366.

[158] R. L. Keiski, T. Salmi, P. Niemistö, J. Ainassaari, and V. J. Pohjola, "Stationary and transient kinetics of the high temperature water-gas shift reaction," *Applied Catalysis A: General*, vol. 137, no. 2, pp. 349–370, 1996. DOI: 10.1016/0926-860x(95)00315-0.

[159] R. G. Munro, "Evaluated Material Properties for a Sintered alpha-Alumina," *Journal of the American Ceramic Society*, vol. 80, no. 8, pp. 1919–1928, 1997. DOI: 10.1111/j.1151-2916.1997.tb03074.x.

[160] P. Tan, J. Müller, and D. Neuschütz, "Gas-Phase Kinetic Modeling of the AlCl3 Decomposition in the AlCl3-CO2-H2-HCl System for a Hot-Wall CVD Reactor," *J. Electrochem. Soc.*, vol. 152, no. 5, p. C288, 2005. DOI: 10.1149/1.1883236.

[161] G. M. Psarellis, I. G. Aviziotis, T. Duguet, C. Vahlas, E. D. Koronaki, and A. G. Boudouvis, "Investigation of reaction mechanisms in the chemical vapor deposition of al from DMEAA," *Chemical Engineering Science*, vol. 177, pp. 464–470, 2018. DOI: 10.1016/j.ces.2017.12.006.

[162] C. R. Kleijn and C. J. Hoogendoorn, "A study of 2- and 3-D transport phenomena in horizontal chemical vapor deposition reactors," *Chemical Engineering Science*, vol. 46, no. 1, pp. 321–334, 1991. DOI: 10.1016/0009-2509(91)80141-K.

[163] C. R. Kleijn, R. Dorsman, K. J. Kuijlaars, M. Okkerse, and H. van Santen, "Multi-scale modeling of chemical vapor deposition processes for thin film technology," *Journal of Crystal Growth*, Proceedings of the Fifth Workshop on Modeling in Crystal Growth, vol. 303, no. 1, pp. 362–380, 2007. DOI: 10.1016/j.jcrysgro.2006.12.062.

[164] K. C. Topka, H. Vergnes, T. Tsiros, P. Papavasileiou, L. Decosterd, B. Diallo, F. Senocq, D. Samelor, N. Pellerin, *et al.*, "An innovative kinetic model allowing insight in the moderate temperature chemical vapor deposition of silicon oxynitride films from tris(dimethylsilyl)amine," *Chemical Engineering Journal*, vol. 431, p. 133 350, 2022. DOI: 10.1016/j.cej.2021.133350.

[165] D. I. Fotiadis, M. Boekholt, K. F. Jensen, and W. Richter, "Flow and heat transfer in CVD reactors: Comparison of Raman temperature measurements and finite element model predictions," *Journal of Crystal Growth*, vol. 100, no. 3, pp. 577–599, 1990. DOI: 10.1016/0022-0248(90)90257-L.

[166] S.-S. Liu and W.-D. Xiao, "CFD–PBM coupled simulation of silicon CVD growth in a fluidized bed reactor: Effect of silane pyrolysis kinetic models," *Chemical Engineering Science*, vol. 127, pp. 84–94, 2015. DOI: 10.1016/j.ces.2015.01.026.

[167] H. Kagermann, "Change Through Digitization—Value Creation in the Age of Industry 4.0," in *Management of Permanent Change*, H. Albach, H. Meffert, A. Pinkwart, and R. Reichwald, Eds., Wiesbaden: Springer Fachmedien, 2015, pp. 23–45. DOI: 10.1007/978-3-658-05014-6_2.

[168] A. Saxena and A. Saad, "Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems," *Applied Soft Computing*, vol. 7, no. 1, pp. 441–454, 2007. DOI: 10.1016/j.asoc.2005.10.001.

[169] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi, "Machine Learning for Predictive Maintenance: A Multiple Classifier Approach," *IEEE Trans. Ind. Inf.*, vol. 11, no. 3, pp. 812–820, 2015. DOI: 10.1109/tii.2014.2349359.

[170] H. Wu, Z. Yu, and Y. Wang, "Experimental study of the process failure diagnosis in additive manufacturing based on acoustic emission," *Measurement*, vol. 136, pp. 445–453, 2019. DOI: 10.1016/j.measurement.2018.12.067.

[171] J. Dalzochio, R. Kunst, E. Pignaton, A. Binotto, S. Sanyal, J. Favilla, and J. Barbosa, "Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges," *Computers in Industry*, vol. 123, p. 103 298, 2020. DOI: 10.1016/j.compind.2020.103298.

[172] D. Kim, P. Kang, S. Cho, H.-j. Lee, and S. Doh, "Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4075–4083, 2012. DOI: 10.1016/j.eswa.2011.09.088.

[173] A. Kim, K. Oh, J.-Y. Jung, and B. Kim, "Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles," *International Journal of Computer Integrated Manufacturing*, vol. 31, no. 8, pp. 701–717, 2018. DOI: 10.1080/0951192x.2017.1407447.

[174] J. A. Carvajal Soto, F. Tavakolizadeh, and D. Gyulai, "An online machine learning framework for early detection of product failures in an Industry 4.0 context," *International Journal of Computer Integrated Manufacturing*, vol. 32, no. 4-5, pp. 452–465, 2019. DOI: 10.1080/0951192x.2019.1571238.

[175] R. Wang, C. F. Cheung, C. Wang, and M. N. Cheng, "Deep learning characterization of surface defects in the selective laser melting process," *Computers in Industry*, vol. 140, p. 103 662, 2022. DOI: 10.1016/j.compind.2022.103662.

[176] P. Priore, B. Ponte, J. Puente, and A. Gómez, "Learning-based scheduling of flexible manufacturing systems using ensemble methods," *Computers & Industrial Engineering*, vol. 126, pp. 282–291, 2018. DOI: 10.1016/j.cie.2018.09.034.

[177] A. Tulsyan, C. Garvin, and C. Ündey, "Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems," *Biotechnology and Bioengineering*, vol. 115, no. 8, pp. 1915–1924, 2018. DOI: 10.1002/bit.26605.

[178] J. Deng, S. Sierla, J. Sun, and V. Vyatkin, "Reinforcement learning for industrial process control: A case study in flatness control in steel industry," *Computers in Industry*, vol. 143, p. 103 748, 2022. DOI: 10.1016/j.compind.2022.103748.

[179] H. Cai, J. Feng, Q. Yang, W. Li, X. Li, and J. Lee, "A virtual metrology method with prediction uncertainty based on Gaussian process for chemical mechanical planarization," *Computers in Industry*, vol. 119, p. 103 228, 2020. DOI: 10.1016/j.compind.2020.103228.

[180] P. Azadi, J. Winz, E. Leo, R. Klock, and S. Engell, "A hybrid dynamic model for the prediction of molten iron and slag quality indices of a large-scale blast furnace," *Computers & Chemical Engineering*, vol. 156, p. 107 573, 2022. DOI: 10.1016/j.compchemeng.2021.107573.

[181] W. Dai, S. Mohammadi, and S. Cremaschi, "A hybrid modeling framework using dimensional analysis for erosion predictions," *Computers & Chemical Engineering*, vol. 156, p. 107 577, 2022. DOI: 10.1016/j.compchemeng.2021.107577.

[182] S. Malley, C. Reina, S. Nacy, J. Gilles, B. Koohbor, and G. Youssef, "Predictability of mechanical behavior of additively manufactured particulate composites using machine learning and data-driven approaches," *Computers in Industry*, vol. 142, p. 103 739, 2022. DOI: 10.1016/j.compind.2022.103739.

[183] Z. He, K.-P. Tran, S. Thomassey, X. Zeng, J. Xu, and C. Yi, "A deep reinforcement learning based multi-criteria decision support system for optimizing textile chemical process," *Computers in Industry*, vol. 125, p. 103 373, 2021. DOI: 10.1016/j.compind.2020.103373.

[184] L. Galvis, T. Offermans, C. G. Bertinetto, A. Carnoli, E. Karamujić, W. Li, E. Szymańska, L. M. C. Buydens, and J. J. Jansen, "Retrospective quality by design r(QbD) for lactose production using historical process data and design of experiments," *Computers in Industry*, vol. 141, p. 103 696, 2022. DOI: 10.1016/j.compind.2022.103696.

[185] H. Boyes and T. Watson, "Digital twins: An analysis framework and open issues," *Computers in Industry*, vol. 143, p. 103 763, 2022. DOI: 10.1016/j.compind.2022.103763.

[186] A. Hürkamp, S. Gellrich, T. Ossowski, J. Beuscher, S. Thiede, C. Herrmann, and K. Dröder, "Combining Simulation and Machine Learning as Digital Twin for the Manufacturing of Overmolded Thermoplastic Composites," *JMMP*, vol. 4, no. 3, p. 92, 2020. DOI: 10.3390/jmmp4030092.

[187] A. Rasheed, O. San, and T. Kvamsdal, "Digital Twin: Values, Challenges and Enablers From a Modeling Perspective," *IEEE Access*, vol. 8, pp. 21 980–22 012, 2020. DOI: 10.1109/access.2020.2970143.

[188] M. Perno, L. Hvam, and A. Haug, "Implementation of digital twins in the process industry: A systematic literature review of enablers and barriers," *Computers in Industry*, vol. 134, p. 103 558, 2022. DOI: 10.1016/j.compind.2021.103558.

[189] S. Urcun, P.-Y. Rohan, W. Skalli, P. Nassoy, S. P. A. Bordas, and G. Sciumè, "Digital twinning of Cellular Capsule Technology: Emerging outcomes from the perspective of porous media mechanics," *PLOS ONE*, vol. 16, no. 7, e0254512, 2021. DOI: 10.1371/journal.pone.0254512.

[190] K. Kalaboukas, D. Kiritsis, and G. Arampatzis, "Governance framework for autonomous and cognitive digital twins in agile supply chains," *Computers in Industry*, vol. 146, p. 103 857, 2023. DOI: 10.1016/j.compind.2023.103857.

[191] S. S. Blakseth, A. Rasheed, T. Kvamsdal, and O. San, "Deep neural network enabled corrective source term approach to hybrid analysis and modeling," *Neural Networks*, vol. 146, pp. 181–199, 2022. DOI: 10.1016/j.neunet.2021.11.021.

[192] S. Deshpande, J. Lengiewicz, and S. P. A. Bordas, "Probabilistic deep learning for real-time large deformation simulations," *Computer Methods in Applied Mechanics and Engineering*, vol. 398, p. 115 307, 2022. DOI: 10.1016/j.cma.2022.115307.

[193] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019. DOI: 10.1016/j.jcp.2018.10.045.

[194] E. D. Koronaki, N. Evangelou, Y. M. Psarellis, A. G. Boudouvis, and I. G. Kevrekidis, *From partial data to out-of-sample parameter and observation estimation with Diffusion Maps and Geometric Harmonics*, 2023. DOI: 10.48550/arXiv.2301.11728.

[195] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Tree-Based Methods," in *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer US, 2021, pp. 327–365. DOI: 10.1007/978-1-0716-1418-1_8.

[196] R. Everson and L. Sirovich, "Karhunen–Loève procedure for gappy data," *J. Opt. Soc. Am. A*, vol. 12, no. 8, p. 1657, 1995. DOI: 10.1364/JOSAA.12.001657.

[197] M. Łępicka and M. Grądzka-Dahlke, "The initial evaluation of performance of hard anti-wear coatings deposited on metallic substrates: Thickness, mechanical properties and adhesion measurements – a brief review," *REVIEWS ON ADVANCED MATERIALS SCIENCE*, vol. 58, no. 1, pp. 50–65, 2019. DOI: 10.1515/rams-2019-0003.

[198] T. Hastie, R. Tibshirani, and J. Friedman, "Additive Models, Trees, and Related Methods," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer Series in Statistics, T. Hastie, R. Tibshirani, and J. Friedman, Eds., New York, NY: Springer, 2009, pp. 295–336. DOI: 10.1007/978-0-387-84858-7_9.

[199] T. Hastie, R. Tibshirani, and J. Friedman, "Boosting and Additive Trees," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer Series in Statistics, T. Hastie, R. Tibshirani, and J. Friedman, Eds., New York, NY: Springer, 2009, pp. 337–387. DOI: 10.1007/978-0-387-84858-7_10.

[200] W. McKinney, "Data Structures for Statistical Computing in Python," in *Python in Science Conference*, Austin, Texas, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[201] P. Kerfriden, O. Goury, T. Rabczuk, and S. P. A. Bordas, "A partitioned model order reduction approach to rationalise computational expenses in nonlinear fracture mechanics," *Computer Methods in Applied Mechanics and Engineering*, vol. 256, pp. 169–188, 2013. DOI: 10.1016/j.cma.2012.12.004.

[202] T. Jo, B. Koo, H. Kim, D. Lee, and J. Y. Yoon, "Effective sensor placement in a steam reformer using gappy proper orthogonal decomposition," *Applied Thermal Engineering*, vol. 154, pp. 419–432, 2019. DOI: 10.1016/j.applthermaleng.2019.03.089.

[203] A. A. Alonso, C. E. Frouzakis, and I. G. Kevrekidis, "Optimal sensor placement for state reconstruction of distributed process systems," *AIChE Journal*, vol. 50, no. 7, pp. 1438–1452, 2004. DOI: 10.1002/aic.10121.

[204] A. A. Alonso, I. G. Kevrekidis, J. R. Banga, and C. E. Frouzakis, "Optimal sensor location and reduced order observer design for distributed process systems," *Computers & Chemical Engineering*, Escape 12, vol. 28, no. 1, pp. 27–35, 2004. DOI: 10.1016/S0098-1354(03)00175-3.

[205] H. Y. Ha, S. W. Nam, T. H. Lim, I.-H. Oh, and S.-A. Hong, "Properties of the TiO2 membranes prepared by CVD of titanium tetraisopropoxide," *Journal of Membrane Science*, vol. 111, no. 1, pp. 81–92, 1996. DOI: 10.1016/0376-7388(95)00278-2.

[206] S. J. Khatib and S. T. Oyama, "Silica membranes for hydrogen separation prepared by chemical vapor deposition (CVD)," *Separation and Purification Technology*, vol. 111, pp. 20–42, 2013. DOI: 10.1016/j.seppur.2013.03.032.

[207] T. Schmauder, K. .-. Nauenburg, K. Kruse, and G. Ickes, "Hard coatings by plasma CVD on polycarbonate for automotive and optical applications," *Thin Solid Films*, Selected Papers from the 5th International Conference on Coatings on Glass (ICCG5)- Advanced Coatings on Glass and Plastics for Large-Area or High-Volume Products, vol. 502, no. 1, pp. 270–274, 2006. DOI: 10.1016/j.tsf.2005.07.296.

[208] J. Karner, M. Pedrazzini, I. Reineck, M. E. Sj\"{o}strand, and E. Bergmann, "CVD diamond coated cemented carbide cutting tools," *Materials Science and Engineering: A*, Proceedings of the 5th International Conference on the Science of Hard Materials, vol. 209, no. 1, pp. 405–413, 1996. DOI: 10.1016/0921-5093(95)10140-3.

[209] E. D. Koronaki, N. Cheimarios, H. Laux, and A. G. Boudouvis, "Non-Axisymmetric Flow Fields in Axisymmetric CVD Reactor Setups Revisited: Influence on the Film's Non-Uniformity," *ECS Solid State Lett.*, vol. 3, no. 4, P37, 2014. DOI: 10.1149/2.002404ssl.

[210] I. G. Aviziotis, N. Cheimarios, T. Duguet, C. Vahlas, and A. G. Boudouvis, "Multiscale modeling and experimental analysis of chemical vapor deposited aluminum films: Linking reactor operating conditions with roughness evolution," *Chemical Engineering Science*, vol. 155, pp. 449–458, 2016. DOI: 10.1016/j.ces.2016.08.039.

[211] C. P. Martin-Linares, Y. M. Psarellis, G. Karapetsas, E. D. Koronaki, and I. G. Kevrekidis, "Physics-agnostic and physics-infused machine learning for thin films flows: Modelling, and predictions from small data," *Journal of Fluid Mechanics*, vol. 975, A41, 2023. DOI: 10.1017/jfm.2023.868.

[212] L. S. Shapley, "A Value for N-Person Games," RAND Corporation, Tech. Rep., 1952.

[213] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012.

[214] M. Sundararajan and A. Najmi, "The Many Shapley Values for Model Explanation," in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, 2020, pp. 9269–9278.

[215] I.-G. Chong and C.-H. Jun, "Performance of some variable selection methods when multicollinearity is present," *Chemometrics and Intelligent Laboratory Systems*, vol. 78, no. 1, pp. 103–112, 2005. DOI: 10.1016/j.chemolab.2004.12.011.

[216] B. Lu, I. Castillo, L. Chiang, and T. F. Edgar, "Industrial PLS model variable selection using moving window variable importance in projection," *Chemometrics and Intelligent Laboratory Systems*, vol. 135, pp. 90–109, 2014. DOI: 10.1016/j.chemolab.2014.03.020.

[217] P. H. Garthwaite, "An Interpretation of Partial Least Squares," *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 122–127, 1994. DOI: 10.1080/01621459.1994.10476452.

[218] K. Kumar, "Partial Least Square (PLS) Analysis: Most Favorite Tool in Chemometrics to Build a Calibration Model," *Reson*, vol. 26, no. 3, pp. 429–442, 2021. DOI: 10.1007/s12045-021-1140-1.

[219] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection – A review and recommendations for the practicing statistician," *Biometrical Journal*, vol. 60, no. 3, pp. 431–449, 2018. DOI: 10.1002/bimj.201700067.

[220] E. D. Koronaki, A. M. Nikas, and A. G. Boudouvis, "A data-driven reduced-order model of nonlinear processes based on diffusion maps and artificial neural networks," *Chemical Engineering Journal*, vol. 397, p. 125 475, 2020. DOI: 10.1016/j.cej.2020.125475.

[221] E. D. Koronaki, N. Evangelou, Y. M. Psarellis, A. G. Boudouvis, and I. G. Kevrekidis, "From partial data to out-of-sample parameter and observation estimation with diffusion maps and geometric harmonics," *Computers & Chemical Engineering*, p. 108 357, 2023. DOI: 10.1016/j.compchemeng.2023.108357.

[222] A. F. Brouwer and M. C. Eisenberg, *The underlying connections between identifiability, active subspaces, and parameter space dimension reduction*, 2018. DOI: 10.48550/arXiv.1802.05641.

[223] N. Evangelou, N. J. Wichrowski, G. A. Kevrekidis, F. Dietrich, M. Kooshkbaghi, S. McFann, and I. G. Kevrekidis, "On the parameter combinations that matter and on those that do not: Data-driven studies of parameter (non)identifiability," *PNAS Nexus*, vol. 1, no. 4, K. E. Nelson, Ed., pgac154, 2022. DOI: 10.1093/pnasnexus/pgac154.

[224] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, RoLoD: Robust Local Descriptors for Computer Vision 2014, vol. 184, pp. 232–242, 2016. DOI: 10.1016/j.neucom.2015.08.104.

[225] C. Fraley and A. E. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002. DOI: 10.1198/016214502760047131.

[226] H. Jia, S. Ding, X. Xu, and R. Nie, "The latest research progress on spectral clustering," *Neural Comput & Applic*, vol. 24, no. 7, pp. 1477–1486, 2014. DOI: 10.1007/s00521-013-1439-2.

[227] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963. DOI: 10.1080/01621459.1963.10500845.

[228] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat Mach Intell*, vol. 2, no. 1, pp. 56–67, 2020. DOI: 10.1038/s42256-019-0138-9.

[229] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014. DOI: 10.1109/TKDE.2013.39.

[230] L. Pellerin, G. Pellegri, P. G. Bittar, Y. Charnay, C. Bouras, J.-L. Martin, N. Stella, and P. J. Magistretti, "Evidence supporting the existence of an activity-dependent astrocyte-neuron lactate shuttle," *Developmental neuroscience*, vol. 20, no. 4-5, pp. 291–299, 1998. DOI: 10.1159/000017324.

[231] H. Kitano, "Computational systems biology," *Nature*, vol. 420, no. 6912, pp. 206–210, 2002. DOI: 10.1038/nature01254.

[232] R. Sánchez-Alvarez, A. Tabernero, and J. M. Medina, "Endothelin-1 stimulates the translocation and upregulation of both glucose transporter and hexokinase in astrocytes: Relationship with gap junctional communication," *Journal of neurochemistry*, vol. 89, no. 3, pp. 703–714, 2004. DOI: 10.1046/j.1471-4159.2004.02398.x.

[233] J. G. Jackson and M. B. Robinson, "Reciprocal regulation of mitochondrial dynamics and calcium signaling in astrocyte processes," *Journal of Neuroscience*, vol. 35, no. 45, pp. 15 199–15 213, 2015. DOI: 10.1523/JNEUROSCI.2049-15.2015.

[234] C. M. Agapakis, P. M. Boyle, and P. A. Silver, "Natural strategies for the spatial optimization of metabolism in synthetic biology," *Nature chemical biology*, vol. 8, no. 6, pp. 527–535, 2012. DOI: 10.1038/nchembio.975.

[235] T. Alexandrov, "Spatial metabolomics: From a niche field towards a driver of innovation," *Nature Metabolism*, vol. 5, no. 9, pp. 1443–1445, 2023. DOI: 10.1038/s42255-023-00881-0.

[236] M. Tantama and G. Yellen, "Imaging changes in the cytosolic ATP-to-ADP ratio," *Methods in enzymology*, vol. 547, pp. 355–371, 2014. DOI: 10.1016/B978-0-12-801415-8.00017-5.

[237] R. Suzuki, K. Hotta, and K. Oka, "Spatiotemporal quantification of subcellular ATP levels in a single HeLa cell during changes in morphology," *Scientific reports*, vol. 5, no. 1, p. 16 874, 2015. DOI: 10.1038/srep16874.

[238] F. Llaneras and J. Picó, "Stoichiometric modelling of cell metabolism," *Journal of bioscience and bioengineering*, vol. 105, no. 1, pp. 1–11, 2008. DOI: 10.1263/jbb.105.1.

[239] S. Cortassa, M. A. Aon, E. Marbán, R. L. Winslow, and B. O'Rourke, "An integrated model of cardiac mitochondrial energy metabolism and calcium dynamics," *Biophysical journal*, vol. 84, no. 4, pp. 2734–2755, 2003. DOI: 10.1016/S0006-3495(03)75079-6.

[240] A. Aubert and R. Costalat, "Interaction between astrocytes and neurons studied using a mathematical model of compartmentalized energy metabolism," *Journal of Cerebral Blood Flow & Metabolism*, vol. 25, no. 11, pp. 1476–1490, 2005. DOI: 10.1038/sj.jcbfm.9600144.

[241] A. Aubert, R. Costalat, P. J. Magistretti, and L. Pellerin, "Brain lactate kinetics: Modeling evidence for neuronal lactate uptake upon activation," *Proceedings of the National Academy of Sciences*, vol. 102, no. 45, pp. 16 448–16 453, 2005. DOI: 10.1073/pnas.0505427102.

[242] A. Szabó and R. M. Merks, "Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution," *Frontiers in oncology*, vol. 3, p. 87, 2013. DOI: 10.3389/fonc.2013.00087.

[243] F. Cleri, "Agent-based model of multicellular tumor spheroid evolution including cell metabolism," *The European Physical Journal E*, vol. 42, pp. 1–15, 2019. DOI: 10.1140/epje/i2019-11878-7.

[244] M. U. Khalid, A. Tervonen, I. Korkka, J. Hyttinen, and K. Lenk, "Geometry-based computational modeling of calcium signaling in an astrocyte," in *EMBEC & NBC 2017: Joint Conference of the European Medical and Biological Engineering Conference (EMBEC) and the Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC), Tampere, Finland, June 2017*, Springer, 2018, pp. 157–160. DOI: 10.1007/978-981-10-5122-7_40.

[245] M. Bell, T. Bartol, T. Sejnowski, and P. Rangamani, "Dendritic spine geometry and spine apparatus organization govern the spatiotemporal dynamics of calcium," *Journal of General Physiology*, vol. 151, no. 8, pp. 1017–1034, 2019. DOI: 10.1085/jgp.201812261.

[246] A. J. Ellingsrud, A. Solbrå, G. T. Einevoll, G. Halnes, and M. E. Rognes, "Finite element simulation of ionic electrodiffusion in cellular geometries," *Frontiers in Neuroinformatics*, vol. 14, p. 11, 2020. DOI: 10.3389/fninf.2020.00011.

[247] G. C. Garcia, K. Gupta, T. M. Bartol, T. J. Sejnowski, and P. Rangamani, "Mitochondrial morphology governs ATP production rate," *Journal of General Physiology*, vol. 155, no. 9, e202213263, 2023.

[248] M. Kobayashi, M. Takeda, T. Sato, Y. Yamazaki, K. Kaneko, K.-I. Ito, H. Kato, and H. Inaba, "In vivo imaging of spontaneous ultraweak photon emission from a rat's brain correlated with cerebral energy metabolism and oxidative stress," *Neuroscience research*, vol. 34, no. 2, pp. 103–113, 1999. DOI: 10.1016/S0168-0102(99)00040-1.

[249] G. Zampieri, S. Vijayakumar, E. Yaneske, and C. Angione, "Machine and deep learning meet genome-scale metabolic modeling," *PLoS computational biology*, vol. 15, no. 7, e1007084, 2019. DOI: 10.1371/journal.pcbi.1007084.

[250] M. Alber, A. Buganza Tepole, W. R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W. W. Lytton, P. Perdikaris, *et al.*, "Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences," *NPJ digital medicine*, vol. 2, no. 1, p. 115, 2019. DOI: 10.1038/s41746-019-0193-y.

[251] J. D. Murray, *Mathematical Biology: I. An introduction*. Springer, 2002.

[252] T. J. Hughes, *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Courier Corporation, 2012.

[253] M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, *et al.*, "The FEniCS project version 1.5," *Archive of numerical software*, vol. 3, no. 100, 2015.

[254] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*. Springer Science & Business Media, 2008, vol. 23.

[255] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[256] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, *Keras Tuner*, 2019.

[257] C.-H. Chu, W.-W. Tseng, C.-M. Hsu, and A.-C. Wei, "Image analysis of the mitochondrial network morphology with applications in cancer research," *Frontiers in Physics*, vol. 10, p. 855 775, 2022.

[258] K. Kontolati, S. Goswami, G. Em Karniadakis, and M. D. Shields, "Learning nonlinear operators in latent spaces for real-time predictions of complex dynamics in physical systems," *Nat Commun*, vol. 15, no. 1, p. 5101, 2024. DOI: 10.1038/s41467-024-49411-w.

[259] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, *et al.*, *Perceiver IO: A General Architecture for Structured Inputs & Outputs*, version 3, 2021. DOI: 10.48550/ARXIV.2107.14795, pre-published.

[260] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018. DOI: 10.1111/ectj.12097.

[261] V. Chernozhukov, C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis, *Applied Causal Inference Powered by ML and AI*, 2024. DOI: 10.48550/arXiv.2403.02467, pre-published.