National Technical University of Athens
School of Chemical Engineering

PhD Thesis

Development of computational methods for the prediction of material properties

Dimitra - Danai Varsou
Chemical Engineer

Athens, July 2021

National Technical University of Athens
School of Chemical Engineering

PhD Thesis

Development of computational methods for the prediction of material properties

Dimitra - Danai Varsou
Chemical Engineer

Athens, July 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Χημικών Μηχανικών

Διδακτορική Διατριβή

Ανάπτυξη υπολογιστικών μεθόδων για την πρόβλεψη ιδιοτήτων υλικών

Δήμητρα - Δανάη Βάρσου
Χημικός Μηχανικός

Αθήνα, Ιούλιος 2021

Η έγκριση της διδακτορικής διατριβής από την Ανώτατη Σχολή Χημικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου δεν υποδηλώνει αποδοχή των γνωμών της συγγραφέως. (Ν. 5343/1932, Άρθρο 202)

The Dissertation is submitted for approval to the seven-member evaluation committee consisting of the following members:

Haralambos Sarimveis Supervisor

Professor, School of Chemical Engineering,

National Technical University of Athens

Eugenia Valsami-Jones Member of the Advisory Committee

Professor, School of Geography, Earth and Environmental Sciences,

University of Birmingham

Constantinos Charitidis Member of the Advisory Committee

Professor, School of Chemical Engineering,

National Technical University of Athens

Doros Theodorou

Professor, School of Chemical Engineering,

National Technical University of Athens

Fotios Tsopelas

Assistant Professor, School of Chemical Engineering,

National Technical University of Athens

Georgia Melagraki

Assistant Professor, Department of Military Sciences, Division of Physical Sciences and Applications,

Hellenic Military Academy

Iseult Lynch

Professor, School of Geography, Earth and Environmental Sciences,

University of Birmingham

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 637)

& by the Onassis Foundation Scholarship Program (Scholarship Code: G ZN 008-1/2017-2018).

Contact information

dimitra.varsou@gmail.com RG Dimitra-Danai Varsou DemetraDanae
Dimitra-Danai Varsou Dimitra-Danai Varsou demetradanae

Cover photo by Cat Han on Unsplash

The present Dissertation is based on the following publications in peer-reviewed scientific journals and international conferences:

List of scientific papers

Dimitra-Danai Varsou, Antreas Afantitis, Andreas Tsoumanis, Georgia Melagraki, Haralambos Sarimveis, Eugenia Valsami-Jones and Iseult Lynch, *A safe-by-design tool for functionalised nanomaterials through the Enalos Nanoinformatics Cloud platform*, Nanoscale Advances, **2019**, 1, 706-718. Accessible via: pubs.rsc.org/en/content/articlehtml/2019/na/c8na00142a [1]

Dimitra-Danai Varsou, Antreas Afantitis, Georgia Melagraki and Haralambos Sarimveis, *Read-across predictions of nanoparticle hazard endpoints: a mathematical optimization approach*, Nanoscale Advances, **2019**, 1, 3485-3498. Accessible via: pubs.rsc.org/en/content/articlehtml/2019/na/c9na00242a [2]

Dimitra-Danai Varsou, Antreas Afantitis, Andreas Tsoumanis, Anastasios Papadiamantis, Eugenia Valsami-Jones, Iseult Lynch, and Georgia Melagraki, *Zeta-Potential Read-Across Model Utilizing Nanodescriptors Extracted via the NanoXtract Image Analysis Tool Available on the Enalos Nanoinformatics Cloud Platform*, Small, **2020**, 16, 1906588. Accessible via: onlinelibrary.wiley.com/doi/full/10.1002/smll.201906588 [3]

Dimitra-Danai Varsou and Haralambos Sarimveis, *Apellis: an online tool for read-across model development*, Computational Toxicology, **2020**, 17. Accessible via: www.sciencedirect.com/science/article/pii/S2468111320300566 [4]

Dimitra-Danai Varsou, Nikoletta-Maria Koutroumpa and Haralambos Sarimveis, *Automated grouping of nanomaterials and read-across prediction of their adverse effects based on mathematical optimization*, Journal of Chemical Information and Modeling, **2021**, 61, 6, 2766–2779. Accessible via: pubs.acs.org/doi/10.1021/acs.jcim.1c00199 [5]

List of conference presentations

Dimitra-Danai Varsou, Georgia Melagraki, Haralambos Sarimveis and Antreas Afantitis, *Read-across in silico investigation of the bioactivity and toxicity behaviour of carbon nanotubes*, 3rd Nanosafety Forum for Young Scientists, Valletta, Malta, 10 and 11 of September **2018**. (Poster)

Dimitra-Danai Varsou, Antreas Afantitis, Georgia Melagraki and Haralambos Sarimveis, *Development and solution of a mathematical optimization problem for the prediction of undesired nanomaterials properties*, 12th National Scientific Conference on Chemical Engineering, Athens, Greece, 29 and 31 of May **2019**. (Oral presentation in Greek)

Nikoletta-Maria Koutroumpa, Dimitra-Danai Varsou and Haralambos Sarimveis, *Read-across automated grouping and hazard endpoint predictions of nanoparticles based on mathematical optimization*, 1st International Young Scientist Forum, Salzburg, Austria, 09 and 10 of September **2019**. (Oral presentation)

Καινούριους τόπους δεν θα βρεις, δεν θάβρεις άλλες θάλασσες.
Η πόλις θα σε ακολουθεί.

Απόσπασμα από το ποίημα του Κωνσταντίνου Π. Καβάφη:
Η Πόλις

Prologue and Acknowledgements

The only thing -I guess- I dislike about Academia, is the fact that I like it so much. Thus, the decision of starting a PhD came naturally, especially due to the fact that I had a constructive and creative research experience when conducting my undergraduate Diploma Thesis.

I always considered that my PhD apart from a simple research project, was an opportunity to unravel my creativity and even my "artistic" nature. I also have to admit that it was a milestone in my personal path to maturity. In this trajectory, I was lucky enough to have by my side, people that supported me and, in their way, helped me overcome any difficulty.

To begin with, I would like to express my gratitude to Professor Haralambos Sarimveis for trusting me and for guiding me during all these years. I have to thank him for this interesting and challenging topic, for giving me freedom when dealing with the timeplan, for his patience in correcting every single detail of my reports and of course of understanding me, when I refused to deal with subjects that were not particularly attractive to me. Following, I would like to thank Professors Eugenia Valsami-Jones and Constantinos Charitidis for their participation in my Advisory Committee.

My PhD experience would not be unique, if I hadn't collaborated in my first PhD years with Professor Georgia Melagraki and Dr. Antreas Afantitis in NovaMechanics Ltd. Through this professional experience apart from working on challenging computational projects, I had the chance to develop my communication skills, and most importantly to visit and work in other European countries and international environments. I am also really grateful towards Georgia and Antreas for their support and guidance from their professional perspective. In addition, I would like to thank my ex-colleague Dr. Andreas Tsoumanis for his great help in the development of the two web services presented in Chapter 7.

Through my international "tour" I had the opportunity to visit the University of Birmingham, where I worked with Professors Eugenia Valsami-Jones and Iseult Lynch and their research teams. I would like to thank them -among others- for their genuine interest in my research and their support when I was stressed. Of course, this visit would be incomplete, if I hadn't met and collaborated with Dr. Anastasios Papadiamantis, who always found a way to cheer me up with his cool way of thinking.

Another person that I would like to wholeheartedly thank is Nikoletta-Maria (Niki) Koutroumpa, who carried out her Diploma Thesis in the Unit of Process Control and Informatics. Niki and I collaborated exceptionally in the development of an automated ENMs grouping methodology (presented in Chapter 5) and due to her patience and persistence (especially when we had to perform long-lasting runs), we were able to produce a high-quality methodology for ENMs grouping.

Furthermore, I would like to express my gratitude and esteem to Professor Antonis Karantonis for welcoming me all these years at the Laboratory of Physical Chemistry and Applied Electrochemistry in NTUA.

Finally, I would like to thank my colleagues in the Unit of Process Control and Informatics: Eleni Strompoula, Periklis Tsiros, Pantelis Karatzas and Marianna Kotzabasaki for offering me a friendly working environment. I would like to give my special thanks to Pantelis for helping me with various technicalities and for encouraging me to learn Python and to use

the Docker for the deployment of my web applications.

Moving on, I would like to express my gratitude to my group of friends and family who supported me in my brightest and darkest days throughout this time. Their help was sometimes more important and necessary in order to overcome any obstacles in my research. Therefore, I would like to thank my friend Dimitris Z. who always listened about my everyday troubles without complaining and always supported me morally, even with his own example and ethos. I would also like to thank my beloved friends Maria Z. and Thodoris P. who, about ten years now, still understand me, care about me, laugh with me and proved that distance is just a number.

I would also like to thank for their company and moral support my friends, George S., Nicholas K., Spyros N. and Takis K. Special thanks I would like to give to Lefteris D. for his help the moment I didn't know I needed it. Next, I would like to thank Raquel G.-F. for insisting on taking a break from time to time for travelling and relaxation. Special thanks I have to give to my childhood friends Maria K. and Eleni D., who always "tolerate" me, support me and have fun with me through the years. I also like to thank Luminița M. for her unconditional interest, help and care.

At this point I would like to express my gratitude to my extraordinary family, who trusts me and supports me with every thoughtful or reckless decision I make. I would like to thank my best friend and brother, Thanasis, for being supportive and patient with anything that happens to me. I would like to thank my mother, Stamatoula, for her brave character, for teaching me the importance of being independent and for showing me that it is never too late to pursue your dreams. Finally, I would like to thank my father, Dimos, for transmitting me his passion for knowledge and comprehension of simple or complicated issues, and his enthusiasm for scientific research. I would be lost without their love.

I hope you enjoy reading this manuscript, as much I enjoyed conducting this Thesis.

Dimitra-Danai Varsou
June 2021

Abstract

The main objective of this PhD program is the development of innovative computational read-across methods for predicting engineered nanomaterial (ENM) properties (with emphasis to toxicity-related endpoints), based on experimental data. The read-across methods aim at determining neighbours (similar samples) to the query ENM in a dataset of ENMs with known properties and creating groups of related substances that have similar biological activity or toxic response.

An important step in all the developed methodologies is the selection of the properties that are relevant to the endpoint of interest, to reduce the dimensionality of the models, avoid over-fitting and generate interpretable models. The automation of all the modelling parameters, is a key goal in this research project, and the proposed methodologies require the minimum information from the users to produce valid and robust read-across models.

Special emphasis was given in the making of the models developed in this program available through repositories or via user-friendly web applications. Implementation of the models as web tools supports their dissemination and actual use by all stakeholders in real-life applications.

To begin with, a novel read-across methodology, related to the prediction of ENMs toxicity was developed. The method selects the most important variables and defines the neighbouring area around the target ENM, using single or multiple similarity criteria. The similarity criteria depend on the available ENM properties (e.g., physicochemical, biological, biokinetics etc.). The read-across prediction is computed as the weighted average of the neighbour ENMs. This novel grouping approach is based on the formulation and the solution of a mixed integer non-linear mathematical programming problem. A specific genetic algorithm scheme was developed to compute an approximate solution, due to the complexity of the problem rendering it practically unsolvable by conventional mathematical algorithms.

The second method constructs, a mixed integer-linear optimisation program, which automatically filters out the noisy variables, defines the grouping boundaries based on one of the available properties -which is automatically chosen- and develops specific to each group LASSO linear regression predictive models. The third computational workflow is based on the formulation of a mathematical optimisation methodology that groups the ENMs into regions -according to their endpoint value-, removes the noisy variables, and incorporates the LASSO method for training predictive linear models specific to each region.

Finally, *k*-Nearest Neighbours machine learning methodology was applied for deriving read-across models predicting the cytotoxicity and the biological activity of decorated multi-walled carbon nanotubes using calculated molecular descriptors of their surface ligands, and the zeta-potential of ENMs using geometrical ENMs properties extracted form transmission electron microscopy images.

All developed methodologies were applied and validated on benchmark datasets, based on OECD principles, and were compared with methodologies already presented in Literature. They proved to be comparable and, in several cases, outperformed other alternative predictive modelling techniques, illustrating this way their good predictive performance and capabilities. Taking also into account that the grouping, feature selection and model generation steps are fully automated, the proposed methods can be considered as promising new approaches in

the field of grouping/read-across modelling.

Keywords Nanoinformatics, read-across, engineered nanomaterials, predictive modelling, safety-by-design, web applications

Περίληψη στα ελληνικά

Εισαγωγή

Οι δομές υλικών σε διαστάσεις 1-100 nm, γνωστές και ως νανοϋλικά (ΝΥ), χάρη ακριβώς στις διαστάσεις αυτές μεταξύ ατόμων και μορίων, αποκτούν νέες και ρυθμίσιμες ιδιότητες σε σύγκριση με αυτές του ίδιου υλικού σε μακροκλίμακα. Ήδη οι μοναδικές ιδιότητες των νανοϋλικών έχουν αξιοποιηθεί σε εφαρμογές στους καταλύτες, στα δομικά υλικά, στις ηλεκτρονικές συσκευές, στους αισθητήρες και στον τομέα των καλλυντικών. Ωστόσο οι ίδιες αυτές ιδιότητες είναι ζωτικής σημασίας για τη συμπεριφορά των ΝΥ κατά τη διάρκεια των διαφόρων σταδίων παραγωγής, επεξεργασίας και τελικής εφαρμογής, καθώς και για τις πιθανές αλληλεπιδράσεις με το περιβάλλον και τον άνθρωπο.

Τα ΝΥ έχουν τη δυνατότητα να εισέρχονται στην κυκλοφορία του αίματος, να φτάνουν στους ιστούς, στα κύτταρα και στα οργανίδια, δηλαδή σε λειτουργικές βιολογικές δομές στις οποίες μεγαλύτερα σωματίδια δε θα είχαν πρόσβαση. Τοξικολογικές μελέτες καταδεικνύουν πως τα ΝΥ είναι δυνητικά επιβλαβή για τους οργανισμούς: μπορούν να διαπερνούν τα κύτταρα μέσω της κυτταρικής μεμβράνης, να συσσωρεύονται σε αυτά ή ακόμα και στον πυρήνα τους. Επίσης τα ΝΥ είναι ικανά να προκαλούν φλεγμονώδεις αποκρίσεις, να αναστέλλουν την κυτταρική ανάπτυξη και να προκαλούν κυτταρικό θάνατο (κυτοτοξικότητα). Επίσης ενδέχεται να οδηγήσουν στην παραγωγή δραστικών μορφών οξυγόνου (Reactive Oxygen Species, ROS), όπως οι ελεύθερες ρίζες, οι οποίες προκαλούν οξειδωτικό στρες, το οποίο είναι υπεύθυνο για βλάβες στο DNA, στις πρωτεΐνες και στα λιπίδια. Τέλος αναφέρονται συχνά περιπτώσεις νευροτοξικότητας και καρκινογενέσεων, λόγω της αλληλεπίδρασης κυττάρων με ΝΥ. Ως εκ τούτου, τα ίδια φυσικοχημικά χαρακτηριστικά που τους προσδίδουν μοναδικές ιδιότητες που αξιοποιούνται σε εμπορικές και ερευνητικές εφαρμογές, μπορούν να τους προσδώσουν και τοξικές ιδιότητες στα βιολογικά συστήματα, που μάλιστα διαφέρουν ανάλογα με τον τύπο της εκάστοτε νανοδομής.

Η πειραματική μελέτη όλων των παραμέτρων της τοξικότητας ενός μόνο τύπου ΝΥ αποτελεί μια χρονοβόρα και πολυέξοδη διαδικασία, γεγονός που την καθιστά ασύμφορη λαμβάνοντας υπόψιν τους διαφορετικούς τύπους των ΝΥ που παράγονται καθημερινά και χρησιμοποιούνται σε διαφορετικές εφαρμογές. Ακόμη η χρήση πειραματόζωων στις πειραματικές μελέτες εγείρει προβληματισμούς σχετικά με τις ψυχοφθόρες επιπτώσεις της χρήσης και θανάτωσής τους στους ίδιους τους ερευνητές. Τέλος, η σύγχρονη ευρωπαϊκή νομοθεσία και οι τάσεις στο πεδίο έρευνας της τοξικότητας των χημικών ουσιών επιτάσσουν την ελαχιστοποίηση των πειραμάτων σε πειραματόζωα (*in vivo*) και την αντικατάστασή τους με πειράματα σε κυτταρικές καλλιέργειες (*in vitro*) αλλά και με μη πειραματικές τεχνικές (*in silico*) με τις οποίες θα γίνεται πρόβλεψη της τοξικότητάς τους.

Κατά συνέπεια δίδεται πρόσφορο έδαφος στο πεδίο της πληροφορικής να αναπτύξει υπολογιστικές τεχνικές για την πρόβλεψη ιδιοτήτων των ΝΥ, συμπεριλαμβανομένων της βιολογικής συμπεριφοράς και της τοξικότητάς τους. Η νανοπληροφορική εστιάζει στην ανάπτυξη υπολογιστικών μοντέλων που θα προβλέπουν με ακρίβεια τις επιβλαβείς ιδιότητες των ΝΥ και ταυτόχρονα επιδιώκει την ανάπτυξη εργαλείων φιλικών-προς-το-χρήστη, ώστε

οι πειραματικοί ερευνητές και οι ενδιαφερόμενοι στους ρυθμιστικούς φορείς να μπορούν να χρησιμοποιούν τα δεδομένα τους στα μοντέλα χωρίς να απαιτείται να έχουν προηγούμενο υπολογιστικό υπόβαθρο.

Υπολογιστικά εργαλεία για την πρόβλεψη ιδιοτήτων υλικών και της τοξικότητάς τους

Για την πρόβλεψη της τοξικότητας των ΝΥ έχει ήδη συζητηθεί η εφαρμογή μοντέλων Ποιοτικής ή Ποσοτικής Συσχέτισης Δομής-Ιδιοτήτων (Qualitative or Quantitative Structure-Activity Relationship models, QSAR), και σε περιπτώσεις εφαρμογής τους σε διάφορους τύπους ΝΥ η πρόβλεψη ήταν επιτυχής. Τα μοντέλα αυτά περισσότερο πλέον γνωστά ως nanoQSAR ή QNAR (Qualitative or Quantitative Nanostructure-Activity Relationship models) βασίζονται κυρίως στην προηγούμενη γνώση που προσφέρει το πεδίο της χημειοπληροφορικής, όπου μοντέλα είχαν αναπτυχθεί και συνεχίζουν να αναπτύσσονται για την πρόβλεψη ιδιοτήτων μικρών οργανικών μορίων.

Τα μοντέλα nanoQSAR έχουν ωστόσο αδυναμίες: αφενός απαιτούν μεγάλο σύνολο δεδομένων για να εκπαιδευτούν, αλλιώς ελλοχεύει ο κίνδυνος της υπερπροσαρμογής του μοντέλου στα δεδομένα, δηλαδή είναι πιθανό να μοντελοποιηθεί ακόμη και το σφάλμα των δεδομένων εκπαίδευσης. Αφετέρου βασίζονται στην ύπαρξη ενός μοναδικού μηχανισμού τοξικότητας για να γίνει η μοντελοποίηση, γεγονός που δεν ανταποκρίνεται στην πραγματικότητα. Τα ΝΥ δεν είναι δομικά ομοιογενή και ως εκ τούτου δεν αναμένεται να υπάρχει ένας κοινός μηχανισμός τοξικότητας. Ενδεικτικά στη βιβλιογραφία αναφέρονται τέσσερις κυρίαρχοι μηχανισμοί τοξικότητας:

- Η απελευθέρωση τοξικών χημικών συστατικών λόγω της διάλυσης των ΝΥ,
- Οι άμεσες επιπτώσεις από τη φυσική επαφή με τα ΝΥ, οι οποίες συσχετίζονται με το μέγεθος, το σχήμα και τις επιφανειακές τους ιδιότητες, και οι οποίες μπορεί να προκληθούν, για παράδειγμα, από την αλλαγή στη δομή των βιομορίων που έρχονται σε επαφή με αυτά,
- Οι οξειδοαναγωγικές επιπτώσεις που ενδεχομένως προκύπτουν από την κρυσταλλική δομή των ΝΥ και,
- Η ικανότητα των ΝΥ να ενεργούν ως φορείς για τη μεταφορά άλλων τοξικών χημικών ουσιών σε ευαίσθητους ιστούς (φαινόμενο του Δούρειου Ίππου).

Η στρατηγική read-across

Δεδομένου ότι τα μοντέλα nanoQSAR αφήνουν περιθώρια αμφισβήτησης ως προς την αξιοπιστία των προβλέψεών τους όταν τα πειραματικά δεδομένα σπανίζουν, η επιστημονική κοινότητα στρέφεται προς εναλλακτικές τεχνικές που βασίζονται στην πρόβλεψη ιδιοτήτων στα πλαίσια συγκεκριμένων ομάδων επαρκώς παρόμοιων ΝΥ που αναμένεται να έχουν παρόμοιες ιδιότητες (μεθοδολογία read-across). Η μεθοδολογία read-across για την πρόβλεψη ιδιοτήτων ενός υλικού βασίζεται στη χρήση δεδομένων συγγενών υλικών με γνωστές ιδιότητες. Κατ' αυτό τον τρόπο είναι δυνατόν να περιοριστεί η πρόβλεψη σε μια μικρή περιοχή του χώρου δεδομένων και κατ' επέκταση να μην αποτελεί πλέον ανάγκη η ύπαρξη μεγάλων συνόλων δεδομένων. Στη συνέχεια η πρόβλεψη μπορεί να επιτευχθεί εφαρμόζοντας «τοπικά» μια μεθοδολογία συσχέτισης εισόδου-εξόδου.

Η μεθοδολογία read-across έχει εφαρμοστεί επιτυχώς σε προβλέψεις ιδιοτήτων και τοξικότητας καρβονυλικών ενώσεων, φωσφο-οργανικών παρασιτοκτόνων, πολικών οργανικών και άλλων χημικών ενώσεων, ωστόσο δεν έχει προχωρήσει σε ισοδύναμο βαθμό

στην πρόβλεψη τοξικότητας ΝΥ. Καθώς οι πρώτες μεθοδολογίες read-across βρίσκονται σε πολύ αρχικά στάδια, το συγκεκριμένο πεδίο έρευνας είναι γόνιμο για την ανάπτυξη και δοκιμή νέων και πρωτότυπων ιδεών στα πλαίσια του read-across. Άλλωστε ο Ευρωπαϊκός Οργανισμός Χημικών Προϊόντων (ΕΟΧΠ) εξέδωσε τον αντίστοιχο κανονισμό Read-Across Assessment Framework όπου περιγράφονται με σαφήνεια και συνέπεια οι αρχές που διέπουν την εν λόγω μεθοδολογία, ώστε όλο και περισσότεροι ερευνητές να ενθαρρύνονται και να διευκολύνονται στην ένταξη του read-across στην πρόβλεψη των ανεπιθύμητων ιδιοτήτων νανοδομών. Επίσης μεθοδολογίες read-across προτείνονται εκτός από την πρόβλεψη της τοξικότητας καθαυτής, και για την πρόβλεψη άλλων ιδιοτήτων όταν υπάρχουν «κενά» στις βάσεις δεδομένων.

Δεδομένου ότι δεν υπάρχει ένας μοναδικός μηχανισμός τοξικότητας και κατά συνέπεια δεν ενδείκνυται η ξεχωριστή μελέτη κάθε ΝΥ, οι ερευνητές προτείνουν την ομαδοποίησή τους (grouping) σε κατηγορίες ανάλογα με τη σύσταση, τα δομικά, τα επιφανειακά και άλλα χαρακτηριστικά και την πρόβλεψη της τοξικότητας στα πλαίσια των ομάδων αυτών. Τα ΝΥ μπορούν να ομαδοποιηθούν αρχικά με βάση τις χημικές τους ιδιότητες (σύνθεση, επιφανειακές ιδιότητες), τις φυσικοχημικές τους ιδιότητες (μέγεθος, σχήμα, ενεργή επιφάνεια), τα χαρακτηριστικά της συμπεριφοράς τους (υδροφοβικότητα, διαλυτότητα, ικανότητα διασποράς, δυναμικό-ζ) και τον τρόπο δράσης τους (βιολογική επίδραση, φωτοχημική επίδραση, τοξική επίδραση).

Προκειμένου να εναρμονιστούν οι διάφορες τεχνικές read-across που προτείνονται από τις διάφορες ερευνητικές ομάδες, ο ΕΟΧΠ πρότεινε μια σειρά επτά βημάτων που πρέπει να ακολουθηθούν ώστε να συγκροτούνται ομάδες παρόμοιων ΝΥ μέσα στις οποίες θα μπορεί να γίνει η πρόβλεψη της ιδιότητας ενδιαφέροντος.[1] Ο ακρογωνιαίος λίθος μιας αποδεκτής διαδικασίας ομαδοποίησης είναι η διαμόρφωση μιας «υπόθεσης»-σεναρίου σύμφωνα με την οποία ταξινομούνται τα διάφορα δείγματα ΝΥ σε κλάσεις (Εικόνα 1.3) και συσχετίζονται οι γνωστές ιδιότητες εισόδου του μοντέλου (π.χ. πειραματικά δεδομένα) με την ιδιότητα εξόδου (π.χ. τοξικότητα).

Η υπόθεση ομαδοποίησης περιλαμβάνει δύο σκέλη: αφενός την επιλογή εκείνων των ιδιοτήτων οι οποίες καταδεικνύουν τις ομοιότητες μεταξύ των δειγμάτων ΝΥ, και αφετέρου τον σαφή καθορισμό των ορίων μεταξύ των ομάδων. Η διαμόρφωση της υπόθεσης αυτής απαιτεί μια χρονοβόρα διαδικασία δοκιμής και σφάλματος, συμπεριλαμβανομένης της πειραματικής συλλογής δεδομένων, έως ότου επιτευχθεί σύγκλιση σε μια επιτυχημένη αλλά ακόμα μη βέλτιστη υπόθεση ομαδοποίησης.

Προτεινόμενες μεθοδολογίες

Οι μέθοδοι read-across που αναπτύχθηκαν στη Διατριβή, στοχεύουν στην εξεύρεση γειτόνων (ΝΥ με παρόμοιες ιδιότητες) μεταξύ ενός συνόλου ΝΥ, λαμβάνοντας ταυτόχρονα υπόψιν μόνο τις χρήσιμες από τις διαθέσιμες πειραματικές ιδιότητές τους, προκειμένου να μην προκύπτουν υπερπροσαρμοσμένα μοντέλα. Η διαδικασία του σχηματισμού των ομάδων, της επιλογής μεταβλητών αλλά και της βελτιστοποίησης των παραμέτρων του τελικού προβλεπτικού μοντέλου, σχεδιάζονται και εκτελούνται με μια αυτοματοποιημένη διαδικασία, ώστε να παράγονται αξιόπιστα μοντέλα με τη λιγότερη αλληλεπίδραση με τον τελικό χρήστη. Η αυτοματοποίηση των διαδικασιών καταργεί την ανάγκη για τη διαδοχική εξέταση διαφόρων σεναρίων ομαδοποίησης, αφού τα όρια των βέλτιστων ομάδων προκύπτουν ως αποτέλεσμα της εφαρμογής των μεθοδολογιών. Στη

[1]Στο παρόν κομμάτι της Διατριβής οι όροι, «εξαρτημένη μεταβλητή», «έξοδος» και «υπό εξέταση ιδιότητα» ή «ιδιότητα ενδιαφέροντος» εναλλάσσονται για να αποδώσουν τον όρο «endpoint» ή «μεταβλητή απόκρισης». Το endpoint είναι η ιδιότητα εκείνη την τιμή της οποίας επιδιώκουν να προβλέψουν οι ερευνητές μέσω της ανάπτυξης ενός μοντέλου (π.χ. κάποια τιμή ή κλάση τοξικότητας, μια βιολογική απόκριση, μια πειραματικά-μετρούμενη ιδιότητα κ.λπ.). Εάν η ιδιότητα αυτή λαμβάνει αριθμητικές τιμές, τότε το μοντέλο που αναπτύσσεται ονομάζεται μοντέλο παλινδρόμησης, ενώ εάν λαμβάνει κατηγορικές τιμές, ονομάζεται μοντέλο ταξινόμησης.

Διατριβή παρουσιάστηκαν δύο κυρίαρχες ιδέες ομαδοποίησης στις οποίες βασίστηκαν οι προτεινόμενες μεθοδολογίες: η ομαδοποίηση χρησιμοποιώντας ένα ή περισσότερα κατώφλια ομοιότητας και η ομαδοποίηση με βάση τους k πλησιέστερους γείτονες.

Ομαδοποίηση μέσω κατωφλιών

Στην μεθοδολογία αυτή τα δείγματα ΝΥ τοποθετούνται στον πολυδιάστατο χώρο, με βάση τις συντεταγμένες που ορίζουν οι τιμές των μεταβλητών/ιδιοτήτων τους. Για κάθε νέα παρατήρηση που εισέρχεται στον πολυδιάστατο χώρο, υπολογίζονται οι αποστάσεις από όλα τα ΝΥ που ήδη βρίσκονται εκεί. Στη συνέχεια με βάση την τιμή ενός κατωφλιού, γείτονες θεωρούνται όσα ΝΥ απέχουν μικρότερη απόσταση από την τιμή του κατωφλιού (Εικόνα 1.4). Η πρόβλεψη της κλάσης ή της τιμής της ιδιότητας-εξόδου υπολογίζεται με βάση την πλειοψηφία των κλάσεων των γειτόνων (ταξινόμηση) ή τον μέσο όρο της ιδιότητας-εξόδου αντίστοιχα (παλινδρόμηση). Συχνά οι γείτονες συμμετέχουν στην πρόβλεψη με κάποιο συντελεστή βαρύτητας ανάλογα με την απόστασή τους από την υπό εξέταση παρατήρηση.

Υπάρχει δυνατότητα να τεθούν περισσότερα από ένα κατώφλια για την επιλογή των γειτόνων, στην περίπτωση που οι διαθέσιμες μεταβλητές μπορούν να ομαδοποιηθούν σε διάφορες κατηγορίες (για παράδειγμα στην περίπτωση των ΝΥ οι μεταβλητές μπορούν να κατηγοριοποιηθούν ανάλογα με το είδος των ιδιοτήτων που περιγράφουν: βιολογικές, φυσικοχημικές, θεωρητικά-υπολογισμένες κ.λπ.). Κατά συνέπεια μπορούν να υπολογιστούν και και διαφορετικά είδη αποστάσεων και να τεθούν τα αντίστοιχα κατώφλια. Δύο δείγματα ΝΥ θεωρούνται γείτονες, μόνο εάν οι υπολογισμένες αποστάσεις για κάθε ομάδα μεταβλητών ικανοποιούν όλες τις τιμές των αντίστοιχων κατωφλιών (Εικόνα 1.5).

Η επιλογή της τιμής ενός ή περισσοτέρων κατωφλιών μπορεί να προκύψει είτε «αυθαίρετα» στην αρχή της ανάλυσης, είτε αυτόματα μέσα από μια διαδικασία αριστοποίησης. Οι μεθοδολογίες που παρουσιάζονται στα Κεφάλαια 4-6 βασίζονται ακριβώς στην ιδέα ομαδοποίησης μέσω κατωφλιών και μάλιστα εστιάζουν στην αυτοματοποίηση και βελτιστοποίηση της εξεύρεσης των ορίων μεταξύ των ομάδων.

Αλγόριθμος των k-πλησιέστερων γειτόνων

Στην προσπάθεια εξεύρεσης μεθοδολογιών που θα παράγουν γρήγορες και αξιόπιστες προβλέψεις, ενσωματώθηκε και η μεθοδολογία μηχανικής μάθησης που βασίζεται στον αλγόριθμο των k-πλησιέστερων γειτόνων (k-Nearest Neighbours, kNN). Σε αυτή την περίπτωση τα ΝΥ τοποθετούνται στο χώρο των ιδιοτήτων και για κάθε άγνωστο δείγμα ΝΥ υπολογίζονται οι αποστάσεις από τα υπόλοιπα δείγματα ΝΥ (όπως και στην προηγούμενη περίπτωση) και επιλέγονται οι k-πλησιέστεροί του γείτονες, όπου k ακέραιος αριθμός (Εικόνα 1.7).

Η πρόβλεψη για το άγνωστο δείγμα προκύπτει από την πλειοψηφική κλάση μεταξύ των γειτόνων στην περίπτωση της ταξινόμησης, ή λαμβάνει τιμή ίση με τον μέσο όρο των τιμών της εξαρτημένης μεταβλητής των γειτόνων στην περίπτωση της παλινδρόμησης. Για να παραχθούν πιο ευαίσθητες προβλέψεις, είναι σύνηθες οι γείτονες να συμμετέχουν με μεγαλύτερο ή μικρότερο συντελεστή βαρύτητας στην παραγωγή των προβλέψεων με βάση την απόστασή τους από την υπό εξέταση παρατήρηση. Στο Κεφάλαιο 7 παρουσιάζονται δύο εφαρμογές της μεθοδολογίας kNN σε δεδομένα νανοπληροφορικής.

Αν και οι δύο τρόποι ομαδοποίησης έχουν κάποια κοινά χαρακτηριστικά, εμφανίζουν και κάποιες ουσιώδεις διαφορές. Σε αντίθεση με τον αλγόριθμο των kNN, στην ομαδοποίηση με χρήση κατωφλιών δεν είναι συγκεκριμένος ο αριθμός των γειτόνων που θα επιλεχθούν για κάθε νέα παρατήρηση. Επίσης αν το κατώφλι είναι αρκετά αυστηρό (σχετικά μικρή τιμή) υπάρχει πιθανότητα να επιλεγούν λίγοι ή και κανένας γείτονας για μια νέα παρατήρηση, και κατά συνέπεια να είναι αδύνατη η παραγωγή προβλέψεων. Με τη μέθοδο kNN πάντα

θα υπάρχει δυνατότητα πρόβλεψης για κάθε δείγμα ΝΥ, ακόμα και αν κάποιοι γείτονες που επιλεγούν έχουν μικρές μόνο ομοιότητες με το υπό εξέταση ΝΥ.

Τεχνικές επικύρωσης προβλεπτικών μεθοδολογιών

Στην προσπάθεια εξεύρεσης μιας συσχέτισης μεταξύ ενός συνόλου ιδιοτήτων και μιας ιδιότητας-εξόδου (εξαρτημένη μεταβλητή), εξετάζονται διάφορες μεθοδολογίες μοντελοποίησης ή γνωστοί αλγόριθμοι. Από αυτές τις μεθοδολογίες και τους αλγορίθμους, λίγοι είναι κατάλληλοι να συσχετίσουν τα δεδομένα εισόδου με την έξοδο του μοντέλου και κατ' επέκταση να παράγουν αξιόπιστες προβλέψεις. Εξάλλου, ο σκοπός της ανάπτυξης ενός μοντέλου είναι να χρησιμοποιηθεί για την παραγωγή προβλέψεων σε νέα δεδομένα τα οποία δεν έχουν ελεγχθεί ως προς την εξαρτημένη μεταβλητή. Για το λόγο αυτό, δεν ενδείκνυται να γίνεται επικύρωση του μοντέλου με κριτήριο την επιτυχία πρόβλεψης της τιμής της εξαρτημένης μεταβλητής στα ήδη γνωστά δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευσή του (ούτως ή άλλως η έξοδός τους είναι γνωστή πριν αναπτυχθεί το μοντέλο), διότι αυτό δεν δίνει καμία πληροφορία ως προς το πώς θα συμπεριφερθεί το μοντέλο σε νέα, άγνωστα δεδομένα και ως προς τη δυνατότητά του να γενικευτεί.

Για να ποσοτικοποιηθεί η επίδοση ενός μοντέλου σε νέα δεδομένα, χρειάζεται να μετρηθεί το ποσοστό των λανθασμένων προβλέψεων ή η απόκλιση από την πραγματική τιμή (ανάλογα με το είδος της εξόδου) με χρήση δεδομένων που επ' ουδενί δεν χρησιμοποιήθηκαν στην ανάπτυξη του μοντέλου. Χρειάζονται λοιπόν δύο ανεξάρτητα σύνολα δεδομένων (εξωτερική αξιολόγηση), ένα που θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου (training set) και ένα για την αξιολόγηση του μοντέλου (test set). Η επιλογή των δύο συνόλων γίνεται από ένα αρχικό σύνολο δεδομένων είτε με τυχαίο τρόπο είτε χρησιμοποιώντας μια μεθοδολογία αντιπροσωπευτικής δειγματοληψίας. Στη Διατριβή χρησιμοποιήθηκε σε μεγάλο βαθμό η επιλογή υποσυνόλων με τη μέθοδο Kennard-Stone: η συγκεκριμένη μεθοδολογία είναι αρκετά διαδεδομένη και εξασφαλίζει μια ομοιόμορφη επιλογή δειγμάτων ξεκινώντας από τα «περιφερειακά» δείγματα του πολυδιάστατου χώρου που ορίζεται από τις μεταβλητές του συνόλου.

Η παραπάνω διαδικασία συχνά εμφανίζεται παραλλαγμένη με τη χρήση τριών αντί δύο συνόλων δεδομένων: το σύνολο βαθμονόμησης (calibration), το σύνολο ελέγχου (validation) και το «τυφλό» σύνολο επαλήθευσης (test). Το σύνολο βαθμονόμησης χρησιμοποιείται όπως και στην περίπτωση του συνόλου εκπαίδευσης, για να βρεθούν οι παράμετροι του μοντέλου. Το σύνολο ελέγχου χρησιμοποιείται για να ελεγχθεί αν οι παράμετροι αυτές οδηγούν πράγματι σε αξιόπιστες προβλέψεις σε νέα δεδομένα. Στην περίπτωση που οι προβλέψεις είναι ικανοποιητικές, ολοκληρώνεται και η φάση της εκπαίδευσης και το μοντέλο είναι έτοιμο για χρήση. Σε αντίθετη περίπτωση, η διαδικασία της εκπαίδευσης επαναλαμβάνεται με νέες παραμέτρους έως ότου οι προβλέψεις στο σύνολο ελέγχου είναι ικανοποιητικές. Το σύνολο ελέγχου δηλαδή εποπτεύει τη διαδικασία της εκπαίδευσης και οδηγεί στη σταδιακή βελτίωση των παραμέτρων. Τέλος, οι προβλέψεις του μοντέλου ελέγχονται ως προς το «τυφλό» σύνολο επαλήθευσης, το οποίο δίνει και το σφάλμα πρόβλεψης υπό κανονικές συνθήκες. Στη Διατριβή το παραπάνω σχήμα χρησιμοποιήθηκε στο Κεφάλαιο 7.

Στην περίπτωση που υπάρχει έλλειψη σχετικά μεγάλων συνόλων δεδομένων ή που επιδιώκεται η εξασφάλιση της ομοιομορφίας κατά την επιλογή των δεδομένων εκπαίδευσης και ελέγχου, προτείνεται η εφαρμογή της μεθόδου της διασταυρούμενης επικύρωσης (εσωτερική αξιολόγηση, cross validation). Με τη μέθοδο αυτή, το αρχικό σύνολο δεδομένων χωρίζεται σε δύο ισοπληθή υποσύνολα με τυχαίο τρόπο. Στη συνέχεια, το ένα υποσύνολο χρησιμοποιείται ως σύνολο εκπαίδευσης και το άλλο ως σύνολο ελέγχου αποθηκεύοντας τις προβλέψεις για τα δείγματα ελέγχου, ενώ η διαδικασία επαναλαμβάνεται αντιστρέφοντας τους «ρόλους» των δύο υποσυνόλων και αποθηκεύοντας τις προβλέψεις για τα υπόλοιπα δείγματα. Η μέθοδος γενικεύεται με διάσπαση του αρχικού συνόλου σε k υποσύνολα (k-

fold cross-validation). Σε αυτή την περίπτωση οι διαδικασίες εκπαίδευσης και αξιολόγησης επαναλαμβάνονται k φορές. Σε κάθε επανάληψη, παράγονται διαδοχικά προβλέψεις για ένα από τα υποσύνολα χρησιμοποιώντας ως δεδομένα εκπαίδευσης τα υπόλοιπα k–1, λαμβάνοντας κατ' αυτό τον τρόπο σταδιακά τα συγκεντρωτικά αποτελέσματα με τις προβλέψεις για κάθε δείγμα από το αρχικό σύνολο δεδομένων. Στην περίπτωση όπου η παράμετρος k ισούται με το πλήθος των διαθέσιμων δεδομένων, κάθε σημείο ελέγχεται διαδοχικά σε ένα μοντέλο που έχει αναπτυχθεί χρησιμοποιώντας σχεδόν όλα τα διαθέσιμα δεδομένα (leave-one-out cross validation).

Τέλος προκειμένου να εξασφαλιστεί ότι δεν έχουν μοντελοποιηθεί τυχαίες συσχετίσεις μεταξύ των δεδομένων εισόδου και εξόδου, εφαρμόζεται ο έλεγχος της τυχαίας επιλογής (y-randomisation ή y-scrambling). Κατά τον έλεγχο αυτό, οι τιμές της εξαρτημένης μεταβλητής ανακατεύονται και μοιράζονται τυχαία ανάμεσα στα δείγματα και στη συνέχεια αναπτύσσεται ένα μοντέλο που συσχετίζει τις ανεξάρτητες μεταβλητές εισόδου με την τυχαία έξοδο. Η διαδικασία επαναλαμβάνεται αρκετές φορές. Αν τα παραγόμενα μοντέλα έχουν καλή απόδοση στην εξωτερική αξιολόγηση, συγκρίσιμη με την απόδοση του μοντέλου που αναπτύσσεται χρησιμοποιώντας το πρωτότυπο σύνολο δεδομένων, τότε το μοντέλο δεν θεωρείται αξιόπιστο τόσο λόγω των δεδομένων όσο και της μεθοδολογίας μοντελοποίησης. Επισημαίνεται ότι οι τεχνικές αξιολόγησης δεν εξαντλούνται σε αυτές που αναφέρονται εδώ.

Αφού παραχθούν προβλέψεις για τα δεδομένα ελέγχου, ανάλογα με το είδος της εξαρτημένης μεταβλητής (αριθμός/κλάση), υπολογίζονται και τα κατάλληλα μέτρα αξιολόγησης όπως προτείνεται από τον Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης (ΟΟΣΑ). Στην περίπτωση αριθμητικής εξόδου (§2.3.3.1) συνηθίζεται να υπολογίζονται το μέσο τετραγωνικό σφάλμα (mean-squared error), η ρίζα του μέσου τετραγωνικού σφάλματος (root mean-squared error), ο συντελεστής συσχέτισης των δεδομένων εκπαίδευσης (correlation coefficient, R^2) και ο δείκτης εξωτερικής ερμηνεύσιμης διακύμανσης (external explained variance, Q^2_{ext}). Στόχος κάθε προβλεπτικής μεθόδου είναι η ελάττωση των σφαλμάτων μεταξύ των προβλεπόμενων τιμών εξόδου ή ισοδύναμα η όσο το δυνατόν επιτυχής τους ταύτιση. Για το λόγο αυτό, οι «επιθυμητές» τιμές των σφαλμάτων που αναφέρθηκαν τείνουν στο μηδέν (0), ενώ οι «επιθυμητές» των R^2 και Q^2_{ext} τείνουν στη μονάδα (1).

Στην περίπτωση κατηγορικής εξόδου (§2.3.3.2), τα μέτρα αξιολόγησης προκύπτουν από συνδυασμούς των συχνοτήτων σωστής ή λανθασμένης κατανομής των δειγμάτων ανάμεσα στις κλάσεις (true positives, true negatives -δείγματα που έχουν κατανεμηθεί σωστά στις δύο κλάσεις, false positives και false negatives -δείγματα που έχουν κατανεμηθεί λανθασμένα στις κλάσεις positive και negative, ενώ ανήκουν στην αντίθετη κλάση). Οι συχνότητες αυτές συχνά εμφανίζονται στις μήτρες σύγχυσης (confusion matrices) και από αυτές υπολογίζεται μια πληθώρα μέτρων όπως η ακρίβεια (accuracy), η ευαισθησία (sensitivity), η εξειδίκευση (specificity), και ο συντελεστής συσχέτισης Matthews (Matthews correlation coefficient). Οι «επιθυμητές» τιμές των στατιστικών αυτών τείνουν προς τη μονάδα (1), δηλαδή επιδιώκεται όσο το δυνατόν η απόλυτη επιτυχία στην πρόβλεψη των κλάσεων.

Εργαλεία λογισμικού

Για την ανάλυση, την επεξεργασία των δεδομένων στα πλαίσια της παρούσας Διατριβής, αλλά και για την υλοποίηση των μεθοδολογιών read-across που σχεδιάστηκαν για την πρόβλεψη ιδιοτήτων υλικών, αναπτύχθηκε κώδικας σε γλώσσες προγραμματισμού R, Python και MATLAB, ενώ χρησιμοποιήθηκε και η πλατφόρμα KNIME.

Στα Παραρτήματα Α΄ και Β΄ περιγράφονται με λεπτομέρεια οι διάφορες γλώσσες προγραμματισμού και τα αντίστοιχα πακέτα που χρησιμοποιήθηκαν για την υλοποίηση των μεθοδολογιών και την ανάπτυξη εφαρμογών στα πλαίσια της Διατριβής.

Μελέτες περιπτώσεων

Οι μεθοδολογίες read-across που αναπτύχθηκαν στην παρούσα Διατριβή, εφαρμόστηκαν σε μια σειρά από δεδομένα που αντλήθηκαν από τη βιβλιογραφία και έχουν χρησιμοποιηθεί κατά κόρον σε μεθόδους και εφαρμογές νανοπληροφορικής, ώστε να διαπιστωθεί και να ποσοτικοποιηθεί η ικανότητά τους να παράγουν αξιόπιστες προβλέψεις.

Τα σύνολα δεδομένων αποτελούνται από ΝΥ χρυσού και αργύρου, ΝΥ με πυρήνα μεταλλικών οξειδίων και νανοσωλήνες άνθρακα πολλαπλών τοιχωμάτων. Τα σύνολα περιλαμβάνουν διάφορες ιδιότητες (περιγραφείς ή descriptors) όπως πειραματικά μετρούμενες ιδιότητες (π.χ. φυσικοχημικοί δείκτες, ανάλυση δεδομένων βιοπληροφορικής), και υπολογισμένοι περιγραφείς (π.χ. κβαντομηχανικοί και θεωρητικοί δείκτες, χαρακτηριστικά από ανάλυση εικόνων). Οι ιδιότητες αυτές αποτελούν τα δεδομένα εισόδου/ανεξάρτητες μεταβλητές στα μοντέλα που αναπτύχθηκαν προκειμένου να προβλεφθεί η εξαρτημένη μεταβλητή (endpoint). Η εξαρτημένη μεταβλητή μπορεί να είναι συνεχής (αριθμητική τιμή) ή κατηγορική (κλάση) και μπορεί να είναι η τοξικότητα των δειγμάτων ή οποιαδήποτε άλλη ιδιότητα ενδιαφέροντος.

Στον Πίνακα 3.1 παρουσιάζονται συνοπτικά οι σημαντικότερες πληροφορίες σχετικά με τα σύνολα δεδομένων: ο αριθμός των δειγμάτων, ο αριθμός και το είδος των περιγραφέων/ιδιοτήτων και το είδος της εξαρτημένης μεταβλητής προς πρόβλεψη. Είναι εμφανές ότι πρόκειται για σχετικά μικρά σύνολα δεδομένων και συνεπώς κατάλληλα για την εφαρμογή των μεθόδων read-across που χρησιμοποιούνται στις περιπτώσεις ελλείψεως δεδομένων.

Μεθοδολογία 1 - Ανάπτυξη και επίλυση μοντέλου μαθηματικής βελτιστοποίησης για την πρόβλεψη ιδιοτήτων υλικών

Όπως αναφέρθηκε, για την ανάπτυξη μιας αξιόπιστης μεθοδολογίας read-across το ενδιαφέρον μας εστιάστηκε αφενός στην επιλογή μεταβλητών που περιέχουν σημαντική πληροφορία για την πρόβλεψη της εκάστοτε απόκρισης και αφετέρου στη βέλτιστη επιλογή των ορίων που θα διαμορφώνουν τις «γειτονιές» συγγενών ΝΥ οδηγώντας σε πιο αξιόπιστες προβλέψεις. Και οι δύο αυτοί στόχοι, θα μπορούσαν να επιτευχθούν μέσω της ανάπτυξης ενός μοντέλου μεικτού-ακέραιου μη-γραμμικού μαθηματικού προγραμματισμού (Κεφάλαιο 4) όπου στόχος είναι να ελαχιστοποιηθεί το μέσο τετραγωνικό σφάλμα (mean squared error, MSE) μεταξύ των πραγματικών τιμών εξόδου (εν προκειμένω της τοξικότητας) και των τιμών που προκύπτουν από την προβλεπτική διαδικασία για κάθε ΝΥ εντός του συνόλου με τουλάχιστον ένα γείτονα. Η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος αποτέλεσε και τον κύριο όρο της αντικειμενικής συνάρτησης (ΑΣ). Ιδιαίτερη μνεία δόθηκε στον έλεγχο της επιλογής μεταβλητών, οπότε προστέθηκε στην ΑΣ και ένας όρος ομαλοποίησης (regularisation) ο οποίος ρυθμίζει και περιορίζει τον αριθμό των μεταβλητών για να εξασφαλιστεί η ανάπτυξη απλών μοντέλων και να αποφευχθεί η υπερπροσαρμογή τους στα δεδομένα εκπαίδευσης (Εξίσωση 4.12).

Το εν λόγω πρόβλημα μαθηματικού προγραμματισμού, μπορεί να επεκταθεί ενσωματώνοντας ένα ή περισσότερα κριτήρια ομοιότητας ανάλογα με τις διαφορετικές κατηγορίες ιδιοτήτων χαρακτηρισμού των ΝΥ που είναι διαθέσιμες (π.χ. βιολογικούς περιγραφείς, φυσικοχημικές ιδιότητες, μεταβλητές που προκύπτουν από προσομοιώσεις μοριακής δυναμικής κ.α.). Τα κριτήρια αυτά μπορούν να επηρεάσουν την επιλογή των γειτόνων θέτοντας περισσότερα κατώφλια για την επιλογή των γειτόνων και εισάγοντας περιορισμό για την ικανοποίηση όλων των κατωφλιών προκειμένου να επιλεγεί ένας γείτονας. Η επέκταση για ένα ή περισσότερα κριτήρια θεωρείται τετριμμένη. Το πλήρες πρόβλημα μαθηματικού προγραμματισμού παρουσιάζεται στις σελίδες 36-40.

Ωστόσο, το πρόβλημα μαθηματικής αριστοποίησης δεν δύναται να επιλυθεί

αποτελεσματικά μέσω συμβατικών μεθόδων αριστοποίησης, λόγω της ύπαρξης μη-γραμμικοτήτων. Έτσι, αναπτύχθηκε ένας καινοτόμος εξελικτικός αλγόριθμος βασιζόμενος στις αρχές των γενετικών αλγορίθμων που έχουν ήδη εφαρμοστεί σε διαδικασίες βέλτιστης επιλογής μεταβλητών.

Κατά τη διαδικασία αυτή κάθε πιθανή λύση (επιλεγμένες μεταβλητές και κατώφλια) απεικονίζεται ως ένα «χρωμόσωμα» που αποτελείται από τόσα «γονίδια» όσο και ο αριθμός των διαθέσιμων μεταβλητών, προσθέτοντας -σε καθορισμένες θέσεις- τον αριθμό των κατωφλιών ανάλογα με τον αριθμό των κριτηρίων ομοιότητας που εφαρμόζονται. Τα γονίδια που εκφράζουν την επιλογή ή μη μιας ιδιότητας είναι δυαδικές μεταβλητές, ενώ τα γονίδια που εκφράζουν την τιμή των κατωφλιών είναι συνεχείς.

Τα δεδομένα αρχικά κανονικοποιούνται προκειμένου όλες οι μεταβλητές να αποκτήσουν το ίδιο εύρος τιμών και να συμμετέχουν ισοδύναμα στην ανάλυση. Στη συνέχεια δημιουργείται ένας τυχαίος πληθυσμός από χρωμοσώματα και καθένα αξιολογείται ως προς την προβλεπτική του ικανότητα: υπολογίζονται οι ευκλείδειες αποστάσεις μεταξύ όλων των δειγμάτων του συνόλου δεδομένων, με βάση τις επιλεγμένες μεταβλητές, και στη συνέχεια επιλέγονται για κάθε δείγμα οι γείτονες μεταξύ των δειγμάτων: εάν η ευκλείδεια απόσταση μεταξύ δύο ΝΥ είναι μικρότερη από την τιμή του κατωφλιού, τότε τα δύο ΝΥ θεωρούνται γείτονες.

Για κάθε δείγμα αναφοράς, υπολογίζεται η πρόβλεψη read-across ως ο σταθμισμένος μέσος όρος των τοξικών αποκρίσεων όλων των γειτόνων του και για όλες τις προβλέψεις υπολογίζεται το μέσο τετραγωνικό σφάλμα. Στο τέλος κάθε χρωμόσωμα βαθμολογείται με την τιμή του αντιστρόφου μέσου τετραγωνικού σφάλματος.

Στις επόμενες επαναλήψεις, μέχρι να ολοκληρωθεί ένα καθορισμένο πλήθος «γενεών», επιλέγονται ζεύγη χρωμοσωμάτων, με βάση τη βαθμολογία τους, και εφαρμόζονται σε αυτά οι γενετικοί τελεστές της αναπαραγωγής και της μετάλλαξης: τα χρωμοσώματα αρχικά διασταυρώνονται σε τυχαίες θέσεις και στη συνέχεια με βάση μια προκαθορισμένη τιμή πιθανότητας, οι τιμές των γονιδίων που αντιστοιχούν σε μεταβλητές μεταβάλλονται από 0 σε 1 και αντίστροφα, ενώ οι τιμές των γονιδίων που αντιστοιχούν σε κατώφλια μεταβάλλονται σύμφωνα με την Εξίσωση 4.29. Ο νέος πληθυσμός χρωμοσωμάτων αξιολογείται ξανά, και η παραπάνω διαδικασία επαναλαμβάνεται. Το χρωμόσωμα με την υψηλότερη βαθμολογία κατά την εξελικτική διαδικασία, δίδεται αυτόματα ως έξοδος του αλγορίθμου.

Ο παραπάνω εξελικτικός αλγόριθμος χρησιμοποιήθηκε και για την ανάπτυξη μοντέλων κατηγοριοποίησης με ανάλογο τρόπο. Οι διαφορές εντοπίζονται στον υπολογισμό της πρόβλεψης (σταθμισμένη ψήφος των γειτόνων ανά κλάση, Εξίσωση 4.35) και στον τρόπο υπολογισμού της επίδοσης των πιθανών λύσεων μέσω της χρήσης του συντελεστή Matthews (ΑΣ, Εξίσωση 4.37). Η αλληλουχία βημάτων για την αξιολόγηση των χρωμοσωμάτων κάθε πληθυσμού σύμφωνα με το είδος της εξαρτημένης μεταβλητής, παρουσιάζεται στην Εικόνα 4.33.

Προκειμένου να αξιολογηθεί, η προτεινόμενη μεθοδολογία εφαρμόστηκε σε πέντε μελέτες περιπτώσεων και ελέγχθηκαν τα αποτελέσματα με και χωρίς τη χρήση παράγοντα ομαλοποίησης αλλά και με τη χρήση περισσοτέρων του ενός κριτηρίων ομαδοποίησης (περισσότερα κατώφλια ομοιότητας). Η μεθοδολογία εφαρμόστηκε σε ένα σύνολο 84 ΝΥ χρυσού, αρχικά λαμβάνοντας υπόψιν όλες τις διαθέσιμες μεταβλητές και ένα κατώφλι ομοιότητας και στη συνέχεια θεωρώντας δύο είδη μεταβλητών και δύο κατώφλια ομοιότητας (ένα φυσικοχημικό και ένα βιολογικό). Τα μοντέλα αξιολογήθηκαν για την ποιότητα των προβλέψεών τους στο σύνολο επαλήθευσης (test set) και η αξιοπιστία τους έφτασε στο $Q^2_{\text{ext}} = 0.78$ με χρήση ενός κατωφλιού και στο $Q^2_{\text{ext}} = 0.83$ με χρήση δύο κατωφλιών ομοιότητας. Η μεθοδολογία εφαρμόστηκε και σε ένα σύνολο 28 νανοσωλήνων άνθρακα πολλαπλών τοιχωμάτων με χρήση ενός κατωφλιού ομοιότητας και η αξιοπιστία των μοντέλων έφτασε στο $Q^2_{\text{ext}} = 0.81$.

Όσον αφορά την ανάπτυξη μοντέλων κατηγοριοποίησης, η μεθοδολογία εφαρμόστηκε

στα δεδομένα 25 ΝΥ με πυρήνα μεταλλικών οξειδίων και η ακρίβεια της πρόβλεψης έφτασε στο 100% χρησιμοποιώντας είτε ένα είτε δύο κατώφλια ομοιότητας. Η εφαρμογή της μεθοδολογίας σε δύο ακόμη σύνολα δεδομένων απέδωσε ικανοποιητικά αποτελέσματα στην εξωτερική αξιολόγηση με ακρίβεια προβλέψεων στο σύνολο επαλήθευσης ίση ή μεγαλύτερη του 80%.

Η διάθεση της παραπάνω αυτοματοποιημένης μεθοδολογίας ανάπτυξης μοντέλων read-across πραγματοποιήθηκε με την ανάπτυξη μιας εφαρμογής με το όνομα Apellis που υλοποιεί την μεθοδολογία και διευκολύνει το ευρύ επιστημονικό κοινό να την χρησιμοποιήσει μέσα από ένα εύχρηστο περιβάλλον. Το γραφικό περιβάλλον από μενού και κουμπιά επιτρέπει την πρόσβαση στη μεθοδολογία ακόμα και από άτομα που δε διαθέτουν βαθιές υπολογιστικές γνώσεις. Οι ενδιαφερόμενοι χρήστες δύνανται να χρησιμοποιήσουν τα δικά τους δεδομένα (που δεν περιορίζονται μόνο σε δεδομένα νανοτοξικότητας) ώστε να αναπτύξουν είτε μοντέλα παλινδρόμησης είτε μοντέλα κατηγοριοποίησης με τη χρήση ενός ή δύο κατωφλιών ομοιότητας. Η εφαρμογή είναι ελεύθερα διαθέσιμη από τον ακόλουθο σύνδεσμο: `https://apellis.jaqpot.org/`, ενώ διατίθεται πλούσιο εκπαιδευτικό υλικό για τη χρήση της. Η ανάπτυξη της εφαρμογής μέσω του Docker αλλά και η λειτουργία της παρουσιάζονται ενδελεχώς στις σελίδες 84-95.

Για την υλοποίηση της μεθοδολογίας αυτής, αναπτύχθηκε κώδικας σε γλώσσα προγραμματισμού MATLAB ενώ η εφαρμογή Apellis αναπτύχθηκε σε γλώσσα R με χρήση του πακέτου `shiny`.

Μεθοδολογία 2 - Ανάπτυξη μεθοδολογίας ομαδοποίησης βάσει βελτιστοποιημένου αλγορίθμου τμηματικής γραμμικής παλινδρόμησης

Η μεθοδολογία που περιγράφηκε στην προηγούμενη παράγραφο, λόγω του στοχαστικού της χαρακτήρα, παράγει λύσεις κοντά στη βέλτιστη. Επίσης ο χρόνος εκπαίδευσης -ειδικά στην περίπτωση μεγάλων συνόλων δεδομένων- είναι αρκετά μεγάλος. Προκειμένου να βελτιωθούν τα αναφερόμενα χαρακτηριστικά της προηγούμενης μεθοδολογίας, αναπτύχθηκε μια ακόμα αυτοματοποιημένη μέθοδος ομαδοποίησης και πρόβλεψης στα πλαίσια του read-across. Η μεθοδολογία αυτή (Κεφάλαιο 5) βασίζεται στην ανάπτυξη και επίλυση ενός προβλήματος μεικτού-ακέραιου γραμμικού προγραμματισμού το οποίο πραγματοποιεί επιλογή μεταβλητών, δημιουργία ομάδων ΝΥ βάσει μιας ή περισσοτέρων γνωστών ιδιοτήτων τους (ανεξάρτητες μεταβλητές) και ανάπτυξη τοπικών γραμμικών μοντέλων πρόβλεψης της τοξικότητας των ΝΥ ανά ομάδα (τμηματική γραμμική παλινδρόμηση).

Ο στόχος αυτής της μεθοδολογίας είναι και πάλι η ελαχιστοποίηση των διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών της ιδιότητας εξόδου και επιτυγχάνεται με την ελαχιστοποίηση της τιμής του μέσου τετραγωνικού σφάλματος (mean absolute error, MAE). Όπως και στην προηγούμενη μεθοδολογία εισάχθηκε στην ΑΣ ένας όρος ομαλοποίησης, ο οποίος ρυθμίζεται από έναν παράγοντα ομαλοποίησης λ. Η ελαχιστοποίηση της ΑΣ του προβλήματος εξασφαλίζει δηλαδή τόσο την ακριβή πρόβλεψη της εξόδου όσο και τη χρήση μόνο των απαραίτητων ιδιοτήτων για την πρόβλεψη της εξόδου, κάνοντας μια έμμεση επιλογή μεταβλητών. Στη συνέχεια, οι περιορισμοί του προβλήματος εξασφαλίζουν ότι τα σημεία διαμέρισης -με βάση την μεταβλητή διαμέρισης- είναι διαδοχικά, ότι κάθε δείγμα ανήκει αποκλειστικά και μόνο σε μια περιοχή και τοποθετείται σε αυτή με βάση την τιμή της ιδιότητας διαμέρισης, ενώ η πρόβλεψη γίνεται εφαρμόζοντας ένα γραμμικό μοντέλο ανά περιοχή. Σε περίπτωση που είναι διαθέσιμα δεδομένα για δύο ή περισσότερες κατηγορίες ιδιοτήτων των ΝΥ, είναι δυνατόν να χρησιμοποιηθούν περισσότερες ιδιότητες (μία από κάθε κατηγορία) που διαχωρίζουν το πεδίο ορισμού σε περιοχές, σε καθεμιά από τις οποίες -κατ' αντιστοιχία- εφαρμόζονται μοντέλα γραμμικής παλινδρόμησης. Το

πρόβλημα μαθηματικού προγραμματισμού που αναπτύχθηκε παρουσιάζεται αναλυτικά στις σελίδες 100-106.

Το παραπάνω πρόβλημα αριστοποίησης εφαρμόζεται στο πλαίσιο μιας συνολικής μεθοδολογίας, η οποία, ως πρώτο βήμα, αναζητά, από τις διαθέσιμες ιδιότητες εκείνη που μπορεί να χωρίσει το πεδίο ορισμού (τα διαθέσιμα δείγματα) σε περιοχές στον πολυδιάστατο χώρο. Αυτό επιτυγχάνεται επιλύοντας το πρόβλημα αριστοποίησης χρησιμοποιώντας κάθε φορά μια από τις διαθέσιμες μεταβλητές ως μεταβλητή διαμέρισης και δημιουργώντας δύο περιοχές. Στη συνέχεια σε κάθε περιοχή εφαρμόζεται ένα γραμμικό μοντέλο που προβλέπει την έξοδο και καταγράφονται τα σφάλματα πρόβλεψης. Από τις διαθέσιμες μεταβλητές εισόδου, αυτή που οδηγεί στα μικρότερα σφάλματα επιλέγεται και ως μεταβλητή διαμέρισης. Στη συνέχεια, εξετάζεται η προσθήκη περισσότερων περιοχών, επιλύοντας και πάλι το πρόβλημα μαθηματικής αριστοποίησης έως ότου να μην υπάρχει ικανοποιητική βελτίωση των σφαλμάτων μεταξύ δύο διαδοχικών προσθηκών επιπλέον περιοχής. Μετά το πέρας της διαδικασίας προσθήκης επιπλέον περιοχών, προκύπτει αυτόματα και η υπόθεση ομαδοποίησης που αποτελείται από τη μεταβλητή διαμέρισης, τα σημεία (συντεταγμένες) διαμέρισης, το πλήθος των περιοχών και τους συντελεστές των γραμμικών μοντέλων σε κάθε περιοχή. Λόγω του αιτιοκρατικού χαρακτήρα της μεθοδολογίας, η λύση του προβλήματος μαθηματικού προγραμματισμού είναι η βέλτιστη (και όχι μια λύση κοντά στη βέλτιστη) οπότε και η προκύπτουσα υπόθεση συνιστά και τη βέλτιστη υπόθεση ομαδοποίησης.

Η μεθοδολογία εφαρμόστηκε με επιτυχία σε δύο σύνολα δεδομένων. Στην πρώτη περίπτωση χρησιμοποιήθηκε όπως και προηγουμένως το σύνολο των 84 ΝΥ χρυσού, αρχικά επιλέγοντας μία μόνο μεταβλητή διαμέρισης και στη συνέχεια επιλέγοντας δύο μεταβλητές μια φυσικοχημική ιδιότητα και μια βιολογική. Τα μοντέλα αξιολογήθηκαν για την ποιότητα των προβλέψεών τους στο σύνολο επαλήθευσης (test set) και η αξιοπιστία τους -σε όρους εξωτερικής ερμηνεύσιμης διακύμανσης- ήταν ίση με $Q^2_{\text{ext}} = 0.88$ με χρήση μίας μεταβλητής και ίση με $Q^2_{\text{ext}} = 0.86$ με χρήση δύο μεταβλητών. Η μεθοδολογία εφαρμόστηκε και στο σύνολο 28 νανοσωλήνων άνθρακα πολλαπλών τοιχωμάτων με χρήση μιας μεταβλητής διαμέρισης και η αξιοπιστία του μοντέλου ήταν ίση με $Q^2_{\text{ext}} = 0.86$.

Τα μοντέλα που παράχθηκαν από τη μεθοδολογία αυτή, βρίσκονται διαθέσιμα μέσω της διαδικτυακής εφαρμογής vythos (`https://vythos.jaqpot.org/`), η οποία δημιουργήθηκε με στόχο την «φιλοξενία» των μοντέλων που προκύπτουν από την προαναφερθείσα μεθοδολογία. Οι ενδιαφερόμενοι χρήστες, μέσω ενός φιλικού περιβάλλοντος και εφαρμόζοντας μια σειρά απλών βημάτων, μπορούν εντός δευτερολέπτων να λάβουν προβλέψεις για άγνωστα δείγματα και πληροφορίες για την αξιοπιστία των προβλέψεων αυτών. Επίσης, παρέχεται η πληροφορία σχετικά με την ομάδα/περιοχή στην οποία ανήκουν τα άγνωστα δείγματα και η θέση τους ως προς τη θέση των δειγμάτων του συνόλου εκπαίδευσης, όπως αυτή ορίζεται από τις τιμές των μεταβλητών διαμέρισης και της ιδιότητας εξόδου.

Για την υλοποίηση της μεθοδολογίας αυτής, αναπτύχθηκε κώδικας σε γλώσσα προγραμματισμού MATLAB ενώ η εφαρμογή vythos αναπτύχθηκε σε γλώσσα R με χρήση του πακέτου `shiny`.

Μεθοδολογία 3 - Ανάπτυξη μεθοδολογίας ομαδοποίησης με βάση τη βέλτιστη διαίρεση της μεταβλητής απόκρισης

Προκειμένου να εντοπίζονται μοτίβα στην ιδιότητα ενδιαφέροντος/εξαρτημένη μεταβλητή (π.χ. τοξικότητα) η προηγούμενη μεθοδολογία επεκτάθηκε με την ανάπτυξη ενός προβλήματος μεικτού-ακέραιου γραμμικού προγραμματισμού που δημιουργεί ομάδες παρόμοιων ΝΥ, χωρίζοντας το πεδίο ορισμού με βάση την ιδιότητα εξόδου (Κεφάλαιο 6).

Ο στόχος και αυτής της μεθοδολογίας είναι η ελαχιστοποίηση των διαφορών μεταξύ

των προβλεπόμενων και των πραγματικών τιμών της ιδιότητας εξόδου και επιτυγχάνεται με την ελαχιστοποίηση της τιμής του μέσου τετραγωνικού σφάλματος (MAE). Όπως και στις προηγούμενες μεθοδολογίες εισάχθηκε στην ΑΣ ένας όρος ομαλοποίησης. Ωστόσο, σε αυτή την περίπτωση το πεδίο των δεδομένων δεν χωρίζεται με βάση μία ή περισσότερες μεταβλητές εισόδου αλλά με βάση τη μεταβλητή απόκρισης (έξοδος). Οι περιορισμοί του προβλήματος εξασφαλίζουν ότι τα σημεία διαμέρισης -με βάση την μεταβλητή εξόδου- θα είναι διαδοχικά, ότι κάθε περιοχή θα περιέχει τουλάχιστον ένα δείγμα, ότι κάθε δείγμα θα ανήκει αποκλειστικά και μόνο σε μία περιοχή και θα τοποθετείται σε αυτή με βάση την τιμή της ιδιότητας εξόδου, ενώ η πρόβλεψη θα γίνεται εφαρμόζοντας ένα γραμμικό μοντέλο ανά περιοχή.

Στη συνέχεια, για την κατανομή άγνωστων δειγμάτων στις διάφορες περιοχές, για καθεμιά από αυτές ορίζεται το «χαρακτηριστικό» της κέντρο, με βάση τις επιλεγμένες μεταβλητές και τα δείγματα που ανήκουν σε αυτή. Για την καταχώριση των αγνώστων δειγμάτων υπολογίζεται η Ευκλείδεια απόστασή τους από όλα τα χαρακτηριστικά κέντρα και το δείγμα τοποθετείται στην περιοχή από την οποία το δείγμα έχει την ελάχιστη απόσταση. Η μεθοδολογία ομαδοποίησης που αναπτύχθηκε παρουσιάζεται αναλυτικά στις σελίδες 125-129.

Η μεθοδολογία grouping/read-across με βάση τη μεταβλητή απόκρισης αποτελεί μέρος μιας ευρύτερης ροής βημάτων (μεθοδολογία demos), η οποία καταλήγει στο βέλτιστο μοντέλο μεταξύ ενός μοντέλου πολλαπλής γραμμικής παλινδρόμησης, ενός μοντέλου τύπου LASSO και της μεθοδολογίας grouping/read-across (Εικόνα 6.1). Αρχικά εφαρμόζονται στα δεδομένα οι μεθοδολογίες γραμμικής παλινδρόμησης και LASSO και καταγράφονται κάποια αρχικά σφάλματα. Στη συνέχεια για ένα εύρος τιμών του παράγοντα ομαλοποίησης λ, εφαρμόζεται η μεθοδολογία grouping. Σε κάθε επανάληψη επιλύεται το πρόβλημα μαθηματικού προγραμματισμού για δύο περιοχές και διαδοχικά ελέγχεται η περαιτέρω αύξηση του αριθμού των περιοχών. Αφού επιλυθεί και η μεθοδολογία grouping/read-across για διάφορες τιμές του λ, επιλέγεται το βέλτιστο μοντέλο, σε όρους ελάχιστου σφάλματος κατά την εξωτερική αξιολόγηση, μεταξύ των τριών μεθόδων που εφαρμόζονται.

Η μεθοδολογία εφαρμόστηκε στο σύνολο των 84 ΝΫ χρυσού το οποίο χωρίστηκε σε δύο περιοχές και η αξιοπιστία των προβλέψεων -σε όρους εξωτερικής ερμηνεύσιμης διακύμανσης- ήταν ίση με $Q^2_{\text{ext}} = 0.83$. Χάρη στη διαίρεση του χώρου των δεδομένων με βάση την ιδιότητα εξόδου, υπήρξε η δυνατότητα να παρατηρηθούν μοτίβα στις σχηματιζόμενες περιοχές. Όπως διαπιστώθηκε στη δεύτερη περιοχή τα δείγματα που συγκεντρώθηκαν είχαν κατιονική επίστρωση στην επιφάνειά τους, γεγονός που βρίσκεται σε συμφωνία με τη βιβλιογραφία, καθώς τα κατιονικά ΝΫ χρυσού έχουν πιο τοξικό χαρακτήρα από τα ανιονικά.

Για την υλοποίηση της μεθοδολογίας αυτής, αναπτύχθηκε κώδικας σε γλώσσα προγραμματισμού Python και αξιοποιήθηκαν τα εργαλεία μαθηματικής αριστοποίησης (πακέτο `mip`) και παράλληλης εκτέλεσης (πακέτο `multiprocessing`) που παρέχονται, εκτός από τα διαθέσιμα πακέτα ανάλυσης δεδομένων (πακέτα `numpy`, `pandas`, `scikit-learn`).

Μεθοδολογία 4 - Ανάπτυξη μοντέλων read-across βάσει της μεθοδολογίας των k-πλησιέστερων γειτόνων

Στην τελευταία ενότητα της Διατριβής (Κεφάλαιο 7) παρουσιάζεται η χρήση της μεθοδολογίας μηχανικής μάθησης των k-πλησιέστερων γειτόνων (kNN), ως μια εναλλακτική μεθοδολογία τύπου read-across. Δεδομένου ότι για την πρόβλεψη της απόκρισης ενός δείγματος χρησιμοποιούνται δεδομένα «συγγενών» δειγμάτων (υπό την έννοια της ομοιότητας των μεταξύ τους ιδιοτήτων) και η πρόβλεψη περιορίζεται σε ένα μικρό μέρος του χώρου των δειγμάτων, ο αλγόριθμος αυτός μπορεί να χρησιμοποιηθεί για τη δημιουργία ομάδων παρόμοιων δειγμάτων υλικών. Περισσότερες λεπτομέρειες για τη μεθοδολογία kNN μπορούν να βρεθούν στο Παράρτημα Ϛ΄.

Η μεθοδολογία *k*NN εφαρμόστηκε σε δύο σύνολα δεδομένων. Στην πρώτη εφαρμογή, πραγματοποιήθηκε μοντελοποίηση των τοξικών και βιολογικών επιδράσεων επικαλυμμένων νανοσωλήνων άνθρακα (decorated multi-walled carbon nanotubes, MWCNTs). Συγκεκριμένα, αναπτύχθηκαν μέσω της πλατφόρμας KNIME δύο μοντέλα κατηγοριοποίησης της μορφής *k*NN/read-across, ένα για την πρόβλεψη της κυτοτοξικότητας («τοξικά»/«μη τοξικά» δείγματα) και ένα για την ιδιότητα της πρωτεϊνικής πρόσδεσης («πρωτεϊνικοί προσδέτες»/«μη πρωτεϊνικοί προσδέτες»). Η πρωτεϊνική πρόσδεση (protein binding) συσχετίζεται άμεσα με την τοξικότητα καθώς υψηλές τιμές συνδέονται με μια αυξημένη τάση ενός νανοσωλήνα άνθρακα να καθιστά τα κύτταρα πιο ευαίσθητα στη φαγοκυττάρωση.

Για την ανάπτυξη των μοντέλων υπολογίστηκαν, μέσω του λογισμικού Mold2, 777 θεωρητικοί περιγραφείς-ιδιότητες που κωδικοποιούν γεωμετρικά και τοπολογικά χαρακτηριστικά των επιφανειακών μορίων επικάλυψης των νανοσωλήνων. Δεδομένου ότι οι νανοσωλήνες έχουν ακριβώς τις ίδιες διαστάσεις και διαφέρουν μόνο ως προς το είδος των μορίων της επιφάνειάς τους, έγινε η παραδοχή ότι οι διαφορές στην τοξική και στη βιολογική τους συμπεριφορά εξαρτώνται μόνο από την επιφανειακή τους επικάλυψη. Τα δεδομένα κανονικοποιήθηκαν με χρήση της γκαουσιάνης κανονικοποίησης (Εξίσωση 2.2) και ο όγκος τους μειώθηκε απομακρύνοντας τις ιδιότητες που περιείχαν τιμές με μικρή διακύμανση.

Για κάθε μεταβλητή απόκρισης (τοξικότητα και πρωτεϊνική πρόσδεση) αναπτύχθηκε ένα μοντέλο χρησιμοποιώντας το σχήμα της Εικόνας 7.1. Το αρχικό σύνολο δεδομένων χωρίστηκε τυχαία σε σύνολο εκπαίδευσης και σύνολο επαλήθευσης με αναλογία 75:25. Στη συνέχεια, το σύνολο εκπαίδευσης χωρίστηκε τυχαία σε δύο επιμέρους υποσύνολα βαθμονόμησης και ελέγχου, με αναλογία 50:25 του αρχικού συνόλου. Το σύνολο βαθμονόμησης χρησιμοποιήθηκε για την επιλογή μεταβλητών και την εύρεση του βέλτιστου αριθμού γειτόνων, *k*. Για την επιλογή μεταβλητών χρησιμοποιήθηκε η μεθοδολογία `InfoGain` σε συνδυασμό με τον αξιολογητή `Ranker`. Για την εύρεση του βέλτιστου αριθμού γειτόνων αξιολογήθηκε η ικανότητα παραγωγής αξιόπιστων προβλέψεων στα δεδομένα του συνόλου ελέγχου, υπολογίζοντας τα στατιστικά της αξιοπιστίας, ευαισθησίας, ειδικότητας, αλλά και πραγματοποιώντας έλεγχο τυχαίας επιλογής. Τέλος η ακρίβεια των μοντέλων αξιολογήθηκε στο «τυφλό» σύνολο επικύρωσης, το οποίο δε συμμετείχε στη διαδικασία της εκπαίδευσης και παρομοιάζει τη χρήση του μοντέλου υπό πραγματικές συνθήκες.

Για το μοντέλο της τοξικότητας επιλέχθηκαν 6 περιγραφείς (Πίνακας 7.1) και ο βέλτιστος αριθμός γειτόνων βρέθηκε ίσος με 7. Η ακρίβεια πρόβλεψης για τα σύνολα ελέγχου και επικύρωσης υπολογίστηκε ίση με 0.78 και 0.84 αντίστοιχα (Πίνακας 7.2). Ομοίως για το μοντέλο της πρωτεϊνικής πρόσδεσης επιλέχθηκαν 6 περιγραφείς και ο βέλτιστος αριθμός γειτόνων βρέθηκε ίσος με 3. Η ακρίβεια πρόβλεψης για τα σύνολα ελέγχου και επικύρωσης υπολογίστηκε ίση με 0.75 και 0.86 αντίστοιχα. Τέλος ορίστηκε και το πεδίο εφαρμογής των μοντέλων μέσω της μεθοδολογίας που περιγράφεται στην Παράγραφο 2.4.

Τα δύο μοντέλα αποτέλεσαν μέρος μιας διαδικτυακής εφαρμογής για την πρόβλεψη ανεπιθύμητων ιδιοτήτων των νανοϋλικών η οποία είναι διαθέσιμη από τον σύνδεσμο: `http://enaloscloud.novamechanics.com/EnalosWebApps/CNT/`. Η εφαρμογή είναι εύκολη στη χρήση ακόμα και από χρήστες χωρίς υπολογιστική εμπειρία, καθώς μέσα από το γραφικό περιβάλλον μπορούν να εισάγουν τα δεδομένα τους και να λάβουν τις αντίστοιχες προβλέψεις με το πάτημα μερικών κουμπιών.

Στη δεύτερη εφαρμογή μοντελοποίησης με χρήση της μεθοδολογίας *k*NN, αναπτύχθηκε ένα μοντέλο πρόβλεψης του δυναμικού-ζ (zeta-potential index) ενός συνόλου δεδομένων νανοσωματιδίων με γνωστές γεωμετρικές ιδιότητες. Συγκεκριμένα, έγινε ανάλυση μιας σειράς 68 εικόνων μικροσκοπίας TEM νανοσωματιδίων μέσω της διαδικτυακής εφαρμογής NanoXtract και εξήχθησαν 18 χρήσιμοι περιγραφείς των γεωμετρικών τους χαρακτηριστικών. Οι δείκτες αυτοί στη συνέχεια χρησιμοποιήθηκαν ως μεταβλητές εισόδου για την ανάπτυξη ενός υπολογιστικού μοντέλου τύπου *k*NN/read-across για την πρόβλεψη της ιδιότητας

του δυναμικού-ζ. Στις ανεξάρτητες μεταβλητές προστέθηκαν το pH του διαλύματος που έγινε η μέτρηση του δυναμικού-ζ και το είδος του πυρήνα των νανοσωματιδίων (καθαρό μέταλλο/μεταλλικό οξείδιο). Οι αριθμητικές τιμές των εξαρτημένων μεταβλητών και της μεταβλητής απόκρισης κανονικοποιήθηκαν με βάση την Εξίσωση 2.2.

Το σύνολο δεδομένων χωρίστηκε τυχαία σε σύνολο εκπαίδευσης και σύνολο επαλήθευσης με αναλογία 75:25 και στο σύνολο εκπαίδευσης πραγματοποιήθηκε επιλογή μεταβλητών με τη μεθοδολογία `BestFirst` σε συνδυασμό με τον αξιολογητή `CfsSubsetEval`. Οι μεταβλητές που χρησιμοποιήθηκαν για τη μοντελοποίηση είναι το είδος του πυρήνα των νανοσωματιδίων και ο γεωμετρικός δείκτης της κύριας επιμήκυνσης (main elongation). Το pH του διαλύματος (6.5 ή 7) είναι επίσης απαραίτητο για την πρόβλεψη. Επιλέχθηκε ο βέλτιστος αριθμός γειτόνων ίσος με 7 για την πρόβλεψη του δυναμικού-ζ.

Το μοντέλο αξιολογήθηκε για την ποιότητα των προβλέψεών του στο σύνολο επαλήθευσης με $Q^2_{\text{ext}} = 0.91$, ενώ πέρασε και από άλλα τεστ αξιολόγησης που προτείνονται στη βιβλιογραφία όπως ο έλεγχος τυχαίας επιλογής. Ορίστηκε επίσης και το πεδίο εφαρμογής του μοντέλου μέσω της μεθοδολογίας που περιγράφεται στην Παράγραφο 2.4.

Η μοντελοποίηση πραγματοποιήθηκε με το ελεύθερο λογισμικό KNIME. Βαρύτητα δόθηκε στη μελέτη της φυσικής διάστασης του προβλήματος, μέσω προσεκτικής μελέτης της βιβλιογραφίας, ώστε να ερμηνευτεί η επιρροή των γεωμετρικών χαρακτηριστικών των νανοσωματιδίων στον δείκτη του δυναμικού-ζ.

Το μοντέλο που αναπτύχθηκε αποτέλεσε τη βάση για την ανάπτυξη μιας διαδικτυακής εφαρμογής, η οποία είναι διαθέσιμη στο σύνδεσμο `http://enaloscloud.novamechanics.com/EnalosWebApps/ZetaPotential/`. Οι ενδιαφερόμενοι χρήστες με μια σειρά απλών βημάτων δύνανται να εισάγουν τα δεδομένα τους και να λάβουν και τις αντίστοιχες προβλέψεις.

Επίλογος-Συμπεράσματα

Η νανοπληροφορική είναι ένα ανερχόμενο πεδίο το οποίο –λαμβάνοντας υπόψιν το πλήθος των εφαρμογών της νανοτεχνολογίας, την ανάγκη για άμεση διερεύνηση των ανεπιθύμητων επιδράσεων των ΝΥ στους ζωντανούς οργανισμούς και την ανάγκη για μείωση των πειραμάτων σε πειραματόζωα- λαμβάνει υποστήριξη από την ερευνητική κοινότητα και τους φορείς όπως η Ευρωπαϊκή Ένωση. Η συμβολή της παρούσας Διατριβής στο πεδίο αυτό είναι σημαντική για τη γρήγορη και αξιόπιστη πρόβλεψη τοξικών και άλλων ιδιοτήτων ΝΥ.

Όλες οι μεθοδολογίες που αναπτύχθηκαν, εφαρμόστηκαν σε βιβλιογραφικά δεδομένα και ελέγχθηκαν βάσει των οδηγιών του ΟΟΣΑ. Ο έλεγχος απέδειξε ότι πρόκειται για καινοτόμες και αξιόπιστες μεθοδολογίες που βελτιώνουν την ικανότητα πρόβλεψης των ήδη υπαρχόντων μοντέλων και οι οποίες θα συμβάλουν καταλυτικά στην έρευνα της τοξικότητας ΝΥ. Στους Πίνακες 8.1 και 8.2 συνοψίζονται τα κυριότερα χαρακτηριστικά των τεσσάρων μεθοδολογιών που αναπτύχθηκαν και ο χρόνος εκπαίδευσής τους στο σύνολο των 84 ΝΥ χρυσού, αντίστοιχα. Με αυτές τις μεθοδολογίες θα επιταχυνθεί η αξιολόγηση των πιθανών κινδύνων των ΝΥ που ήδη υπάρχουν στην αγορά ή βρίσκονται σε φάση ανάπτυξης, ενώ θα ελαχιστοποιηθούν οι απαιτούμενοι πόροι (κόστος και εργασία) για την πειραματική τους αξιολόγηση.

Τα αποτελέσματα της έρευνας διατίθενται ελεύθερα, είτε ως πηγαίος κώδικας στο αποθετήριο GitHub, είτε μέσω διαδικτυακών εφαρμογών, σε όλη την επιστημονική κοινότητα ώστε να επωφεληθεί από την υπολογιστική μελέτη της τοξικότητας των ΝΥ. Με αυτό τον τρόπο μπορούν να χρησιμοποιούνται άμεσα και δωρεάν. Μάλιστα, μέσω ενός γραφικού περιβάλλοντος φιλικού-προς-το χρήστη, τα μοντέλα είναι εύκολα προσβάσιμα και σε ερευνητές που δεν έχουν εξειδικευμένες υπολογιστικές γνώσεις (π.χ. πειραματιστές), ώστε να μπορούν εύκολα να εφαρμόζουν τις τεχνικές απευθείας στα πειραματικά τους δεδομένα. Με τον τρόπο αυτό μεγιστοποιείται η συμβολή της νανοπληροφορικής στην επιστημονική

έρευνα στο πεδίο της τοξικότητας.

Αναλυτικότερα, οι τεχνικές που αναπτύχθηκαν και παρουσιάζονται σε αυτή τη Διατριβή, μπορούν να εφαρμοστούν στις ακόλουθες περιπτώσεις, διευκολύνοντας την επιστημονική έρευνα στο πεδίο της νανοτοξικότητας και συμβάλλοντας στη μείωση των πειραμάτων σε πειραματόζωα:

- Κατά τις διαδικασίες εκτίμησης του κινδύνου χρήσης των ΝΥ που ήδη βρίσκονται στο εμπόριο,
- Για την ιεράρχηση κατά την πειραματική αξιολόγηση των ΝΥ, αποκλείοντας ΝΥ που ήδη έχουν προβλεφθεί ως τοξικά, εξοικονομώντας χρόνο και κόστος,
- Για την ανάπτυξη νέων ασφαλέστερων και αποδοτικότερων ΝΥ από τη φάση του σχεδιασμού και πριν την μαζική παραγωγή τους (safety-by-design),
- Για τον εντοπισμό των ιδιοτήτων των ΝΥ οι οποίες πρέπει να ρυθμιστούν ώστε να παράγονται ασφαλή ΝΥ,
- Για τον εντοπισμό των ιδιοτήτων των ΝΥ οι οποίες πρέπει να ρυθμιστούν ώστε να παράγονται ΝΥ με βελτιωμένα χαρακτηριστικά,
- Για την κάλυψη των «κενών» στο χώρο των ιδιοτήτων των ΝΥ και,
- Για διευκόλυνση των Ρυθμιστικών Αρχών ώστε να εντοπίζονται τα όρια ομαδοποίησης ΝΥ και να εντάσσονται στους κανονισμούς.

Τέλος, θα πρέπει να τονιστεί ότι οι μεθοδολογίες αυτές έχουν καθολικό χαρακτήρα και θα μπορούσαν να εφαρμοστούν και σε άλλα προβλήματα πρόβλεψης ιδιοτήτων υλικών και γενικότερα της Επιστήμης των Δεδομένων.

Στο Παράρτημα της παρούσας Διατριβής, περιλαμβάνεται στα ελληνικά ένα μικρό ερμηνευτικό «λεξικό» των κυριότερων όρων που απαντώνται στο κείμενο, για την πλήρη κατανόηση των νοημάτων της Εργασίας.

Λέξεις κλειδιά Νανοπληροφορική, συγκριτικό πλαίσιο read-across, νανοϋλικά, ανάπτυξη προβλεπτικών μοντέλων, ασφάλεια στο στάδιο του σχεδιασμού, διαδικτυακές εφαρμογές

Table of contents

List of Tables

List of Figures

Abbreviations

AGPL	Affero General Public License
AOP	Adverse Outcome Pathway
APD	Applicability Domain
API	Application Programming Interface
BD	Biological Descriptor
BEGM	Bronchial Epithelial Cell Growth Medium
BSA	Bovine Serum Albumin
CA	Carbonic Anhydrase
CDK	Chemistry Development Kit
CRAN	the Comprehensive R Archive Network
CSF	Corrected Shape Factor
CSS	Cascading Style Sheets
CSV	Comma-Separated Values file
CT	ChymoTrypsin
DDT	1-DoDecaneThiol
DMEM	Dulbecco Modified Eagle's medium
EC	European Commission
ECHA	European Chemicals Agency
ENM	Engineered Nanomaterial
FN	False Negative
FP	False Positive
GA	Genetic Algorithm
gif	Graphics Interchange Format
GNU	GNU's Not Unix!
GO	Gene Ontology
GPL	General Public License
GSVA	Gene Set Variation Analysis
GUI	Graphical User Interface
HB	HaemogloBin
HD	Hydrodynamic Diameter
HOMO	Highest Occupied Molecular Orbital
HTML	HyperText Markup Language
IDE	Integrated Development Environment
ILP	Integer Linear Programming
InChI	International Chemical Identifier
ISO	International Organization for Standardization

IUPAC	International Union of Pure and Applied Chemistry
JSON	JavaScript Object Notation
KNIME	Konstanz Information Miner
*k*NN	*k*-Nearest Neighbours
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
LDH	Lactate Dehydrogenase
LOO	Leave-One-Out
LP	Linear Programming
LSPRi	Localized Surface Plasmon Resonance index
LUMO	Lowest Occupied Molecular Orbital
L5O	Leave-Five-Out
MAE	Mean Absolute Error
MCC	Matthews Correlation Coefficient
MeOx	Metal Oxide
MILP	Mixed-Integer Linear Programming
MINLP	Mixed-Integer Non-Linear Programming
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MWCNTs	Multi-Walled Carbon Nanotubes
NLP	Non-linear Programming
OECD	Organization for Economic Cooperation and Development
OF	Objective Function
OPLRA	Optimal Piece-wise Linear Regression Algorithm with regularisation
OPLRAreg	Optimal Piece-wise Linear Regression Algorithm
OS	Operating System
PCA	Principal Component Analysis
PCF	Protein Corona Fingerprint
PD	Physicochemical Descriptor
PDB	Protein Data Bank
PEG	PolyEthylene Glycol
PVP	PolyVinylPyrrolidone
QNAR	Quantitative Nanostructure Activity Relationship
QSAR	Quantitative/Qualitative Structure-Activity Relationships
RAAF	Read-Across Assessment Framework
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RMSE	Root Mean Squared Error
ROS	Reactive Oxygen Species
RPMI	Roswell Park Memorial Institute medium
SDF	Standard Database Format
SEM	Scanning Electron Microscopy
SF	Shape Factor
SMILES	Simplified Molecular-Input Line-Entry System
SPIONs	Super-Paramagnetic Iron Oxide Nanoparticles
SSA	Specific Surface Area

TEM	Transmission Electron Microscopy
TN	True Negative
TNF	Tumour Necrosis Factor
TP	True Positive
UNIPROT	Universal Protein Resource
WEKA	Waikato Environment for Knowledge Analysis
XML	eXtensible Markup Language
ΑΣ	Αντικειμενική Συνάρτηση
ΕΟΧΠ	Ευρωπαϊκός Οργανισμός Χημικών Προϊόντων
ΝΥ	Νανοϋλικά
ΟΟΣΑ	Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης

Chapter 1

Introduction

In earlier times, the discovery of new materials was the result of an iterative process consisting of three phases: design, synthesis and testing. Extensive modification and testing of proposed candidate materials was required, before the material was actually fielded, in order to assure the quality, safety and consistent performance of the material in the desired application. [6] This iterative process was time consuming, labor-intensive and expensive. [7] Ethical questions also arose, concerning the use of laboratory animals in *in vivo* experiments for assessing the safe use of new materials on human and environmental applications. [8] These limitations and shortcomings resulted in growing demands for developing accurate models to predict properties and behaviour of materials. Predictive models allow for screening and eliminating unpromising candidates before committing time and resources on synthesis and testing.

Among the various approaches that have been taken in computer-aided material design, the development and application of data-driven methods for training predictive models has gained increased popularity during the last few years, due to rise of computational power and the rapid advancements in the field of artificial intelligence, machine learning, and data storage technologies. [9]

Electronic structural representations of materials, such as the simplified molecular-input line-entry system (SMILES) or Protein Data Bank (PDB) 3D representations are combined with advanced physics-based computational tools and algorithms to calculate specific to the field structural descriptors. [10]–[12] These descriptions can be combined with experimental information and with descriptors calculated using image or omics data analysis techniques, in order to create full characterisation fingerprints of materials. These fingerprints can be organised in tubular formats, which are then introduced to machine learning or statistical algorithms for developing predictive models that correlate the materials structure with their activities, functionalities or possible adverse effects. These models are known as Quantitative Structure-Activity Relationships (QSARs). [6] Many examples of successful machine learning models applied in chemistry and materials science exist in Literature, including for example models that predict the thermodynamic parameters in catalytic processes, the critical temperatures of superconductors, crystals structures etc. [9] The development of QSAR models can be coupled with virtual screening techniques, [6] to prioritize compounds with desired properties in the circle of design, synthesis and testing and exclude compounds with undesired properties in the early stages of the analysis. Therefore, experimental time and cost can be reduced and the process of developing novel materials can become more efficient.

1.1 Engineered nanomaterials and nanotoxicity

A class of materials of particular interest are engineered nanomaterials (ENMs). According to European Commission's (EC) definition a nanomaterial is "a natural, incidental or manufactured material containing particles, in an unbound state or as an aggregate or as an

agglomerate and where, for 50% or more of the particles in the number size distribution, one or more external dimensions is in the size range 1 nm - 100 nm". [13] Due to their small size and their tunable physicochemical and biological properties, (E)NMs exhibit novel and unique behaviour that is not observed to their bulk counterparts. [14] For instance, the ENMs surface properties can be adjusted and enhanced by adding a coating to the "core" of the particle. The coating of ENMs can regulate their solubility, their charge and hydrophilic or hydrophobic character and consequently regulate their flow within a fluid. The type of coating also determines the level of interactions between ENMs and other molecules, such as proteins. These properties render them adequate for successful applications in industrial catalytic processes, energy storage, manufacturing, medical devices, diagnostics, therapeutics, cosmetics and sun-creams. [15]–[17] However these exact same properties in combination with their small size, facilitates the translocation of ENMs or other active chemical species bound to them (Trojan Horse effect), to organisms' tissues and organs. [14], [18]

1.1.1 Nanotoxicity

While the use of ENMs is expanding to various applications and commercial products, parallel results in the area of nanotoxicity have increased public concerns regarding their possible hazardous effects on human health and the environment. [19], [20]

The level of interactions between living organisms and ENMs are driven by the ENMs' physicochemical properties such as size, shape, surface chemistry and aggregation state. Available data from *in vivo* studies show that ENMs can penetrate cells through the cell membrane, accumulate there and in their nucleus. *In vitro* studies in cell cultures demonstrate the ability of nanostructures to elicit inflammatory responses, inhibit cell growth, and induce cell death (cytotoxicity). They may also lead to the production of reactive oxygen species (ROS), such as free radicals, which cause oxidative stress, which is responsible for DNA damage. Finally, cases of neurotoxicity and carcinogenesis are often reported due to the interaction of cells with ENMs. [15], [16], [21], [22] Therefore, prior to their broad release into the market great effort should be placed into the assessment of environmental and human health risks caused by the exposure to ENMs.

Figure 1.1 presents schematically the life cycle, the possible transformations and some of the possible toxicity effects on living organisms and Figure 1.2 depicts the routes of exposure of humans to ENMs and their possible risks of this exposure.

Interested readers may find more information on nanotoxicity in the publication of Yang *et al*, [16] which presents in a coherent manner all the routes of exposure, the parameters affecting nanotoxicity, the possible nanotoxicity effects on organisms, and the experimental assays that are adequate for the evaluation of different aspects of nanotoxicity.

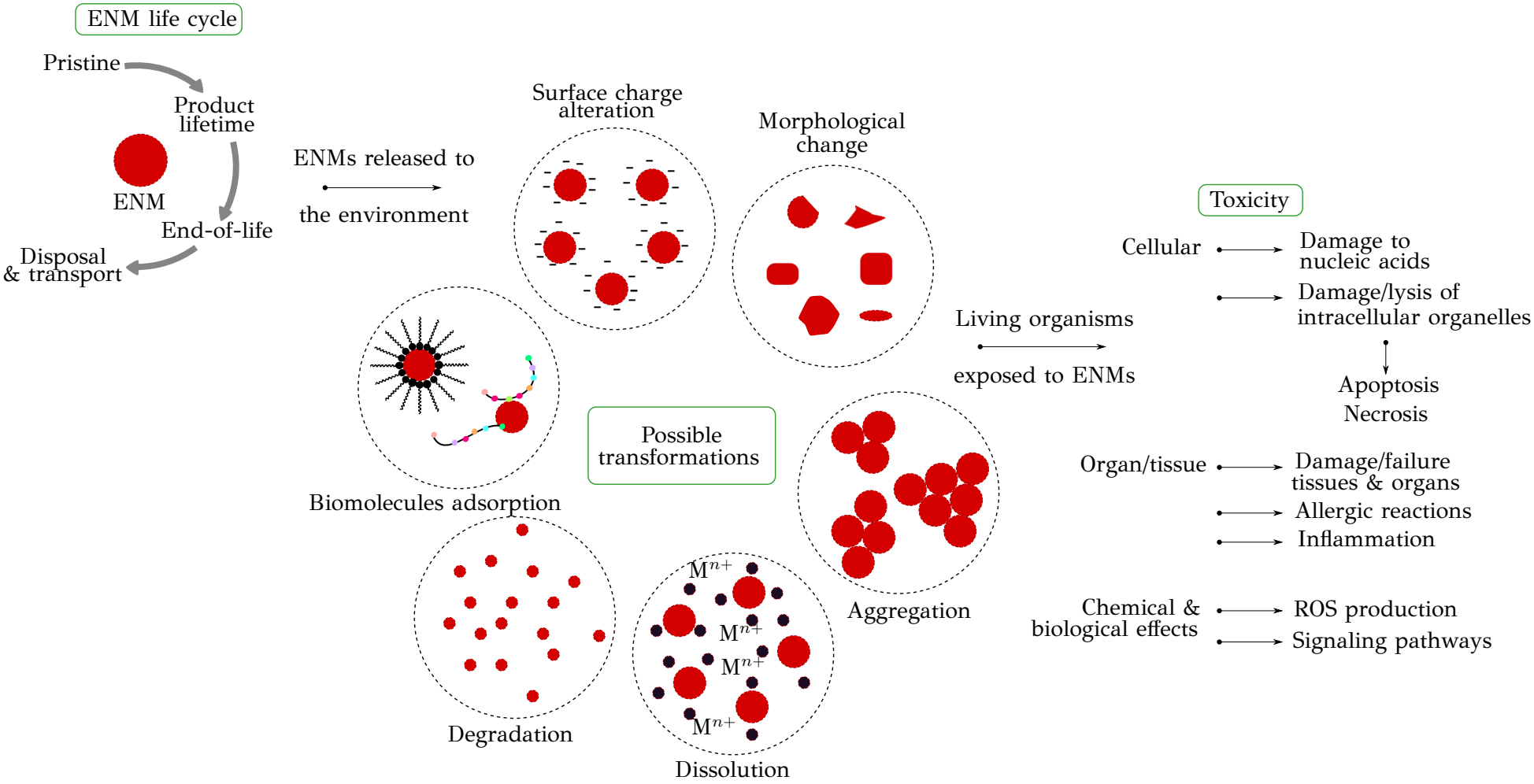


Figure 1.1: Life cycle, possible transformations, and toxicity of ENMs. The image is based on [15] and [16]. This image can be used under the following terms: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

1.1.2 Specific issues and challenges of the nanotoxicity assessment

The rapid increase in the use of novel ENMs in a noticeably short period, renders almost unattainable their complete and systematic experimental evaluation, due to the complexity of intrinsic and extrinsic properties to be considered. In addition to the time constraints, experimental techniques are particularly expensive regarding the cost of the consumables and the laboratory equipment. [8], [24] It is reported in Literature that the conventional full-scale testing of chemical substances costs about $5 million per substance and it may require decades to be completed. In the US alone, it has been estimated that toxicity testing for the existing ENMs in the market has cost between $249 million and $1.18 billion. These numbers may change as novel ENMs are introduced. [25], [26]

Furthermore, the use of animals in preclinical trials is an issue which is subjected to national and international legal restrictions regulating their use and sacrifice. In fact, according to the European REACH Regulation[1] experiments on vertebrate animals "shall be undertaken only as a last resort". [27] Last but not least, another aspect that should be considered during ENMs risk evaluation, is the researchers' personal ethical dilemmas. The use and killing of animals (e.g. euthanasia to reduce animals' suffering) for research purposes, apart from the fact that is opposed to the animal prosperity, may greatly affect researchers' mental well-being. [8]

Over the past few years, the nanosafety community has encouraged the development of alternative non-testing methods for the toxicological investigation of ENMs introducing *in vitro* and *in silico* methods. The so-called "nanoinformatics" field includes novel, computational approaches which can produce reliable predictions for the toxic and biological behaviour of ENMs (endpoints). These computer-aided methods aim to contribute to the prioritization of ENMs and to support the regulatory decision-making. [19], [28] In addition these methods facilitate the estimation of either long or short-term hazardous effects, prior to their production. Therefore, they make possible the targeted design of all the novel nano-structures aspects including their functionality and their safety, in advance (safety-by-design).

One successful approach for nanotoxicity computational assessment, is the adaptation of the QSAR modelling methodologies [29] to the special requirements of ENMs, which are due to their complex structures. The produced models are presented in the Literature as nanoQSARs or QNARs (quantitative nanostructure-activity relationship) models. Comprehensive reviews of nanoQSAR modelling methods and produced predictive models have been published recently in the Literature. [30], [31] A repository of nanoQSAR models is included in the final report of the Nanocomput project and is freely available through the European Commission Science Hub. [32]

However, in order to ensure the functionality of the nanoQSAR approaches, sufficiently large (more than 20 samples for just one endpoint) and diverse datasets should be provided, in order to avoid over-fitting of the model to the training data. [33], [34] Thus, the development of nanoQSAR models is often limited by the size of the available experimental nanotoxicity datasets, which are not big enough or are unbalanced regarding the number of available ENMs samples and descriptors to allow sufficient testing of the accuracy and the robustness of the produced models. [33] In addition, nanoQSAR approaches assume a common mechanism of toxicity which is also a prerequisite for modelling. Nevertheless, ENMs present a wide variety of structural groups and considering all the possible transformations that may occur (see Figure 1.1), a common mechanism of toxicity should not be expected. [35]

[1] Regulation (EC) No 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals

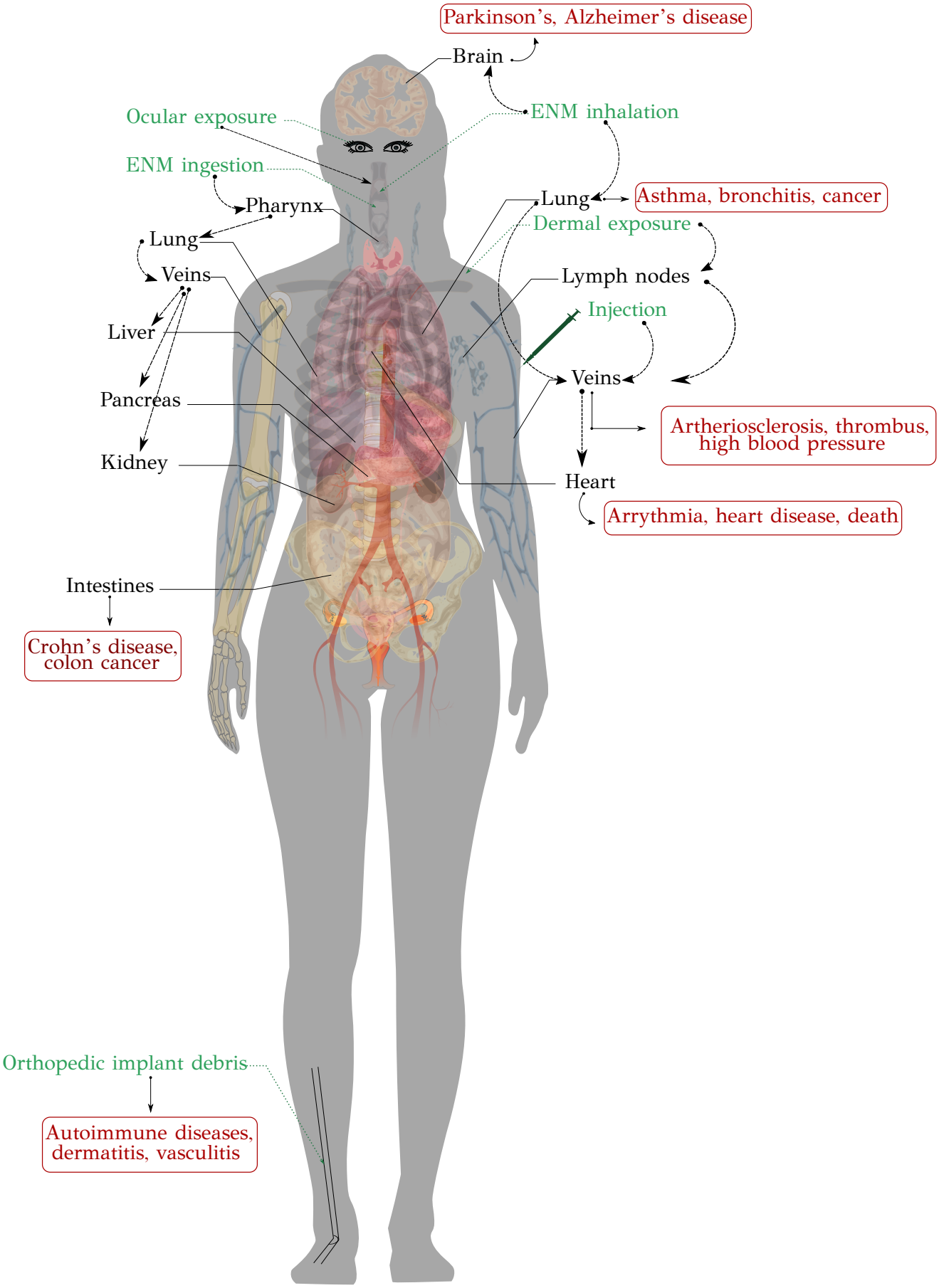


Figure 1.2: Routes of ENMs human exposure and uptake (green), and potential ultimate risks (red). The image is a remix from `https://openclipart.org/` based on [22], [23]. This image can be used under the following terms: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

1.2 The alternative read-across approach for predicting properties and adverse effects of chemical substances

Read-across is an alternative approach for endpoint predictions and hazard assessment of chemical substances. In contrast to QSAR modelling, read-across can rely on small datasets. The read-across concept is based on the empirical knowledge that similar substances may exhibit comparable properties thus, the estimation of the hazardous effects of non-tested chemical substances ("target" substances) can be achieved using data within a group of comparable substances ("analogous-source" substances). [33], [36], [37] Read-across is one of the most commonly used alternative approaches for data gap filling in registrations submitted under REACH.

Although the read-across concept is simple, the actual application of the read-across methodology for obtaining endpoint predictions for unknown substances, is still an open scientific field. The most challenging task is the selection of properties that can be used effectively for defining the similarity between two chemical substances and the partitioning of substances into groups of structurally similar constituents.

The European Chemicals Agency (ECHA) developed the Read-Across Assessment Framework (RAAF) as an internal tool for assessing the predictions, based on read-across, of properties of substances in the context of the REACH Regulation. [38] RAAF has been made publicly available to support the development and improvement of the read-across methods in a way that they can fulfill the requirements of the REACH Regulation. A key concept in this framework is the definition of similarity between substances. According to RAAF, [38] similarity should be based on a variable that is straightforwardly connected to a structural property. Various similarity definitions between substances have been proposed, including chemical similarity (common functional groups), similarity through (bio)transformation (common predecessors and/or decomposition products) and the existence of a pattern in biological, physicochemical, toxicological and/or environmental fate properties. These similarities may lead to the formation of groups of substances used for data gap filling.

The read-across approach has been used successfully in the area of cheminformatics in various applications including the assessment of the genotoxicity of pesticides [39] and the mutagenicity of aldehydes and ketones. [40] The application of the read-across concept in the field of nanoinformatics is more challenging compared to simple chemicals, due to the more complex structures and the multi-perspective characterisation that includes structural, kinetics and biological properties.

1.2.1 Read-across approaches

There are two read-across approaches, supported by ECHA and Organization for Economic Cooperation and Development (OECD), namely the analogue and the category/grouping approach. The definitions of the two approaches slightly differ between ECHA and OECD, [41] however their eminent characteristics are presented in Table 1.1.

The two approaches are explained next in more detail with emphasis on use of read-across for the prediction of ENM properties. It should be noted that the borderline between the two approaches is still vague and depends on the number of available samples. [20], [41], [42]

1.2.1.1 The analogue approach

In the analogue approach the prediction is limited to a small area of the data space, where regular patterns or trends cannot be established. One source ENM can be used for the endpoint estimation of a single or more target ENMs, or two or more source ENMs can be used to make predictions for a single or several target ENMs. More specifically, the toxicity endpoint for an untested ENM can be estimated using "similar" source ENMs, where "similarity" is

based on mathematical definitions (Euclidean distance, cosine similarity etc.), data mining techniques or the critical opinion of experts. [20], [41]

The read-across methodologies apply an interpolation strategy "locally" among similar samples which, depending on the provided data -numerical or discrete-, can be quantitative or qualitative. [35] The methods for the prediction of each endpoint range from simple average value calculations, or simple linear interpolations to more complicated methods applying (e.g. *k*NN, partial least squares, random forests). [33], [43]

1.2.1.2 The category approach

In the category approach, the ENM samples are organized into groups of similar compounds. As a result, the toxicity endpoint properties will either all be similar or follow a regular pattern. [38], [44] Groups are formed considering structural similarities or/and dissimilarities between samples, and it is assumed that due to these similarities, the biological or toxic activity of the ENMs within a group follows a regular pattern. Groups of ENMs can be further divided into subgroups based on interdependencies in nanodescriptors and the formation of these subgroups can be "tuned" in order to gain satisfactory predictions. [42], [45] Other studies have investigated alternative grouping possibilities including Principal Component Analysis (PCA) [46], linear discriminant analysis (LDA), [42] two-dimensional hierarchical clustering [35] or considering the ENM mode-of-action. [47] For the estimation of the endpoint of a target ENM in a group, the analogue approach can be applied.

Table 1.1: Overview of the two read-across officially supported approaches. [38], [41]

Analogue approach	**Grouping approach**
Employed between a small number of structurally similar substances (source and target substances).	Employed between several substances that have structural similarity.
No trend or regular pattern on the properties.	A trend or a regular pattern is expected (in order to accept or reject the grouping hypothesis).
Evaluation of each sample independently.	Evaluation of the category as a whole.
Worst case: single source substance (one neighbour).	Worst case: the strength of effects in a target sample within the group is expected to be lower than the strength of effects observed for the source sample.

ECHA -to harmonize the different approaches- has recently presented a systematic ENMs specific workflow for grouping and read-across in the document titled "Recommendations for nanomaterials applicable to the guidance on QSARs and Grouping". The entire approach consists of seven well defined steps (Figure 1.3): [48]

1. Determination of structural characteristics of ENMs (composition, including surface chemistry and any impurities, size, shape etc.).

2. Development of an initial grouping hypothesis that correlates an endpoint (e.g. a toxicity index), to different behaviour and reactivity properties, including solubility, zeta potential, dispersibility, hydrophobicity, dustiness, biological activity (redox formation, gene expression), photoreactivity etc. [35], [47], [48] Assignment of the samples to groups.

3. Gathering of the above properties (depending on the grouping hypothesis) for each ENM.

4. Construction of a data matrix including properties and endpoints.
5. Assessment of the applicability of the approach using computational techniques (e.g. PCA, hierarchal clustering, [35], [48] random forests, [48] LDA, [42] etc.) and data gaps filling. If no regular pattern can be emerged, an alternative grouping hypothesis must be proposed (step 2).
6. In case that the grouping hypothesis is robust, but adequate data are not available, additional testing should be considered for data gap filling.
7. Justification of the method.

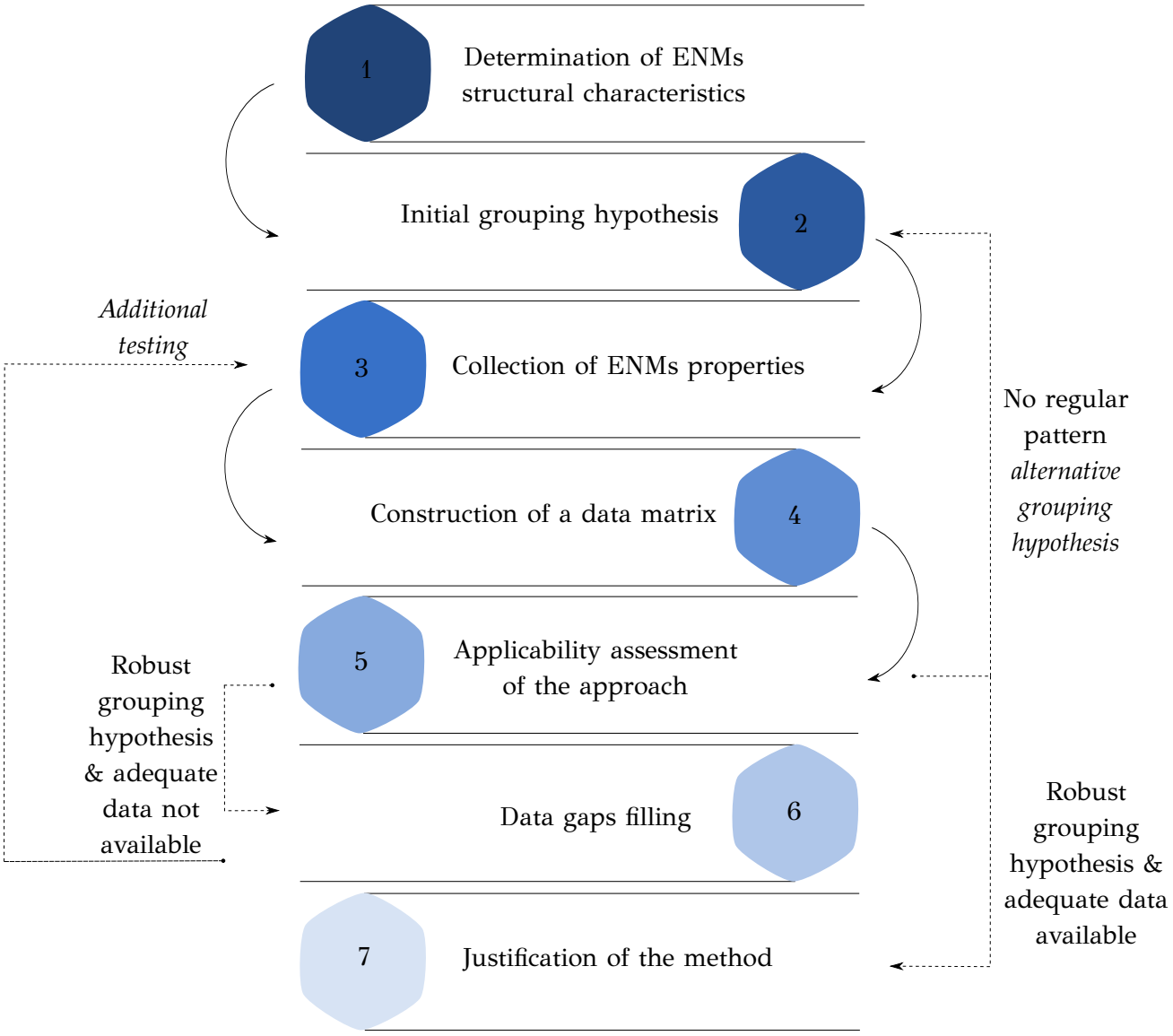


Figure 1.3: ECHA proposed grouping/read-across step-wise approach

This workflow was slightly modified by Lamon *et al.* [37] and Aschberger *et al.* [49] who presented a simplified version consisting of four steps:

1. Identify the (nano)forms of the substance.

2. Gather the available data, evaluate them for adequacy and reliability and build the data matrix.

3. Develop a grouping hypothesis and assign the source analogues to groups.

4. Assess the applicability of the grouping hypothesis and fill data gaps.

The simplified workflow was used to develop case studies for the read-across prediction of hazard endpoints of nanoforms of TiO_2 and of Multi-Walled Carbon Nanotubes (MWCNTs) respectively. The first of these studies has been released as an official OECD document. [50]

The read-across workflow proposed by ECHA assumes a hypothesis, which is evaluated and assessed in terms of its adequacy to fill data gaps (step 2). The read-across hypothesis may involve both the selection of the most informative descriptors and the definition of the source ENMs, that can be considered as neighbours to the target ENM. This procedure is iterated in a manual trial-and-error fashion, including experimental data collection, until a hypothesis producing successful read-across predictions is determined. The procedure is time-consuming and due to the complexity of the problem, it does not guarantee that the produced read-across model is optimal.

An important step in the proposed ECHA workflow is the identification of the parameters which may affect the risk posed by nanoforms (step 1). [48] This step can be connected to the variable selection process in predictive modelling, which aims at removing non-informative or noisy features, reducing the dimensionality and improving the reliability and the performance of the produced model. [51]

1.3 Scope

The present PhD Thesis aims to contribute to the field of read-across by developing novel, fully automated and validated approaches. Although the methods developed in this Thesis can be applied to all types of chemical substances, particular emphasis has been given to the prediction of ENM properties. Our primary goal is to provide the scientific community with reliable predictive modelling methodologies that can reduce the time and cost spent on experimental testing during the risk assessment of ENMs, e.g. prioritize ENMs for experimental evaluation with a rapid screening and reduce the use of animals in experiments, which is also harmonized with the European Union directives.

Understanding the specific properties and the modes of action that lead to undesirable behaviour in organisms or in the environment, allows material scientists to design these properties out, early in the design phase (safety-by-design). All proposed read-across workflows and models have been implemented as user-friendly web applications, which makes them open, transparent and easily accessible to all interested stakeholders in Academia, in the Industry and in Regulatory Agencies.

The proposed methods consider both key components of the read-across procedure as optimisation parameters: the variable selection and the boundaries that define the neighbourhood of the query ENM, for which a read-across prediction is sought. The developed methodologies are constructed around two main strategies; the "thresholding" (§1.3.1) and the "*k*-nearest neighbours" (§1.3.2) strategy.

1.3.1 [m] The thresholding strategy

In the thresholding strategy two ENMs are considered as neighbours if they satisfy a predefined similarity threshold (Figure 1.4). In detail, the distances between the available ENMs are calculated, according to their known properties. If the distance between two ENMs is lower or equal to the threshold value, then these ENMs are neighbours.

With this strategy, the number of neighbours is not the same for all query ENMs due to local similarities that may lead to a dense or a spare neighbourhood. In fact, in cases of a strict threshold, no neighbours may be found for a query ENM.
Another advantage of this proposed methodology, is that it takes into account -whenever possible- the multi-perspective characterisation of ENMs by grouping ENM descriptors into categories (e.g. physicochemical, biological, quantum mechanical, image or biokinetics) and by using multiple similarity criteria for defining neighbours to the target ENM.

More specifically, distance measures are calculated between all ENMs separately for the available categories of descriptors. For each query ENM in the available set, the train ENMs for which both distance measures satisfy the corresponding thresholds are selected as neighbours. Therefore, in order to characterize two ENMs as neighbours they need to satisfy all similarity criteria (Figure 1.5).

1.3.2 [m] The *k*-Nearest Neighbours strategy

The *k*-Nearest Neighbours (*k*NN) method belongs to the "lazy" (instance based) learning techniques, that classify an instance based on the *k* closest training examples (neighbours) in the feature space (Figure 1.7). In case of classification problems, each instance is assigned to the class indicated by the weighted majority vote of the *k* closest neighbours. [52] In regression modelling problems, prediction is calculated as the weighted average of the endpoint of the *k* selected neighbours. This prediction scheme places the *k*NN method among the read-across strategies, as it requires only a few neighbouring – in terms of similarity – ENMs, in order to predict the respective endpoint class. [53]

In Figures 1.6 and 1.7 and in Table 1.2 the differences between the two developed strategies are presented.

Table 1.2: Overview of the two developed read-across strategies.

Thresholding approach	***k*NN approach**
Closest neighbours have greater impact on the final prediction	Closest neighbours have greater impact on the final prediction
Susceptible to local similarities	Possible selection of too "distant" neighbours (great *k* value) or overfitting (small *k* value)
Prediction may be impossible for some input ENMs (no neighbours found)	Prediction for any input ENM

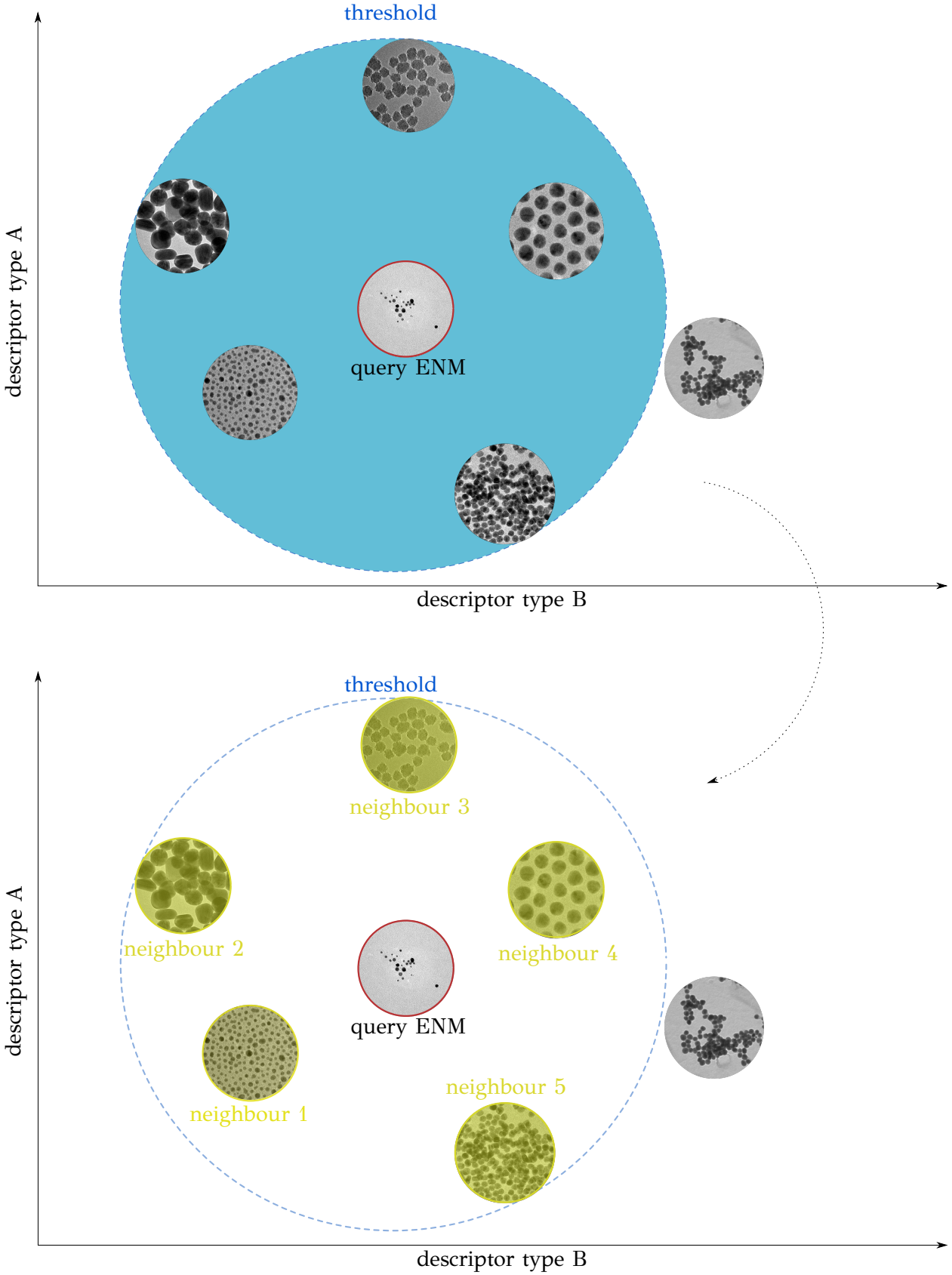


Figure 1.4: Schematic representation of the thresholding approach using a single similarity threshold. The untested ENM (red-framed ENM) and the tested ENMs are placed into the multidimensional space (for simplicity a 2D coordinate system) defined by their descriptors. The similarity threshold is the distance limit that tested ENMs should meet in order to be considered as neighbours of the query ENM (yellow-shaded ENMs). This image can be used under the following terms: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

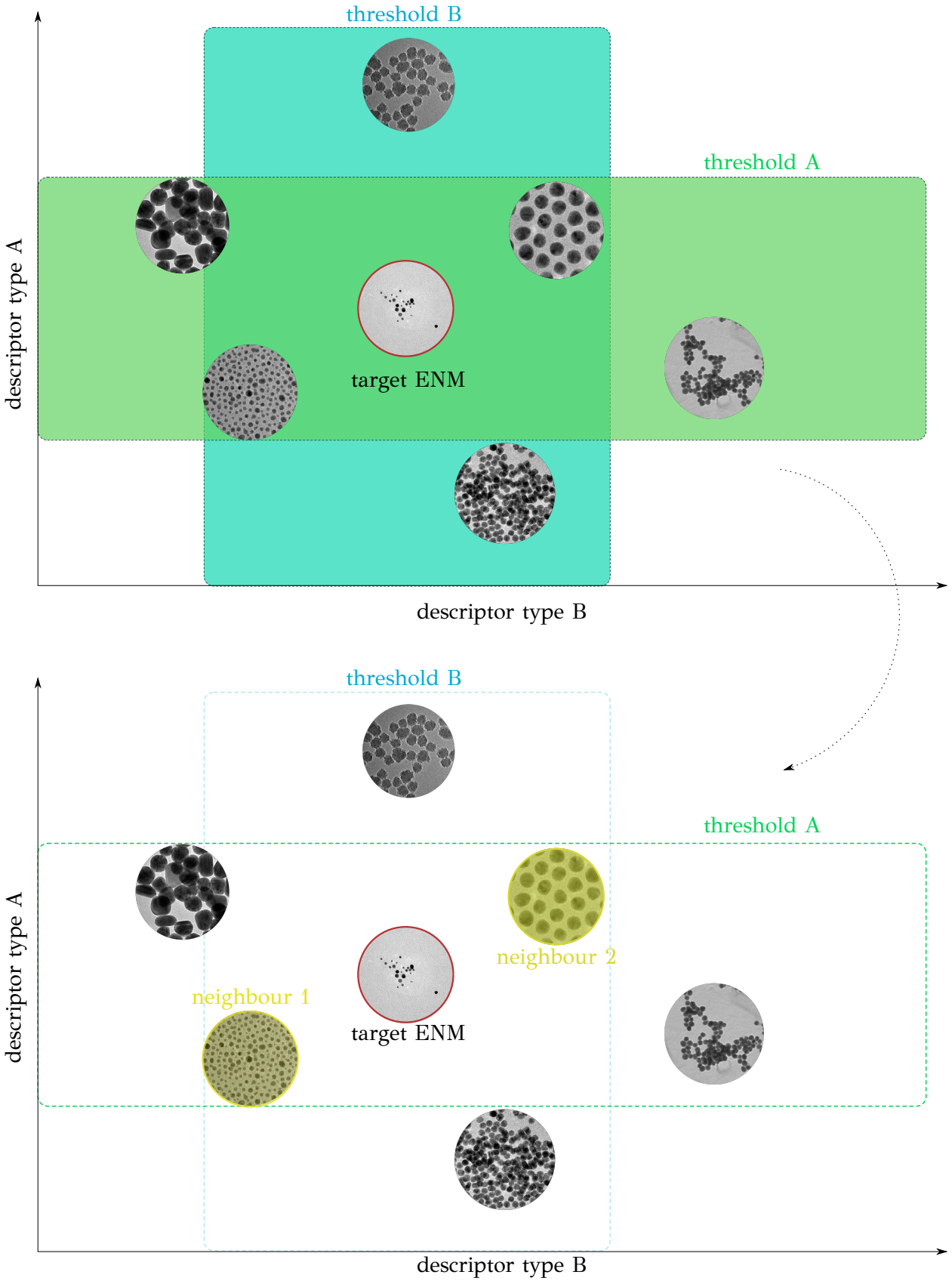


Figure 1.5: Schematic representation of the thresholding approach using two similarity thresholds. The untested ENM (red-framed ENM) and the tested ENMs are placed into the multidimensional space (for simplicity a 2D coordinate system) defined by their descriptors. The similarity thresholds define the candidate neighbours of the query ENM considering only one type of descriptors. A tested ENM should meet both threshold limits in order to be considered as a neighbour of the query ENM (yellow-shaded ENMs). This image can be used under the following terms: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

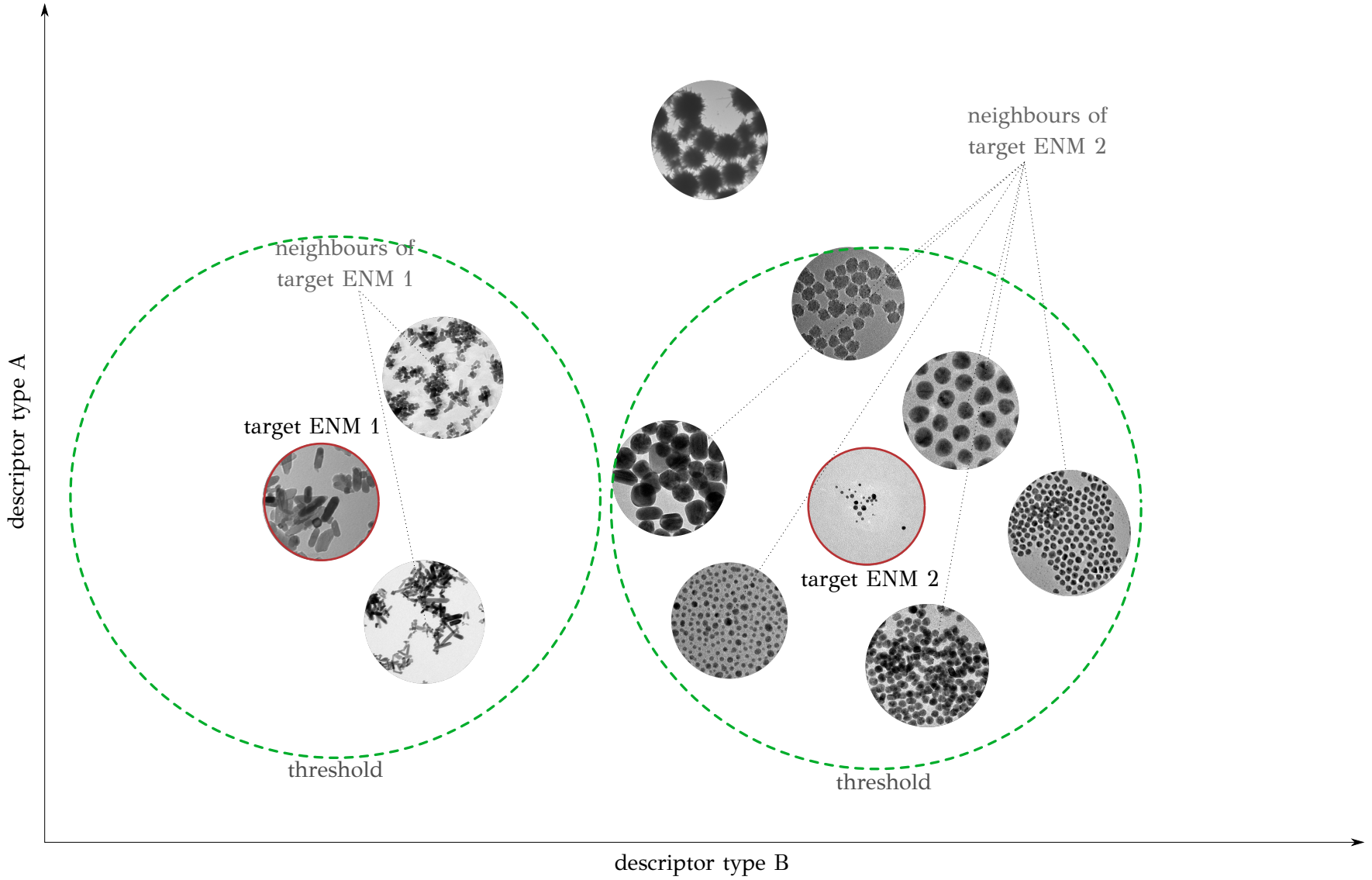


Figure 1.6: Schematic representation of the thresholding read-across approach in a collection of ENMs using a single similarity threshold. The untested ENM (red framed ENMs) and the tested ENMs are placed into the multidimensional space defined by their descriptors. The similarity threshold defines the neighbours of each untested ENM. In this approach, the number of neighbours of each ENM is not pre-determined. This image can be used under the following terms: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

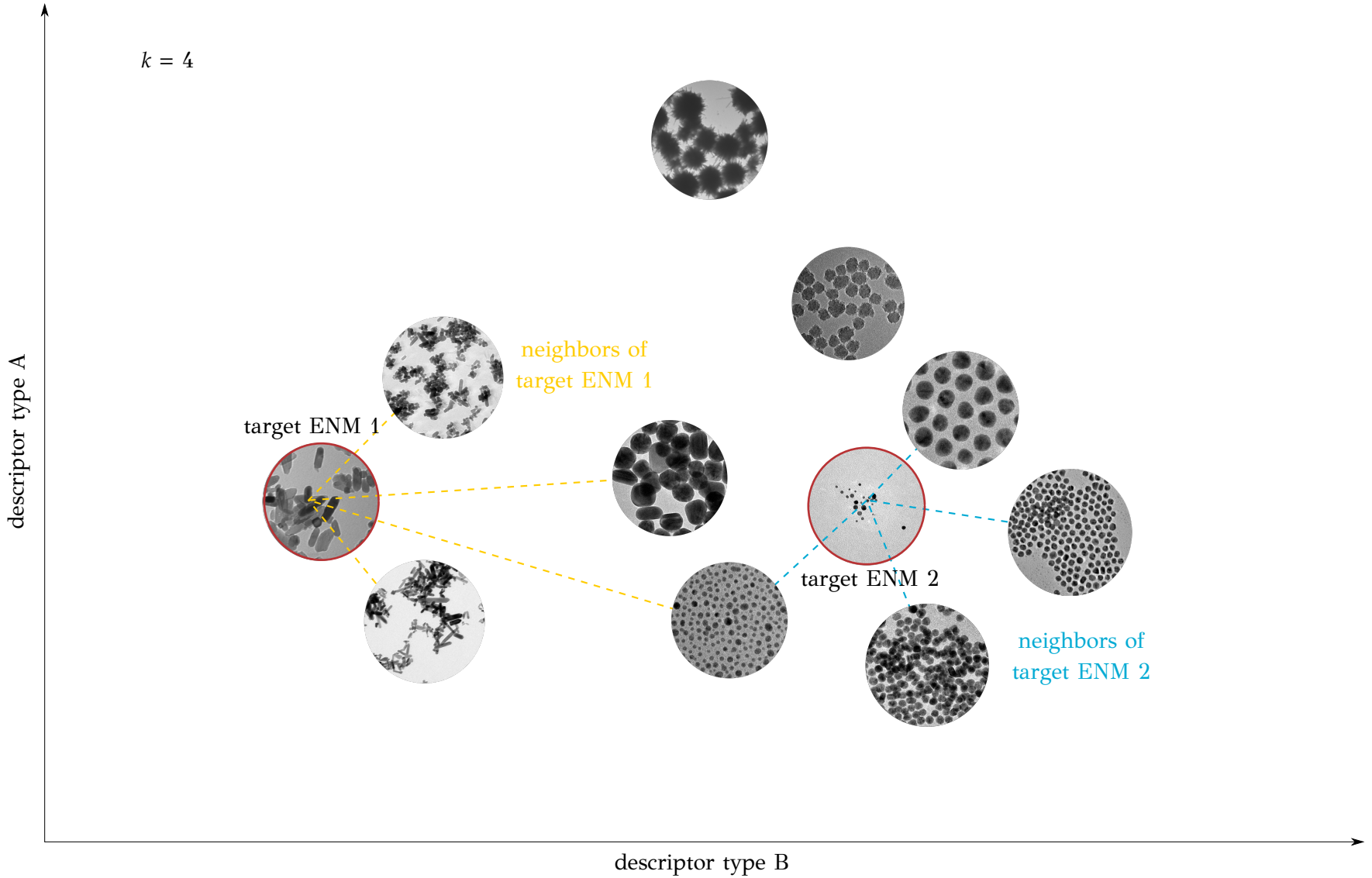


Figure 1.7: Schematic representation of the *k*NN read-across approach in a collection of ENMs. The untested ENM (red framed ENMs) and the tested ENMs are placed into the multidimensional space defined by their descriptors. The *k* nearest tested ENMs are the neighbours of the untested ENM. This image can be used under the following terms: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

1.4 Structure of the Thesis

The Dissertation consists of eight Chapters. The first Chapter introduces the readers to the topic of this PhD Thesis. Chapter 2 provides a review of the computational methods and workflows in data-driven computational modelling with emphasis on data preprocessing steps and validation techniques that are used extensively in the next chapters for preparing the data and testing the proposed read-across methods. Chapter 3 presents the benchmark datasets on which the read-across methods will be applied and validated.

The developed methodologies are presented in detail in Chapters 4-7. Chapter 4 presents a novel read-across methodology for the prediction of toxicity related endpoints of ENMs. The proposed method lies in the interface between the two main read-across approaches, namely the analogue and the grouping methods, and can employ a single criterion or multiple criteria for defining similarities among ENMs. The main advantage of the proposed method is that there is no need of defining a prior read-across hypothesis. Based on the formulation and the solution of a mathematical optimisation problem, the method searches over a space of alternative hypotheses, and determines the one providing the most accurate read-across predictions. The procedure is automated and only two parameters are user-defined: the balance between the level of predictive accuracy and the number of predicted samples, the number of selected variables and the similarity criteria, which define the neighbours of a target ENM. This methodology is adequate for the prediction of both numerical and categorical endpoints.

The developed automated workflow is integrated in a web application named Apellis, presented in the same Chapter. During the training procedure, the application selects the most important ENM properties of concern that affect their toxic behaviour. In the process of grouping ENMs for performing read-across predictions, the multi-perspective characterisation of ENMs can be taken into account, by defining more than one similarity criteria. The workflow converges to the grouping hypothesis that leads to the most accurate read-across estimations. Visualisation tools are included in the application, which offer better and more clear understanding of grouping and similarities among ENMs. The trained models can be saved in an electronic format, so that they can be easily retrieved, for calculating new predictions. In addition, this allows model developers to disseminate and share the produced models with the community.

In Chapter 5, a computational workflow is presented for grouping ENMs and for predicting their toxicity-related endpoints. A mixed integer linear optimisation program (MILP) problem is formulated, which automatically filters out the noisy variables, defines the grouping boundaries and develops specific to each group predictive models. The method is extended to the multi-dimensional space, by considering the ENM characterisation categories as different dimensions. The performance of the proposed method is illustrated through the application to benchmark datasets and comparison with alternative predictive modelling approaches. The trained models using the above datasets were made publicly available through a user-friendly web service named vythos. The MILP methodology was further extended by making the necessary adaptations in order to define ENM grouping boundaries according to their endpoint as presented in Chapter 6.

Finally, in Chapter 7 two read-across models using the *k*NN strategy are presented. The first one predicts the biological and toxicological profile of decorated MWCNTs using only computational descriptors based on their surface ligands. The second one predicts the ENMs zeta-potential utilizing as input data their type of core and an image descriptor derived from transmission electron microscopy images. Both models are released to the scientific community as web services.

The Appendices present technical details of the software and additional results related to the present Thesis. Appendices A′ and B′ present all the informatics tools and programming software used in order to build all the devised modelling strategies. Appendix Γ′ presents

information about the descriptors of the datasets and Appendix Δ′ additional results regarding the methodology of Chapter 4. The results, the data and the developed software of the present Thesis are freely-available to any interested researcher. The sources of these data are presented in Appendix Ε′. Finally, an index containing small explanations in Greek of the main concepts of the Thesis is attached by the end of the manuscript.

The present manuscript is structured in a way that every Chapter can be read and understood independently. An encoding has been used, before every title and subtitle of the Chapters, in order to guide the readers concerning the content of each part:

t theory
m devised methodology
r results

Chapter 2

Computational tools for the prediction of material properties and adverse effects

This Chapter provides an overview of the most important data-driven computational methods and tools that are used in the field of cheminformatics for developing models for the prediction of properties, functionalities and adverse effects of materials. It includes a presentation and description of the validation methods and metrics proposed by OECD [54] for testing the accuracy, the performance and for defining the domain of applicability of the predictive models. The seminal work of Puzyn *et al*. [55] was the first attempt to adapt previous knowledge on data-driven methods to the development of predictive models for ENMs. Since then, many predictive models have been developed in the context of the young, but rapidly evolving field of "nanoinformatics". The computational tools will be presented this Chapter, with references to nanoinformatics applications. [29], [35], [56]–[58]

In general, a predictive modelling workflow consists of a sequence of steps including data collection and preparation, model development, internal and external validation of the produced model, definition of the applicability domain and release to the community. [59] These steps are presented in the following paragraphs. In particular, the validation methods and criteria will be used extensively in the next Chapters to assess, validate and compare the performance of the read-across methods on nanoinformatics case studies.

2.1 [t] Data preprocessing

Data quality is the cornerstone of a successful and reliable predictive modelling. For this reason, data preprocessing -including manipulation and filtering methodologies- should be applied to the datasets prior to their analysis.

2.1.1 Normalisation

It is commonly observed that within a dataset attribute values have considerably different numerical ranges. However, unless data transformation is performed, attributes with a broad range of values will outweigh attributes with lower range of values and will influence the analysis results. This is the particular case of distance-based machine learning methods (e.g. *k*NN classifier). [6] Additionally, differences in features' range may affect the results of loss function when regularisation is used.

By applying data normalisation[1], each attribute has an equal chance or "weight" of con-

[1]Also referred as standarisation or scaling, in data analysis.

tribution to the analysis. [6] It is also advisable to perform normalisation when there is no former knowledge of the analysed data (e.g. to avoid the relativity on the selection of an attribute's measurement units). [60] There are several methods of data normalisation; we are presenting below the two used in the present Thesis.

2.1.1.1 Range normalisation

In this type of normalisation, also known as min-max normalisation- the Eq. 2.1 is used in order to transform the data to fall within the range of [0, 1]. Other small standarisation ranges, such as [-1, 1] or [-0.5, 0.5], are commonly used with the appropriate alterations in this linear expression.

$$\tilde{\mathrm{X}}_i = \frac{\mathrm{X}_i - \min(\mathbf{X})}{\max(\mathbf{X}) - \min(\mathbf{X})} \tag{2.1}$$

where $\tilde{\mathrm{X}}_i$, is the scaled value of the ith instance of $\mathbf{X}$ variable vector,
X_i, is the original value of the ith instance of $\mathbf{X}$ variable vector and,
$\min(\mathbf{X})$ and $\max(\mathbf{X})$, are the minimum and the maximum instance values of $\mathbf{X}$ variable vector.

2.1.1.2 Gaussian normalisation

In Gaussian, or z-score normalisation, the original data are transformed, in order to gain mean value equal to 0 and standard deviation equal to 1 (Eq. 2.2).

$$\tilde{\mathrm{X}}_i = \frac{\mathrm{X}_i - \bar{\mathbf{X}}}{S} \tag{2.2}$$

where $\tilde{\mathrm{X}}_i$, is the scaled value of the ith instance of $\mathbf{X}$ variable vector,
X_i, is the original value of the ith instance of $\mathbf{X}$ variable vector,
$\bar{\mathbf{X}}$, is the mean value of $\mathbf{X}$ variable vector and,
S, is the standard deviation of $\mathbf{X}$ variable vector.

2.1.2 Variable selection

The efficiency of a data-driven predicting methodology is largely affected by the information provided through the input variables. Inclusion of non-informative variables may increase the computational time needed by the modelling algorithm and deteriorate the performance of the produced predictive model. The filtering of noisy variables, also known as variable selection, is a key preprocessing step that speeds up the analysis and contributes to avoiding over-fitting. [6], [52] One or more of the following actions are routinely performed in data analysis to eliminate non-informative descriptors from the training set.

- Treatment of missing and inaccurate values, and duplicate data.
- Removal of features with low variance.
- Removal of features with no correlation (e.g. Pearson correlation) to the endpoint.
- Removal of highly correlated features (avoid over-representation of information).
- Dimensionality reduction (e.g. performing PCA).

In most of the read-across methodologies presented in this Dissertation, the variable selection process has taken into account from the starting conceptualization and design phase, and has been integrated as a component in the mathematical programming models that will be presented in the next Chapters.

2.2 [t] Computational modelling

This part of the workflow aims at establishing correlations between the ENMs properties and their adverse behaviour, by analysing and modelling the available data. Various modelling algorithms can be employed for developing the predictive models, including standard statistical modelling methods and advanced machine learning methodologies. [61] Depending on the type of the endpoint to be predicted (continuous or categorical), machine learning methods are divided into regression and classification methods respectively.

In Table 2.1 some machine learning algorithms that are commonly used in nanoinformatics (including regression and classification) algorithms are presented.

2.3 [t] Modelling validation methods

The use of *in silico* approaches in the field of nanotoxicity assessment may be necessary in order to reduce experimental time, resources and animal testing. However, a proposed methodology should fulfil a series of requirements that ensure its reliability and validity. [62] In brief, ECHA proposes that predictive models must meet the OECD principles for the validation of predictive models: [54], [62]

> » *To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:*
>
> 1. *a defined endpoint;*
> 2. *an unambiguous algorithm;*
> 3. *a defined domain of applicability;*
> 4. *appropriate measures of goodness-of-fit, robustness and predictivity;*
> 5. *a mechanistic interpretation, if possible.*

In general in the present Thesis the developed nanoinformatics methodologies are always validated both internally and externally.

2.3.1 Internal validation

In order to evaluate the robustness (stability of the parameters in slight alterations of the training data) and predictivity of the developed models, and to reduce the bias produced from a possible unbalanced representation of the classes between the training and the test set, cross-validation procedures are applied. In its general form (k-fold cross validation), different proportions (k subsets of approximately the same size) of the samples are successively excluded from the training process. The remaining samples are used in model training. The model is later used to generate predictions for the omitted samples. This procedure is repeated k times and gradually the prediction of the endpoint of each compound is calculated.

In the case that k equals the number of the samples of the dataset (N), the method is called leave-one-out (LOO) cross-validation. Each sample is removed, one at a time, and N models are built, using every time the rest $N-1$ samples. The developed model is applied

Table 2.1: Machine learning algorithms used in the nanoinformatics field.

Type	Algorithm
Regression	Simple Linear Regression
	Multiple Linear Regression (MLR)
	Least Absolute Shrinkage and Selection Operator (LASSO)
	Artificial neural networks
	*k*NN
Classification	Support Vector Machines
	Bayesian Methods
	Decision trees
	*k*NN
	Logistic Regression
	Random Forests
	Naïve Bayes
	Partial Least Squares (classification)

to the remaining sample in order to compute the endpoint prediction. This method uses the maximum possible number of training data and also can be used to compare different model methodologies fairly, because the deletion scheme is unique. [52], [54]

2.3.2 External validation

When datasets are quite large, it is advisable not to use only cross-validation for testing the predictive power of the modelling: although a low accuracy statistic from cross-validation implies a poor modelling result, the opposite is not equivalent [63]. Therefore, an external validation using two separate sets is considered necessary. In this case the original dataset is divided into training and test sets. The training set is used to estimate the model's parameters and the test set is excluded from modelling and is only used to measure model's performance. According to the reliability of the predictions, adjustments on the model may be needed. [64]

When the available volume of data allows it, the external validation scheme can be extended, by splitting of the original dataset in two phases: [65] first the dataset is split into training and test set, and the training set is further divided into calibration and validation subsets. As before, the training set is used for model development, however this time due to the use of the two subsets an inner feedback loop is applied. The model parameters definition is performed using the calibration set and their efficiency is measured on the validation data. According to the results on the validation subset, adjustment of the parameters may take place in order to improve predictions accuracy. The test set is not involved in the modelling workflow, and is used only as an external-blind set. In that way the modelers can simulate the performance of the model on new-untested data. Dataset partition into training-test sets or train-validation-test sets may be performed using a random partition, a stratified-random partition (in order to ensure balanced representation of the classes) or a method of representative subset selection.

2.3.2.1 Representative subset selection

In order to achieve a reliable external validation the training and test subsets should comprise instances of the original dataset, uniformly covering the available data space. The formation of such representative subsets is also required when the original datasets are so large that cross-validation schemes require extensive computational resources. This representative data

partition can be performed using a random -but stratified as to the dependent variable- sampling, a "uniform design" or a "cluster-based design" technique. [66]

One of the renowned techniques of uniform design is the Kennard and Stone algorithm [67] that is based on the selection of "dissimilar" (in terms of distance) instances starting from the "boundaries" of the data space and achieving in this way a regularity in subset formation. Starting from a dataset of N instances and P properties (dependent variables), in order to form a subset of n instances, the squared distance (Eq. 2.3) is calculated between all instances based on the P available properties. The instances that are the farthest apart (most dissimilar) are the first two selected instances for the subset of interest. In continuation, for each unassigned instance ν the squared distance from all instances k already in the subset is calculated and assigned to array $\Delta^2_\nu(k)$ (Eq. 2.4). The instance $k+1$ that will be included to the subset, is the most distant one from the existing instances in the subset (criterion of Eq. 2.5). This procedure is repeated until the desired number of instances n has been reached.

$$D^2_{i,j} = \sum_{p=1}^{P} (x_{i,p} - x_{j,p})^2 \tag{2.3}$$

where $D^2_{i,j}$ is the squared distance between instances i and j and,
$x_{i,p}$ is the value of the independent variables for instance i on property p.

$$\Delta^2_\nu(k) = \min\{D^2_{\nu,1^*}, D^2_{\nu,2^*}, \cdots D^2_{\nu,k^*}\}_\nu, \nu \neq i^* = 1, 2, \cdots N \tag{2.4}$$

where $\Delta^2_\nu(k)$ is an array containing the squared distances from candidate point ν, not yet in the subset, to the nearest subset point i^* and,
k is the number of instances already included in the subset.

$$\Delta^2_{k+1} = \max\{\Delta^2_\nu(k)\}, \nu \neq i^* \tag{2.5}$$

where Δ^2_{k+1} is the distance of the candidate instance that will be included in the subset.

In the present Thesis, the Kennard-Stone partition method was applied using the function written by Michal Daszykowski in MATLAB®[2] [68], the `Enalos+ Kennard and Stone` node in KNIME [69] which is based on the function of Michal Daszykowski, the `kenStone` (`prospectr` package) function in R [70] and a custom-written function in Python.

2.3.3 Quantitative measures of goodness-of-fit and predictivity

According to the 4th OECD principle [54] the statistical model validation is necessary for the assessment of a model's performance. Appropriate "fitness" metrics are used to quantify the accuracy of the model (approximation of the prediction to its actual value), and avoid under-fitting or over-fitting situations.

The produced model after training is applied in both training and test sets and computes the endpoint estimations. The training set is used to investigate the goodness-of-fit of the model, and the test set is used to assess its predictivity (predictive ability on new data). [54]

In the next paragraphs the statistical metrics for the assessment of performance are presented for both regression and classification models. As the present project deals with nanotoxicity-related datasets, the following functions refer to ENM samples.

[2]This function is based on a variation of the original Kennard and Stone algorithm: the two initially selected instances are the "central" one and its most dissimilar one in the data space. The rest of the instances are selected as in the original methodology.

2.3.3.1 Regression models

The quality-of-fit between the predicted and experimental values of the training set is expressed by the squared Pearson correlation coefficient R^2 (Eq. 2.6). R^2 values closer to 1, correspond to fitter models.

$$R^2 = \left(\frac{\sum_{i=1}^{N} (y_i - \bar{y})(\widehat{y_i} - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2 \left(\widehat{y_i} - \bar{\hat{y}}\right)^2}} \right)^2 \tag{2.6}$$

where, y_i and $\hat{y}_i$ are the experimental and predicted endpoint values over the N training samples,
$\bar{y}$ and $\bar{\hat{y}}$ are the averages of the experimental and predicted values respectively.

The mean absolute error (MAE, Eq. 2.7), the mean squared error (MSE, Eq. 2.8) and the root mean squared error (RMSE, Eq. 2.9) index can be computed on both the training and test sets to assess the validity and accuracy of the produced models. All these indexes together, provide a complete and thorough validation of the prediction accuracy, independently of the distribution level of the training-test endpoint values. [71]

$$MAE = \frac{1}{N} \sum_{i=1}^{N} | \, (y_i - \hat{y}_i) \, | \tag{2.7}$$

where, N is the number of ENMs in the training set and,
y_i and $\hat{y}_i$ are the actual and predicted endpoint values for the ith ENMs.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.8}$$

where, N is the number of ENMs in the training set and,
y_i and $\hat{y}_i$ are the actual and predicted endpoint values for the ith ENMs.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{2.9}$$

where, N is the number of ENMs in the training set and,
y_i and $\hat{y}_i$ are the actual and predicted endpoint values for the ith ENMs.

The external explained variance metric, (Q^2_{ext}, Eq. 2.10), which compares the predictions for the test ENMs with the observed endpoint values, is used to quantify the credibility of predictions on new data. [54]

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y_{tr}})^2} \tag{2.10}$$

where, y_i and $\hat{y}_i$ are the experimental and predicted endpoints of the N test samples and,
$\bar{y_{tr}}$ is the averaged value of the experimental endpoints of the training set.

In the LOO method, the explained variance in prediction (Q^2_{LOO}, Eq. 2.11) is calculated (instead of Q^2_{ext}). [54]

$$Q^2_{\text{LOO}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_{i/i}})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.11}$$

where, y_i is the experimental endpoint value of the ith sample,
$\hat{y_{i/i}}$ is the predicted endpoint estimated by using a model obtained without using the ith sample of the N samples and,
$\bar{y}$ is the averaged value of the experimental endpoints of the samples.

Test of the predictive ability of QSAR models Golbraikh and Tropsha [63] proposed a set of statistical indices to assess the predictive power of regression models (Eqs. 2.13 to 2.16), including the coefficient of multiple determination (R^2_{pred} - Eq. 2.12) and the external explained variance metric (Q^2_{ext}, Eq. 2.10). [54]

$$R^2_{\text{pred}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.12}$$

$$k = \frac{\sum_{i=1}^{N} y_i \widehat{y_i}}{\sum_{i=1}^{N} \widehat{y_i}^2} \tag{2.13}$$

$$k' = \frac{\sum_{i=1}^{N} y_i \widehat{y_i}}{\sum_{i=1}^{N} y_i^2} \tag{2.14}$$

$$R^2_0 = 1 - \frac{\sum_{i=1}^{N}(\widehat{y_i} - y_i^{r0})^2}{\sum_{i=1}^{N}(\widehat{y_i} - \bar{\widehat{y}})^2} \tag{2.15}$$

$$R'^2_0 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y_i}^{r0})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.16}$$

where N, is the number of ENMs that constitute the test dataset,
y_i, is the experimental (observed) endpoint value for the ith test ENM,
$\bar{y}$, is the average value of the endpoint in the test set,
$\widehat{y_i}$, is the predicted endpoint value for the ith test ENM,
$\bar{\widehat{y}}$, is the average over all $\widehat{y_i}$ ($i = 1, \ldots, N$), $y_i^{r0} = k \cdot \widehat{y_i}$ and,
$\widehat{y_i}^{r0} = k' \cdot y_i$.

Tropsha *et al*. [64] considered that a regression model is predictive if the following conditions (Eqs. 2.17 to 2.20) are satisfied:

$$R^2_{\text{pred}} > 0.6 \tag{2.17}$$

$$Q^2_{\text{ext}} > 0.5 \tag{2.18}$$

$$\frac{R^2_{\text{pred}} - R^2_0}{R^2_{\text{pred}}} < 0.1 \text{ or } \frac{R^2_{\text{pred}} - R'^2_0}{R^2_{\text{pred}}} < 0.1 \tag{2.19}$$

$$0.85 < k < 1.15 \text{ or } 0.85 < k' < 1.15 \tag{2.20}$$

2.3.3.2 Classification models

The most common representation of a model's classification results is the "confusion matrix" (Table 2.2). The rows in a confusion matrix represent the actual classes and the columns the predicted classes thus, the main diagonal represents the correct classification cases whereas the non-diagonal elements represent misclassifications. [54]

Table 2.2: Example of a confusion matrix for a case of two classes.

		Predicted class	
		TRUE	FALSE
Actual class	TRUE	TP	FN
	FALSE	FP	TN

where, TP (true positive) is the frequency of class TRUE ENMs correctly classified as "TRUE",
TN (true negative) is the frequency of class FALSE ENMs correctly classified as "FALSE",
FP (false positive - Type I error) is the frequency of class FALSE ENMs incorrectly classified as "TRUE" and,
FN (false negative - Type II error) is the frequency of class TRUE ENMs correctly classified as "FALSE".

Additional metrics which are used for assessing the performance of classification models are the following:

Sensitivity Sensitivity (Sn, Eq. 2.21) -also referred to as TP rate- expresses the ability of the model to distinguish the TRUE-class ENMs and correctly assign them to the "TRUE" class.

$$Sn = \frac{TP}{TP + FN} \tag{2.21}$$

Specificity Specificity (Sp, Eq. 2.22) -also referred to as TN rate- expresses the ability of the model to distinguish the FALSE-class ENMs and correctly assign them to the "FALSE" class.

$$Sp = \frac{TN}{TN + FP} \tag{2.22}$$

Accuracy Accuracy (Ac, Eq. 2.23) expresses the overall ability of the model to correctly assign the ENMs in their actual class. For a case of two classes the statistics -in order to consider a model reliable- should be greater than 0.5 which is the possibility for a random guessing prediction.

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.23}$$

Matthews correlation coefficient Especially for two-class (binary) problems, Matthews correlation coefficient (MCC, Eq. 2.24), is considered an adequate and reliable statistic even if the two classes are imbalanced (in terms of their size). [72] It expresses the correlation between actual and predicted values and takes values between [-1,1]; 1 demonstrates a

perfect agreement between predictions and actual observations, -1 a complete disagreement and 0 the lack of correlation between predictions and observations.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.24}$$

2.3.4 Response permutation

An additional robustness check of a proposed QSAR methodology and to eliminate the possibility of chance correlation between the independent variables (descriptors) and the response variable (endpoint), is the response permutation testing or Y-scrambling. In this technique, all -and the exact same- modelling steps are repeated using the original values of the independent variables, but scrambled sequence of values for the endpoint and. The quality of the predictions is again assessed using the R^2 and Q^2_{ext} metrics in case of regression, or Ac, Sn, Sp metrics in case of classification. If the new models have statistically lower predictive power compared to the model built with the original endpoints sequence, the initial model is considered reliable, because the possibility of random correlation is eliminated. [54], [73]

2.4 [t] Domain of applicability

Predictive models are trained on a subset of the space of input data (descriptors, physicochemical properties etc.) and can be used with confidence for predictions only if the untested data fall into the same subset, which is named applicability domain (APD). Predictions for substances located outside the APD are computed by extrapolating the model thus, are prone to be unreliable. The exact definition of the applicability domain is one of the five OECD computational modelling validation principles.

There are several strategies for the determination of the APD characterised as "range-based" (such as convex hull, bounding box, principal component analysis bounding box) and "geometric" methods (such as *k*NN, decision trees, probability density-based methods). [74] In the preset Thesis, similarity measurements based on the Euclidean distance among all training and test samples are used to define the APD of the proposed models when needed (see Chapter 7). The distance of a test sample to its nearest neighbour in the training set is compared to a predefined APD threshold (Eq. 2.25). In the case where this distance for a test sample exceeds the APD limit, its prediction is considered unreliable. [56]

$$APD = \langle d \rangle + Z\sigma \tag{2.25}$$

where $\langle d \rangle$ is the average of all distances included in the subset of distances which are lower than the mean value,
σ is the standard deviation of all distances included in the subset of distances that are lower than the mean value and,
Z is an empirical cut-off value (in most cases equal to 0.5). [75]

All the above steps can be employed using different software, programming languages and informatics platforms, that are presented in the next Chapters and in Appendices A′ and B′.

Chapter 3

Case studies

In this chapter we briefly present the datasets on which the proposed read-across methodologies will be applied. The datasets comprise gold and silver ENMs, various metal-oxide (MeOx) cored ENMs and MWCNTs. These data were derived from the Literature and in some cases were enriched with theoretical descriptors, calculated by appropriate software tools. The datasets consist of experimentally measured properties (e.g. physicochemical and biological descriptors, toxicity-related behaviour or biological activity), "periodic table descriptors" and calculated descriptors (quantum-mechanical, theoretical, image descriptors).

3.1 Gold ENMs cell association dataset

In their publication, Walkey *et al*. [76] presented an extensive characterisation of the protein corona in blood of a diverse set of gold ENMs. The library consists of 105 gold ENMs of different diameters (15, 30 and 60 nm) which were coated with surface ligands classified as neutral, anionic and cationic. A series of instrumental analysis experiments were used to measure their physicochemical characteristics; 40 descriptors were directly or indirectly generated from this analysis.

The ENMs were incubated with undiluted human serum, which is a proxy for the biomolecular environment of both *in vitro* cell culture experiments and the "biological" conditions when intravenous administration of ENMs is performed. The created protein corona (Figure 3.1) was characterized qualitative and quantitatively using poly(acrylamide) gel electrophoresis and high-resolution label-free shotgun tandem mass spectrometry, respectively. The cell association (in mL/μg(Mg)) of the ENMs with A549 human lung epithelial carcinoma cells was measured and used as an endpoint in predictive models. The ENM-cell association -especially when studied in relevance to protein corona formation- can be considered an important initiating event leading to disperse biological interactions such as inflammatory responses, biodistribution and toxicity. [76]–[78] A pseudopartition coefficient has been calculated (Eq. 3.1) considering the magnesium content of the cells.

For the model development only the anionic and cationic ENMs were used (84 samples), because neutral ENMs resisted serum protein adsorption, and 129 proteins of the protein corona fingerprint (PCF). The cell association data were transformed to $\log_2$ values prior to modelling (*net.cell*).

$$y = \frac{m_{\text{cell}}/m_{\text{well}}}{m_{\text{cells}}} \tag{3.1}$$

where, m_{cell} is the total atomic gold/silver content associated with cells,
m_{well} is atomic gold/silver content in the well where the cells were plated and,
m_{cells} is the total mass of magnesium per sample.

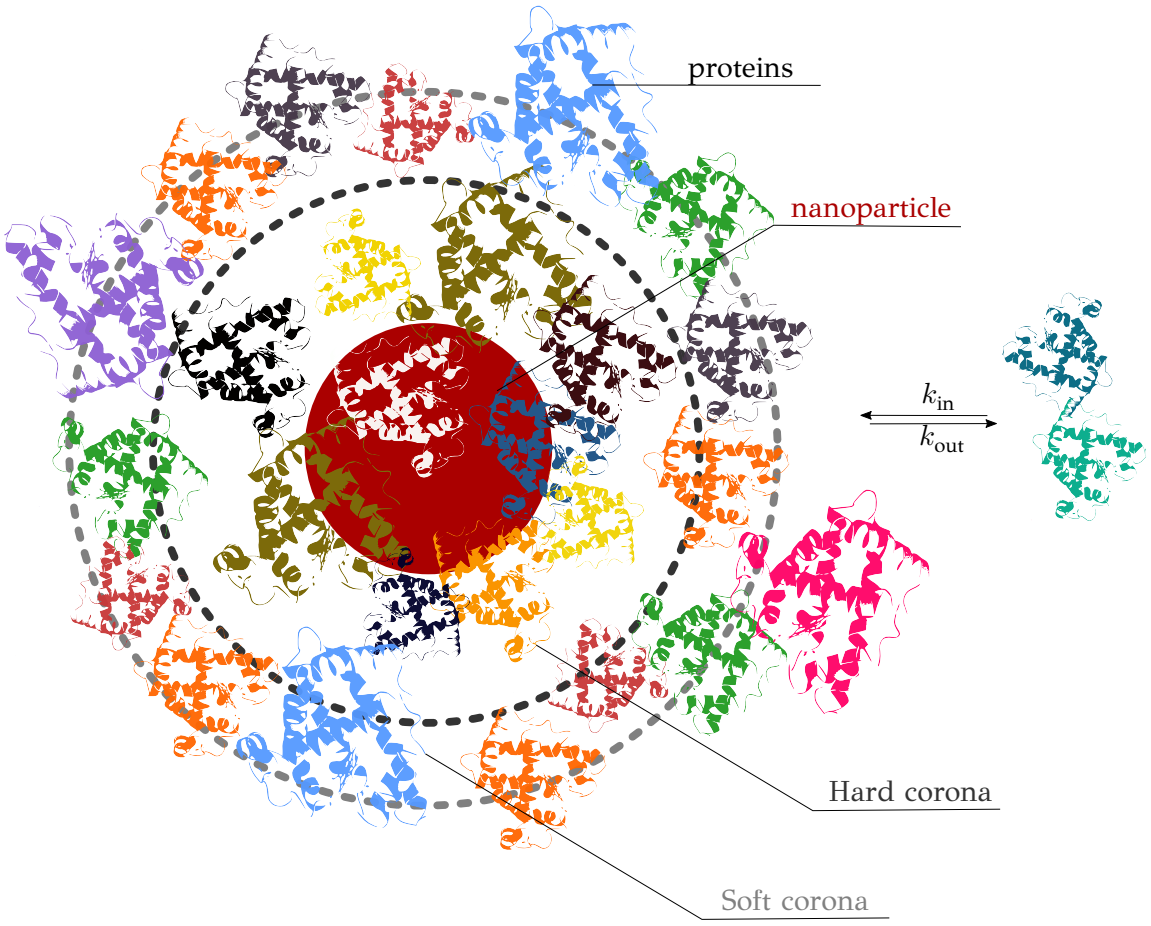


Figure 3.1: A schematic qualitative representation of the ENM-protein corona complex. ENMs due to their high surface free energy adsorb biomolecules when inserted in biological media. The ENM acquires a new biological identity that determines further interaction of the ENM with other molecules. This corona is subject to the size and shape of the ENM, its surface charge and solubility as well as the route of administration of the ENM to the organism. The protein corona is a dynamic bilayer consisting of the "hard" corona (tightly bound proteins and/or other biomolecules) and the "soft" corona (biomolecules not directly bound to the ENM). [79], [80]

In our analysis, from the available physicochemical descriptors *zp.serum.sign* was removed from the set as it would not offer any additional information to the models (same value for all samples). In addition, the PCF was filtered by applying the Gene Set Variation Analysis (GSVA) method [81] in toxFlow application (see §A′.2.1). From the original set only 63 proteins were included in the statistically significant gene sets, and these proteins, along with the physicochemical descriptors, were used in the developed methodologies. [45] The availability of two different types of descriptors (Tables Γ′.1 and Γ′.2), rendered this dataset suitable for testing the proposed methodologies with one or two similarity criteria (see §1.3.1).

3.2 Metal (hydr)oxide ENMs cytotoxicity dataset

Forest *et al*. [82] present a nanoQSAR approach in order to create simple and reliable predictive models for ENMs experimental prioritization especially effective in case of small datasets. For this reason, 25 MeOx and metal hydroxide ENMs were synthesized and extensively characterized under ISO guidelines. The ENMs have different chemical composition from six metal (hydr)oxide families (SiO_2, TiO_2, CeO_2, AlOOH, ZnO, $Ni(OH)_2$) with different particle

dimensions and shapes (spheres or rods).

In total 12 descriptors (Table Γ′.3) are included in this dataset belonging in two categories: dimensional (nano) and composition-related (non-nano) descriptors. The dimensional descriptors decode information about ENMs size (minimum, d_{min} and maximum d_{max} diameter), solubility, shape (shape factor, SF and corrected shape factor, CSF), zeta-potential, specific surface area (SSA) and agglomeration state. The non-dimensional descriptors are related to the ENMs chemical composition including the hydration rate (to encode chemical dissimilarity between oxides and hydroxides), the oxidation degree of the metallic element, the radius of the metallic cation and the Pauling electronegativity of the metallic element.

The toxicity of the 25 ENMs was investigated on the murine cell line RAW 264.7: the lactate dehydrogenase (LDH) release was measured and was considered as an endpoint relevant to cytotoxicity, especially for binary classification of ENMs to "toxic" and "non-toxic". For this reason, an LDH score cut-off value of 1.5 was settled on to categorize ENMs as "cytototoxic" (LDH score above 1.5) or "non-cytotoxic/safe" (LDH score bellow 1.5). The publication of Forest *et al.* also includes the toxicity indexes of the TNF-α production (signalling a pro-inflammatory reaction) and ROS production (indicative of an oxidative stress) that can be used as endpoints in predictive nanoinformatics modelling.

3.3 Metal oxide ENMs cytotoxicity classification dataset

In their studies Zhang *et al*. [83] and Liu *et al.* [84] presented a dataset of MeOx ENMs, which were tested in a multi-parametric series of toxicity experiments and studied for nanoQSAR modelling. The original dataset consists of 24 MeOx ENMs with a detailed multi-perspective toxicity profile, from seven different assays for two different cells; human bronchial epithelial (BEAS-2B) and murine myeloid (RAW 264.7) cell lines. The toxicity studies included a single-parameter cytotoxic assay, where cell viability was measured, and a multi-parameter toxicity assay for oxidative stress responses assessment of cells. [83] The toxicity profile of the ENMs was summarized to a toxicity class ("toxic"/"non-toxic") based on dose-response analysis and consensus Self-Organizing Map clustering. From the original dataset the F_3O_4 sample was excluded due to high rate of impurities.

For our work, 24 descriptors were selected for modelling purposes (Table Γ′.4) including size descriptors (the actual ENMs diameter and the hydrodynamic diameter in different media (water, Bronchial Epithelial Cell Growth medium - BEGM and Dulbecco Modified Eagle's medium - DMEM)), surface charge descriptors (the zeta-potential at pH=7.4 and the isoelectic point), fundamental MeOx descriptors (the number of metal and oxygen atoms, the atomic mass of metals and the molecular weight, the metal electronegativity and the ionic index), MeOx energy descriptors (atomization energy and sublimation energy, standard molar enthalpy of formation, lattice energy, ionization energy and first molar ionization energy) and ENMs energy descriptors (conduction band and valence band energies, chemical potential, hardness, electrophilicity and electronegativity of MeOx).

3.4 Multi-walled carbon nanotubes surface adsorption dataset

Xia *et al.* [85] studied and modelled the surface adsorption properties of a 40 nm diameter MWCNT, coated with hydroxyl derivatives. Adsorption energy is responsible for the interaction of the nanomaterial with various biological molecules, including proteins, and can contribute to the interpretation of its behaviour in different biological media. Five nanodescriptors representing the surface adsorption interactions (lipophilicity, hydrogen bond acidity and basicity, polarity/polarizability, and lone-pair electrons, also presented in Table Γ′.5) were used to train a model for predicting the adsorption coefficient (k) of this particular MWCNT.

The dataset contained adsorption coefficient (k) values for 28 probe compounds with various physicochemical properties. The k values were converted to the logarithmic scale.

3.5 Functionalized multi-walled carbon nanotubes toxicity dataset

Zhou *et al.* [86] presented a library of 83 surface-modified MWCNTs, with an identical core of a controlled size distribution (diameter of 40 $\pm$ 10 nm and length of 250 $\pm$ 120 nm). Combinational chemistry modifications were performed, by covalently attaching copies of different molecules to the surface of the MWCNTs, whereas the size and the shape of each nanotube remained intact (Figure 3.2).

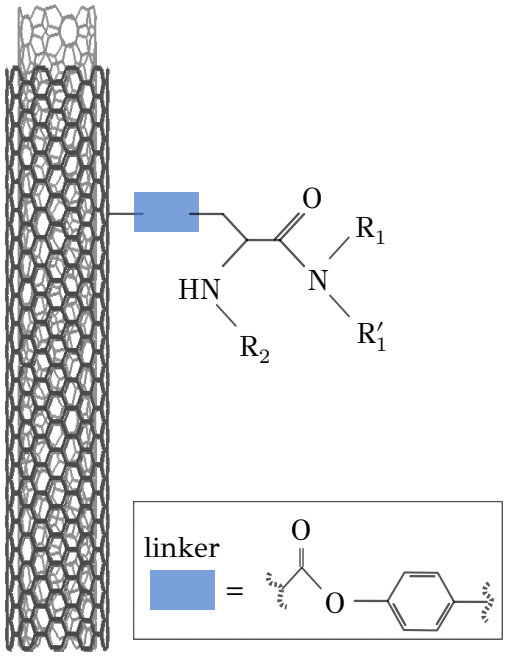


Figure 3.2: The core MWCNT structure (*MWCNTs* dataset) along with the organic modifier and its substituents position. The substituents R_1/R'_1 and R_2 account for eight amines and nine acylators (and one without acylation) respectively, and with their combination, the library of the decorated MWCNTs is created. [86]

As the all studied samples had an identical core and further Transmission Electron Microscopy (TEM) characterisation demonstrated that there has been no change of the general MWCNT structure throughout the functionalization route, [86] a reasonable assumption [87] was made that the differences in their biological behaviour were mostly due to the structural characteristics of their surface ligands. This hypothesis can be considered realistic, especially taking into consideration the near- and long-term assessment goals, and the time and resources requirements for the complete characterisation -experimental and/or computational- of all available nanostructures. This hypothesis has already been used in different studies found in the Literature. [56], [88] Since the surface modification differentiated the MWCNTs, each sample was represented by its ligands. Mold2 software was used in order to calculate the necessary descriptors. This software calculates a large and diverse set of molecular descriptors for each decorator encoding two-dimensional chemical structure information. Mold2 descriptors were calculated using KNIME and `Enalos+` nodes. [89], [90]

In total 777 descriptors per MWCNT-decorated molecule were generated accounting for the topological, geometric and structural characteristics of the organic modifiers. In general, an important step in the modelling procedure is the reduction of the original pool of descriptors before the feature selection, in order to increase the model quality. [91] Thus, the descriptors containing the same values at a percentage equal or higher than 20% among the samples were excluded from further analysis using the `Enalos+/Remove column` node.

The MWCNTs were experimentally tested in six *in vitro* assays including MWCNT binding of the proteins bovine serum albumin (BSA), carbonic anhydrase (CA), chymotrypsin (CT), and haemoglobin (HB), as well as acute toxicity and immune toxicity properties. In our analysis (§7.1) we followed the division of the data into categories, as proposed by Fourches *et al*. [87] The CA binding affinity values varied from 0.53 to 5.29 at a MWCNT concentration of 15 mg/mL, thus a separation cut-off limit of 2.0 was chosen, in order to produce two classes of balanced distribution; in total, 44 MWCNTs were assigned as "binders" (CA protein binding activity greater than 2.0) and 39 as "non-binders" (CA protein binding activity less than 2.0). Similarly, for the toxicity endpoint the cellular survival percentage measured experimentally ranged between 2% and 68% at the high MWCNT concentration of 200 mg/mL. MWCNTs with cell survival values lower than 37% were labelled as "toxic" (38 samples), whereas samples with cell survival values greater than 42% were labelled as "non-toxic" (35 samples). The MWCNTs around the median cell survival range (37–43%) were not included in the refined modelling set, as it was difficult to define a clear threshold for the division of the two classes. [87]

3.6 NanoMILE zeta-potential dataset

37 ENMs of different cores (pure metal/MeOx), coating compositions (uncoated/anionic/cationic/neutral coatings) and different shapes (circular for spheres, rods, and plates), were selected and characterized extensively in the EC-funded project NanoMILE. [3], [92] An important aspect in ENMs toxicity assessment, is the study of their extrinsic properties, such as the agglomeration of the ENMs under certain conditions, taking into consideration that a large agglomerate of ENMs may dissociate or break up in a cellular environment and later release smaller (and potentially more bioavailable) particles in the body. [93], [94] The agglomeration phenomena are greatly affected by the surface charge of ENMs, as encoded by the zeta-potential index; high zeta-potential values either negative or positive, produce stable ENM suspensions, whereas ENMs with low zeta-potential values tend to form agglomerates in the absence of either steric stabilization resulting from polymer coatings or association of biomolecules with the ENMs via the formation of an acquired corona. [95] Thus the zeta-potential index is a critical factor in ENMs characterisation and the study or prediction of their toxicity. [93], [96], [97] For the aforementioned set of ENMs the experimental values for the zeta-potential descriptor were used in the predictive procedure as an endpoint.[1]

Besides the various physicochemical characterisation properties, the dataset contained 68 TEM images of the ENMs. The captured TEM images can be an invaluable source of information, from the *in silico* point of view, because a range of image descriptors can be extracted with the use of appropriate software tools. The extraction of image nanodescriptors can be a complex process due to the wide variety of microscopy images with varying resolution, mixed sizes and shapes, as well as the agglomeration and aggregation of the ENMs depicted within a TEM image, which can either be a drying artifact or indicative of the presence of agglomerates/aggregates in the sample prior to deposition on the TEM grid. [100]

[1] It is noteworthy that the meaningfulness of zeta-potential measurements on non-spherical particles is debatable, given that the underpinning mathematical assumptions assume spherical particles. [98], [99] Being able to predict zeta-potentials for non-spherical ENMs from TEM images would thus improve the meaningfulness of this datapoint for prediction of ENMs behaviour.

The NanoXtract image analysis web application was employed in our work for the analysis the TEM images. In particular, 18 image descriptors (Table Γ΄.6) were extracted from each image accounting for the geometrical and morphological characteristics of the ENMs. Each image was treated separately in NanoXtract, according to the shape of the depicted ENMs (Figure 3.3) and by tuning a series of parameters including noise reduction, thresholding etc. until satisfactory segmentation was achieved.

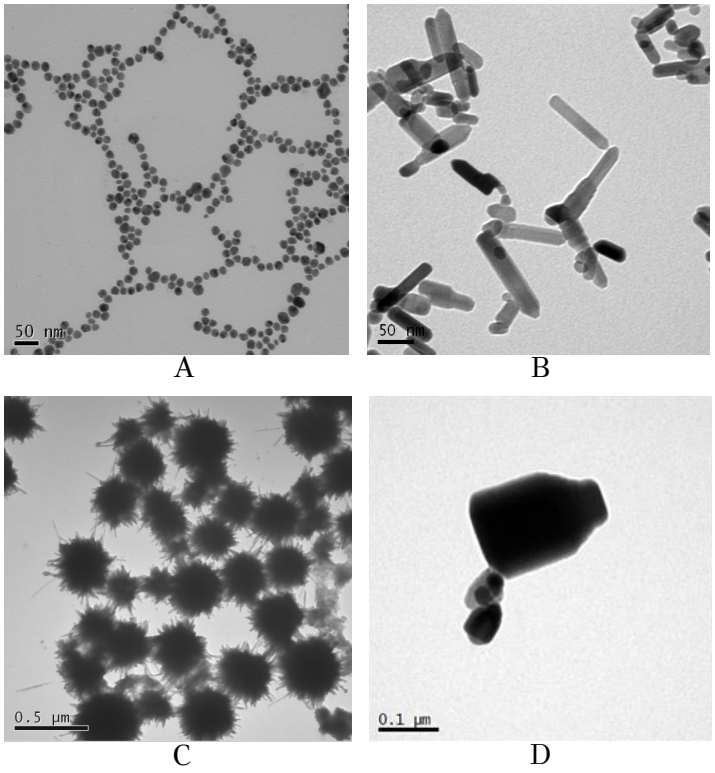


Figure 3.3: Representative examples from the *NanoMILE ENMs* dataset, including three different shape types of ENMs. [A] Sample-"AgPURE" which are primarily spheres. [B] Sample-"prom ZnO" which are nanorods. [C] Sample-"Aged PROM_CeO2_PO43" which are sea-urchin shaped ENMs. [D] Sample-"ZnO 110" which are plates.

Only two additional descriptors from full physicochemical characterisation dataset were considered, namely the information about the core of the ENMs (pure metal or MeOx) and the pH where the zeta-potential was measured (as pH affects the degree of charge neutralisation). More details about this dataset are presented in Table Γ΄.7.

3.7 Super-paramagnetic iron oxide ENMs cell viability classification dataset

In the publication of Kotzabasaki *et al.* [101] a data collection of physicochemical descriptors of super-paramagnetic iron oxide nanoparticles (SPIONs) along with a toxicity profile, is presented. Data were collected and curated after an extensive Literature research and the final dataset consists of 16 coated SPION samples, with 6 known physicochemical descriptors (the SPIONs magnetic core, their zeta-potential, their size, the field strength, the concentration of iron per cell and the relaxivity) and additional measurements of their cell viability.

From the available descriptors, zeta-potential was excluded from modelling activities due to high rate of missing values. The used descriptors are presented in Table Γ΄.8. The magnetic core was encoded in a binary variable using 1 for "maghemite" and 0 for "magnetite". From the SPIONs cell viability, a categorical toxicity endpoint was derived; a cut-off value (75%) of

cell viability was defined and samples with cell viability value less than the threshold value, were characterized as "toxic", otherwise were characterized as "non-toxic". Cell viability is a commonly evaluated biological endpoint in *in vitro* and *in silico* nanotoxicology and is usually the percentage of live cells in comparison to a control sample, after exposure to ENMs. [101], [102]

In our analysis (see §4.3.6.4) we additionally removed the duplicate SPIONs from the dataset in order to ensure non-biased models. Samples with identifications 6 and 7 (both iron oxide-loaded cationic nanovesicles) differentiate in their type of coating, however the coating is not used as a training descriptor, due to the diversity of coatings that does not permit a binary encoding and, as a descriptor, has no predictive power.

After the removal of sample with ID 7, the missing relaxivity values were calculated following the agglomerative hierarchical clustering steps performed in the aforementioned study, including pre-scaling of data between 0 and 1, using the `scipy.cluster.hierarchy` Python package (v1.5.1). [103] Three SPIONs clusters were defined (same as in the original study -apart from sample with ID 7), and the relaxivity values that were missing (samples with IDs 0, 9 and 12) were calculated from the average relaxivity value of the rest of the samples belonging to the same cluster.

3.8 Chapter summary

This Chapter presents the datasets that will be used in the rest of the Thesis as case studies for testing and validating the proposed read-across methods in terms of their accuracy and performance. Table 3.1 summarizes the important information about the datasets: number of samples, number and types of descriptors, type of endpoint (numerical or categorical) and the references where the datasets have been presented. The first column contains short names, for quick and easy reference to the datasets in the results sections of the next chapters.

Table 3.1: Summary of processed case studies. In the first column the short name of each dataset is presented.

Dataset	Samples	Descriptors	Endpoint(s)	Endpoint type	Reference
Gold ENMs	84 gold ENMs	40 physicochemical descriptors & 63 PCF	$\log_2$ cell association	numerical	Walkey *et al*. (2014) [76], Varsou *et al*. (2017) [45]
MeOx ENMs [a]	25 metal (hydr)oxide ENMs	12 physicochemical descriptors	LDH release, TNF-α production, ROS production, & cytotoxicity (based on LHD release values)	categorical/ numerical	Forest *et al*. (2019) [82]
MeOx ENMs [b]	23 MeOx ENMs	8 physicochemical descriptors, 6 fundamental descriptors, 10 quantum-mechanical descriptors	toxicity (summarized multi-perspective profile)	categorical	Zhang *et al*. (2012) [83], Liu *et al*. (2013) [84]
MWCNTs [a]	28 probe compounds	5 descriptors (encoding surface interactions)	adsorption coefficient ($\log k$)	numerical	Xia *et al*. (2011) [85]
MWCNTs [b]	83 surface modified MWCNTs	777 mold2 descriptors	CA protein binding affinity & acute toxicity (cell survival)	categorical	Zhou *et al*. (2008) [86]
NanoMILE ENMs	37 diverse-core ENMs	18 image descriptors, solution pH & ENMs core	zeta-potential	numerical	Varsou *et al*. (2020) [3]
SPIONs	15 SPIONs	5 physicochemical descriptors	toxicity (cell viability)	categorical	Kotzabasaki *et al*. (2020) [101]

Chapter 4

A mathematical programming strategy for the development of read-across models

In this Chapter a novel read-across methodology for the estimation of toxicity-related endpoints of ENMs is presented, according to the thresholding strategy (see also §1.3.1). The proposed methodology defines neighbours of ENMs based on similarities -in the multidimensional ENM space - on single or multiple levels- and generates endpoint predictions for untested ENMs using only information from their neighbourhood. The proposed method automates and optimises the read-across framework proposed by ECHA, which has been presented in the Introduction section. Via the formulation and solution of a mathematical optimisation problem, the method searches over the space of alternative solutions (hypotheses) and converges to one providing the most accurate read-across predictions. In this search for the optimal grouping hypothesis, the proposed methodology considers as pivotal optimisation parameters the variable selection and the boundaries that define the neighbourhood of the query ENM, for which a read-across prediction is sought. Another advantage of the proposed methodology is that it considers the multi-perspective characterisation of ENMs by grouping ENM descriptors into categories and by using multiple similarity criteria for defining neighbours to the target ENM. The different categories may include physicochemical, biological, quantum-mechanical, image, biokinetics descriptors etc.

The methodology is based on the formulation of a rigorous mixed integer non-linear mathematical optimisation problem (MINLP), which is solved, using a specifically designed Genetic Algorithm (GA). The methodology was extended by including a regularisation parameter in the constructions of the objective function (OF). The methodology was also modified and adapted to the case of categorical endpoints. In this Chapter we first present the formulation of the MINLP problem, then the solution strategy using the GA and finally the two extensions concerning the regularisation term and the prediction of categorical endpoints. The theoretical background is provided whenever necessary. The presented methodology has been the core of the web application named Apellis, which is also presented at the end of the Chapter.

4.1 [t] Mathematical optimisation

Mathematical programming is one of the most important decision-making methods in the field of Operations Research. Mathematical programming deals with the optimal distribution of limited resources b_i which result from a series of m constraints imposed by the studied problem and the objective function (OF, Eq. 4.1). The OF, which expresses the function of the x_j decision variables of the problem (n decision variables), needs to be minimised

or maximised, subject to the constraints. Constraints (Eq. 4.2) are written as equalities or inequalities that express the distribution of resources in the various activities.

When the mathematical representation of the problem consists exclusively of linear functions with respect to unknown variables which are continuous, then it is a Linear Programming (LP) problem. LP is a useful tool for dealing with complex decision problems. Decisions depend entirely on the accuracy of the problem description and the adequacy of the model. A typical form of an LP model is presented below. [104]

Objective function

$$\min \ or \max \ Z = \sum_{j=1}^{n} c_j x_j \tag{4.1}$$

subject to

$$\sum_{j=1}^{n} a_{ij} x_j \ \leq \ b_i \qquad (i = 1, 2, \dots, m) \qquad x_j \leq 0 \qquad \forall j = 1, 2 \dots n \tag{4.2}$$

In this relationship the coefficients c_j participate in the OF in a linear fashion. Similarly, decision variables participate linearly with a technological factor a_{ij} to the formulation of the constraints.

In the case of linear functions and integer variables, an Integer Linear Programming (ILP) problem is formulated, while if both continuous and integer decision variables are involved, a Mixed Integer Linear Programming (MILP) problem is constructed. If the mathematical representation of the problem consists of non-linear functions, then it is a Non-linear Programming (NLP) problem.

4.2 [m] Development of the MINLP problem

As explained before, the development of a robust and reliable read-across workflow for the prediction of ENM undesired properties had a twofold objective: First, the reduction of the available dataset, by removing the variables that add noise rather than useful information to the analysis. Second, the definition of the neighbour boundaries (in the form of one or more thresholds) which indicate the source ENMs that are considered similar to the target ENM.

These two different goals can be achieved simultaneously through the development of a MINLP problem, where the objective is to minimize the mean squared error (MSE) between the experimental endpoint values and the produced predictions of a set of ENMs with respect to selecting the most informative descriptors and defining the neighbour boundaries. The problem is explained thoroughly in the next paragraphs.

4.2.1 [m] One similarity measure

4.2.1.1 Available data

The methodology assumes the availability of a dataset containing the values of L descriptors and the endpoint for N_{tr} ENMs. The data are first scaled using a standardisation (e.g. Gaussian normalisation) or a normalisation (e.g. min-max) method, to ensure that scaled descriptors contribute equally to the overall prediction analysis. [6] The dataset is denoted by $\{\mathbf{x}_i, y_i\}$, $i = 1, \dots, N_{\text{tr}}$, where $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \dots x_{i,L}\}$, is a vector containing the values of the L descriptors and y_i is the endpoint value of the ith ENM.

4.2.1.2 Set of variables

The main results of the solution of the MINLP problem are the following:

- $attr_\ell$: a binary variable indicating if the descriptor ℓ is selected or not, $\ell = 1, \dots, L$.
- thr: a continuous variable that defines a threshold for the selection of neighbour ENMs. Only if the Euclidean distance between two ENMs is equal or less than thr, these two ENMs are considered as neighbours.

A number of additional variables are used for the construction of the MINLP problem:

- $dist_{i,j}$: a continuous variable containing the Euclidean distance between ENMs i and j, $i = 1, \dots, N_{\text{tr}}$, $j = 1, \dots, N_{\text{tr}}$.
- $neib_{i,j}$: a binary variable taking the value of 1 if ENMs i and j are neighbours and 0 if they are not, $i = 1, \dots, N_{\text{tr}}$, $j = 1, \dots, N_{\text{tr}}$.
- $pred_i$: a binary variable taking the value of 1, if ENM i has at least one neighbour and 0 if it has no neighbours, $i = 1, \dots, N_{\text{tr}}$.
- y_i: a continuous variable containing the experimental (actual) endpoint value for the ith ENM, $i = 1, \dots, N_{\text{tr}}$.
- $\hat{y}_i$: a continuous variable containing the predicted read-across endpoint value for the ith ENM, $i = 1, \dots, N_{\text{tr}}$.
- $predFactor$: a user-defined percentage, indicating the balance level between the predictive accuracy and the number of ENMs with a prediction.

4.2.1.3 Mathematical formulation

The mathematical formulation of the optimisation problem, consists of a set of well-defined constraints that should be satisfied by the solution of the problem and the OF to be minimized.

Eq. 4.3 computes the Euclidean distance between all pairs of ENMs considering only the selected descriptors:

$$dist_{i,j} = \sqrt{\sum_{\ell=1}^{L} attr_\ell \left(x_{i,\ell} - x_{j,\ell}\right)^2} \quad i = 1, \dots, N_{\text{tr}}, \quad j = 1, \dots, N_{\text{tr}}, \quad i \neq j \tag{4.3}$$

The following set of equations ensures that two ENMs i and j are considered as neighbours only if their Euclidean distance $dist_{i,j}$ is equal or lower than the threshold. In this case the corresponding binary variable $neib_{i,j}$ takes the value of 1, otherwise the value of 0 is assigned to this variable. In Eqs. 4.4, 4.5, m is a very small positive real number (equal to 10^{-3})[1]:

$$neib_{i,j} \geq \ m \cdot \left(thr - dist_{i,j}\right), \ \forall \ i,j \in \{1, \dots, N_{\text{tr}}\}, \ i \neq \ j \tag{4.4}$$

$$1 - neib_{i,j} \geq -m \cdot \left(thr - dist_{i,j}\right), \ \forall \ i,j \in \{1, \dots, N_{\text{tr}}\}, \ i \neq \ j \tag{4.5}$$

$$neib_{i,i} = 0, \ \forall \ i \in \{1, \dots, N_{\text{tr}}\} \tag{4.6}$$

[1]bigM reformulation: In this case the "sufficiently large" or "small" positive number ensures the equality of the variables participating in the corresponding constraints (usually displayed in pairs), only if a particular binary variable gets a certain value, but leaves the variables "open" if the binary variable has the opposite value.

Eq. 4.7 computes the read-across predictions as weighted averages of the endpoint values of neighbour ENMs:

$$\hat{y}_i = \frac{\sum_{j=1}^{N_{\text{tr}}} y_j \cdot \frac{neib_{i,j}}{1+dist_{i,j}}}{\sum_{j=1}^{N_{\text{tr}}} \frac{neib_{i,j}}{1+dist_{i,j}}}, \forall\ i \in \{1, \dots, N_{\text{tr}}\} \tag{4.7}$$

For ENMs without any neighbour, read-across predictions are not possible. An additional set of constraints (Eqs. 4.8, 4.9, 4.10) guarantees that the % percentage of ENMs with at least one neighbour, is greater than or equal to a predefined percentage denoted by $predFactor$. In these equations, $pred_i$ is a binary variable that becomes equal to 1, when a read-across prediction is achieved for the ith ENM, and 0, if no prediction is possible:

$$\sum_{i=1}^{N_{\text{tr}}} pred_i \geq\ predFactor \cdot\ N_{\text{tr}} \tag{4.8}$$

$$pred_i \geq\ neib_{i,j}, \forall\ i \in \{1, \dots, N_{\text{tr}}\}, \forall\ j \in \{1, \dots, N_{\text{tr}}\} \tag{4.9}$$

$$pred_i \leq \sum_{j=1}^{N_{\text{tr}}} neib_{i,j}, \forall\ i \in \{1, \dots, N_{\text{tr}}\} \tag{4.10}$$

Objective function The OF to be minimized (Eq. 4.11) is the MSE between the endpoint read-across predictions and the actual endpoint values over all the ENMs with at least one neighbour.

$$\min \frac{1}{\sum_{i=1}^{N_{\text{tr}}} pred_i} \sum_{i=1}^{N_{\text{tr}}} pred_i \left(y_i - \hat{y}_i\right)^2 \tag{4.11}$$

In order to inspect the influence of the total number of selected variables on the accuracy of the produced models, we included an adjustment of the OF with the addition of a regularisation parameter. The addition of the regularisation term was inspired by the penalization of the coefficient estimates in LASSO and ridge regression in order to avoid the risk of over-fitting (see also page 131).

The altered OF (Eq. 4.12) has two components the MSE term and a regularisation term, accompanied by a user-defined regularisation factor wf_{OF}, that controls the influence of the number of selected variables on the final score. Higher wf_{OF} values lead to the selection of fewer variables and this reduces the complexity of the produced model.

$$\min \frac{1}{\sum_{i=1}^{N_{\text{tr}}} pred_i} \sum_{i=1}^{N_{\text{tr}}} pred_i \left(y_i - \hat{y}_i\right)^2 + wf_{\text{OF}} \cdot \sum_{\ell=1}^{L} attr_\ell \tag{4.12}$$

4.2.2 [m] Extension of the MINLP problem to multiple similarity criteria

Due the complex structure of ENMs, different types of data and descriptors are often used for ENM characterisation. In the study of Varsou *et al*, [45] it has been demonstrated how two or more similarity criteria can be used, for defining thresholds and for selecting the neighbours, if different types of characterisation data are available. Our extended read-across approach follows the same concept and computes optimal threshold values for different types of characterisation data. Two ENMs are considered as neighbours if both distances are lower than the corresponding thresholds.

The MINLP formulation described before is extended in this subsection to account for multiple similarity criteria. For brevity and for simplified notation, the extended formulation is presented for two similarity criteria. Inclusion of additional criteria is trivial.

4.2.2.1 Available data

The descriptors are grouped into sets A and B containing L_A and L_B descriptors respectively. The dataset is presented to the algorithm in the form $\{\mathbf{xA}_i, \mathbf{xB}_i, y_i\}$, $i = 1, \dots, N_{\text{tr}}$, where $\mathbf{xA}_i = \{xA_{i,1}, xA_{i,2}, \dots, xA_{i,L_A},\}$, and $\mathbf{xB}_i = \{xB_{i,1}, xB_{i,2}, \dots, \ xB_{i,L_B}\}$, $i = 1, \dots, N_{\text{tr}}$.

4.2.2.2 Set of variables

The main outcomes of the MINLP problem are:

- $attrA_\ell$: a binary variable indicating if the descriptor ℓ in group A is selected or not, $\ell = 1, \dots, L_A$.
- $attrB_\ell$: a binary variable indicating if the descriptor ℓ in group B is selected or not, $\ell = 1, \dots, L_B$.
- thr_A, thr_B: two continuous variables defining threshold for the selection on neighbouring ENMs for the two similarity criteria. Only if both Euclidean distance between two ENMs are equal or less than the respective thresholds, these two ENMs are considered as neighbours.

The following additional variables are used for the construction of the MINLP problem:

- $distA_{i,j}$,$distB_{i,j}$: two continuous variables containing the Euclidean distance between ENMs i and j for the two similarity criteria, $i = 1, \dots, N_{\text{tr}}$, $j = 1, \dots, N_{\text{tr}}$.
- $neibA_{i,j}$,$neibB_{i,j}$: two binary variables taking the value of 1 if ENMs i and j are neighbours with respect to similarity criteria A or B and 0 if they are not, $i = 1, \dots, N_{\text{tr}}$, $j = 1, \dots, N_{\text{tr}}$.
- $neib_{i,j}$: a binary variable taking the value of 1 if ENMs i and j are neighbours and 0 if they are not, $i = 1, \dots, N_{\text{tr}}$, $j = 1, \dots, N_{\text{tr}}$.
- $pred_i$: a binary variable taking the value of 1, if ENM i has at least one neighbour and 0 if it has no neighbours, $i = 1, \dots, N_{\text{tr}}$.
- y_i: a continuous variable containing the experimental (actual) endpoint value for the ith ENM, $i = 1, \dots, N_{\text{tr}}$.
- $\hat{y}_i$: a continuous variable containing the predicted read-across endpoint value for the ith ENM, $i = 1, \dots, N_{\text{tr}}$.
- $predFactor$: a user-defined percentage, indicating the balance level between the predictive accuracy and the number of ENMs with a prediction.

4.2.2.3 Mathematical formulation

The set of constraints is similar to the formulation in §4.2.1. The next equations (4.13, 4.14) compute the Euclidean distances between all pairs of ENMs taking into account only the selected descriptors for groups A and B.

$$distA_{i,j} = \sqrt{\sum_{\ell=1}^{L_A} attrA_\ell \left(xA_{i,\ell} - xA_{j,\ell}\right)^2}, i = 1, \dots, N_{\text{tr}}, \quad j = 1, \dots, N_{\text{tr}}, \quad i \neq j \tag{4.13}$$

$$distB_{i,j} = \sqrt{\sum_{\ell=1}^{L_B} attrB_\ell \left(xB_{i,\ell} - xB_{j,\ell}\right)^2}, i = 1, \ldots, N_{\text{tr}}, \quad j = 1, \ldots, N_{\text{tr}}, \quad i \neq j \tag{4.14}$$

The following set of equations ensure that two ENMs are considered as neighbours with respect to the different similarity criteria only if the Euclidean distances are lower than the respective threshold. In this case the corresponding binary variable takes the value of 1, otherwise the value of 0 is assigned to this variable. In Eqs. 4.15, 4.16, 4.18 and 4.19 m, is a very small positive real number:

$$neibA_{i,j} \geq \ m \cdot \left(thr_A - distA_{i,j}\right), \ \forall \ i,j \in \{1, \ldots, N_{\text{tr}}\}, \ i \neq \ j \tag{4.15}$$

$$1 - neibA_{i,j} \geq -m \cdot \left(thr_A - distA_{i,j}\right), \ \forall \ i,j \in \{1, \ldots, N_{\text{tr}}\}, \ i \neq \ j \tag{4.16}$$

$$neibA_{i,i} = 0, \ \forall \ i \in \{1, \ldots, N_{\text{tr}}\} \tag{4.17}$$

$$neibB_{i,j} \geq \ m \cdot \left(thr_B - distB_{i,j}\right), \ \forall \ i,j \in \{1, \ldots, N_{\text{tr}}\}, \ i \neq \ j \tag{4.18}$$

$$1 - neibB_{i,j} \geq -m \cdot \left(thr_B - distB_{i,j}\right), \ \forall \ i,j \in \{1, \ldots, N_{\text{tr}}\}, \ i \neq \ j \tag{4.19}$$

$$neibB_{i,i} = 0, \ \forall \ i \in \{1, \ldots, N_{\text{tr}}\} \tag{4.20}$$

The set of Eqs. 4.21-4.23 define two ENMs i and j are neighbours if they satisfy both similarity criteria, i.e. only if both $neibA_{i,j}$ and $neibB_{i,j}$ are equal to 1.

$$neib_{i,j} \geq \ neibA_{i,j} + neibB_{i,j} - 1, \ \forall \ i,j \in \{1, \ldots, N_{\text{tr}}\} \tag{4.21}$$

$$neib_{i,j} \leq \ neibA_{i,j}, \ \forall \ i,j \in \{1, \ldots, N_{\text{tr}}\} \tag{4.22}$$

$$neib_{i,j} \leq \ neibB_{i,j}, \ \forall \ i,j \in \{1, \ldots, N_{\text{tr}}\} \tag{4.23}$$

Eq. 4.24 computes the read-across predictions as weighted averages of the endpoint values of neighbour ENMs by selecting one distance metric (here we assume the metric based on group A):

$$\hat{y}_i = \frac{\sum_{j=1}^{N_{\text{tr}}} y_j \cdot \frac{neib_{i,j}}{1+distA_{i,j}}}{\sum_{j=1}^{N_{\text{tr}}} \frac{neib_{i,j}}{1+distA_{i,j}}}, \forall \ i \in \{1, \ldots, N_{\text{tr}}\} \tag{4.24}$$

Constraints 4.8, 4.9, 4.10 are used again to guarantee that the % percent of ENMs with at least one neighbour is greater than or equal to a predefined percentage denoted by $predFactor$.

Objective function The OF is the same as in the previous MINLP formulation (Eq. 4.11).

4.3 [m] Solution strategy: an evolutionary algorithm

The MINLP problems described above cannot be solved efficiently by conventional optimisation methods. For the solution of the problem, we developed a tailor-made novel evolutionary algorithm based on the concept of Genetic Algorithms (GAs) which is described in detail in this section. GAs have been used successfully for the variable subset selection in different optimisation problems. [105] Before presenting the details of our GA workflow, we are including a small theoretical overview for the GAs.

4.3.1 [t] Genetic algorithms

GAs are one of the most popular categories of evolutionary optimisation strategies and are inspired by the natural selection principles. The basic genetic algorithm was invented by Holland [106] and his associates during the 60s and 70s, based on the concept that living organisms are examples of successful optimisation trough a natural or artificial selection. A detailed comparison of the evolution terminology in both natural systems and optimisation strategies can be found in Table 4.1.

By mimicking the biological processes, a GA is using an initial *population* (usually randomly selected) of candidate solutions, encoded in *chromosome* sequences, and a corresponding degree of *fitness* that summarizes their "quality". In continuation, the most *fitted chromosomes* survive and produce offspring according to the operational parameters (genetic operators) of *selection*, *crossover* and *mutation*, during a cycle of iterations (*generations*). In that way the candidate solutions are evolving, converging to a near-optimal solution in the search space. These biological procedures are applied to the potential solutions with a degree of probability, thus GAs are stochastic optimisation methods.

The main genetic operators can be summarized below: [104], [107], [108]

Selection Similarly to natural selection process, where the fittest individuals have more than average chances to survive, reproduce and transfer their genetic information to the next generations, the *selection* operator points out the *chromosomes* of the *population* for reproduction. Better performing (or *fitter*) *chromosomes* regarding their ability to solve the problem on call, are more likely to be selected for reproduction than poor performing ones. A practical -but not common- method to *select chromosomes* for *reproduction* is to keep the 50% of the best-performing ones (top mate selection). A more common *selection* method is the "roulette wheel sampling" which is a *fitness*–balanced method: to each *chromosome* is given a proportional to its *fitness* "wedge" in a circular roulette-wheel; when the roulette-wheel is turned around, the "ball" stops randomly on one of the slices of the wheel and the corresponding *chromosome* is *selected*. The well-performing *chromosomes* are more likely to be *selected*.

Crossover This operator is approximately mimicking the biological recombination between two haploid organisms where genetic information is inherited to the offspring. In this case a random *locus* is selected and the two *mate chromosomes* exchange their *genes* before and after the *locus* in order to produce two *child chromosomes*.

Mutation Adopting the idea of mutation from natural systems, where genetic information can be inherited susceptible to random errors, this operator randomly inverts the values of some *genes* in a *chromosome*. The *mutation* operator can be applied on every bit position, however with a very small probability.

Table 4.1: Comparison of evolution terminology in natural systems and mathematical optimisation. [104], [109]–[111]

Concept	Natural/Social system	Genetic algorithms
gene	A specific DNA sequence (transcribed as a single unit) that encodes a distinct hereditary characteristic and corresponds to a single or a group of related proteins, or to a single or a group of related RNA molecules.	Single bits or specific sequences of bits that encode the natural parameters of the optimisation problem.
chromosome	A structure consisting of a DNA molecule and a set of related proteins that conveys part of (or all) the inherited information of an organism.	A bit (*gene*) sequence of finite length that encodes a potential solution of the optimisation problem.
fitness	The ability of organisms to survive and reproduce in the conditions where they live into and thus, contribute their genes to the future generations.	The quantification of the optimality of a potential solution as encoded by a *chromosome*. The *fitness* function is usually based on the OF of the mathematical optimisation problem.
individual	A single organism capable of independent existence. (A member of a compound organism or colony can be also considered as an individual unit).	A pair of a *chromosome* and its corresponding *fitness* value. *Individuals* are subject to genetic operators in order to perform an optimisation process.
population	The individuals of one species or all the organisms of a same group that live in a given area and interbreed.	A set of *individuals* in a *generation*.
parents	The direct ancestor of a living species (human, animal or plant).	*Population* members of the current *generation* (set of potential solutions).
mates	Pairs of animal sexual partners.	Pairs of *parents* selected to produce offspring and pass on their genetic material, based on their *fitness*.
child	An offspring.	*Population* members of the next *generation* (new set of potential solutions).

(*continued from Table 4.1*)

Concept	Natural/Social system	Genetic algorithms
selection	An evolutionary process that results in the reproduction between the individuals of a population. In this process more offsprings are produced than the ones that survive thus, "competitive" conditions exist between the members of the population. The individuals that are more "fit" are able to pass in greater proportion their inheritable characteristics to the future generations.	The process of picking two appropriate *parent chromosomes* based on their calculated *fitness* scores, for reproduction (*crossover*).
crossover	The recombination of chromosomes during sexual reproduction by exchanging genetic materials between two homologous chromosomes non-sister chromatids.	The combination of two potential solutions, by exchanging complementary set of *genes* between two *mate chromosomes* from the *parent generation*. Two *child chromosomes* are produced from *crossover*.
mutation	Heritable alteration in the sequence of nucleotides in a chromosome.	The alteration of the crossovered *chromosome*, by inverting the values of the *genes* (in case of binary encoding) or by applying a more sophisticated *mutation* function (in case of non-binary encoding). *Mutation* ensures the genetic diversity of the child populations.
locus	A specific position on a chromosome indicating a gene or a genetic marker.	The point of *crossover*.
elitism	"The belief that some things are only for a few people who have special qualities or abilities".	Forcing the most-fit *individual* to be included in the *child generations* until a better one is found. This heuristic addition ensures desirable evolution.
generations	A period of time between the child phase and the parent phase of an individual in a population of organisms.	A number of cycles of *selection*, *crossover* and *mutation* transforming a *parent population* to a *child population*, converging to an optimal solution.
genome	The DNA that conveys all the hereditary information of a cell or an organism.	The solution in which the optimisation problem converges after a cycle of *generations*.

The most appropriate choice of optimisation parameters in a GA scheme is subject to the type of the problem. The *population* size, the number of *generations*, the number of *genes* and the *mutation* frequency along with the time needed to evaluate its *chromosome's fitness*, define the total execution time of the workflow. Therefore, these result to a counterbalance between large and varied *populations* that explore extensively the solution space and limited *populations* that explore for longer time periods. [104]

4.3.2 [m] Development of a GA workflow

The GA scheme assumes that a dataset with known descriptors and toxicity endpoint values is available. For validation purposes, the dataset is partitioned into training and test sets. The training set is used in the optimisation process, whereas the test set is used to assess the performance of the final model.

A subset of the available descriptors along with the threshold value, encoded in a hybrid array of binary and real values (*genes*), defines precisely a potential solution of the optimisation problem (*chromosome*). Different combinations of variables and thresholds are encoded in different *chromosomes* that comprise a *population*. The *population* evolves through a number of cycles (*generations*) of "biological" operations (*selection*, *crossover* and *mutation*) between the potential solutions, leading to an optimal solution (*genome*). All the biological operations are controlled by user-defined probability values.

Each *chromosome* in every level of the evolutionary process is tested for read-across prediction and, depending on the accuracy and robustness of the generated predictions, a score number is assigned to it. When all *chromosomes* of a *population* are scored, the ones with the highest scores are selected in pairs and are combined in order to exchange *genes* in random *crossover* points. The two new *chromosomes* are subject to *mutation*, where the *gene* values are altered according to a uniform or non-uniform scenario. The above process of *selection*, *crossover* and *mutation* is repeated until a new *population* is created. The process ensures that if a *chromosome* of the old *population* has higher score than all the *chromosomes* of the new *population*, the chromosome with the highest score is included in the new *population* (*elitism*). It also ensures that at least one variable will be selected, and that the combination of variables and threshold(s) will produce predictions for at least a predefined number of training samples (*predFactor*).

The particular GA developed in this work uses the user-defined parameters depicted in Table 4.2 and is composed of three main steps including the creation of an initial population, its evaluation and the natural selection process that converges to the optimal solution. The algorithm is schematically described in Figure 4.1. The steps presented in the next paragraphs refer mainly to the training process.

Table 4.2: Initial user-defined parameters for the developed *genetic algorithms* scheme.

Initial parameter	Details
nChrom	The size of the *population*, total number of *chromosomes* per *generation*
maxGenerations	The total number of *generations*
initGeneProb	The probability for a *gene* to have value 1 initially
crossProb	The probability of *crossover*
mutProb	The probability for *mutation* of each *gene* (uniform)
nonUnf	The *mutation* probability of the threshold(s) (non-uniform)
$thr_{\min}^{GA}$	Lower bound of the threshold(s) value
$thr_{\max}^{GA}$	Upper bound of the threshold(s) value
bGA	Freezing parameter
predFactor	Minimum number of samples with produced prediction

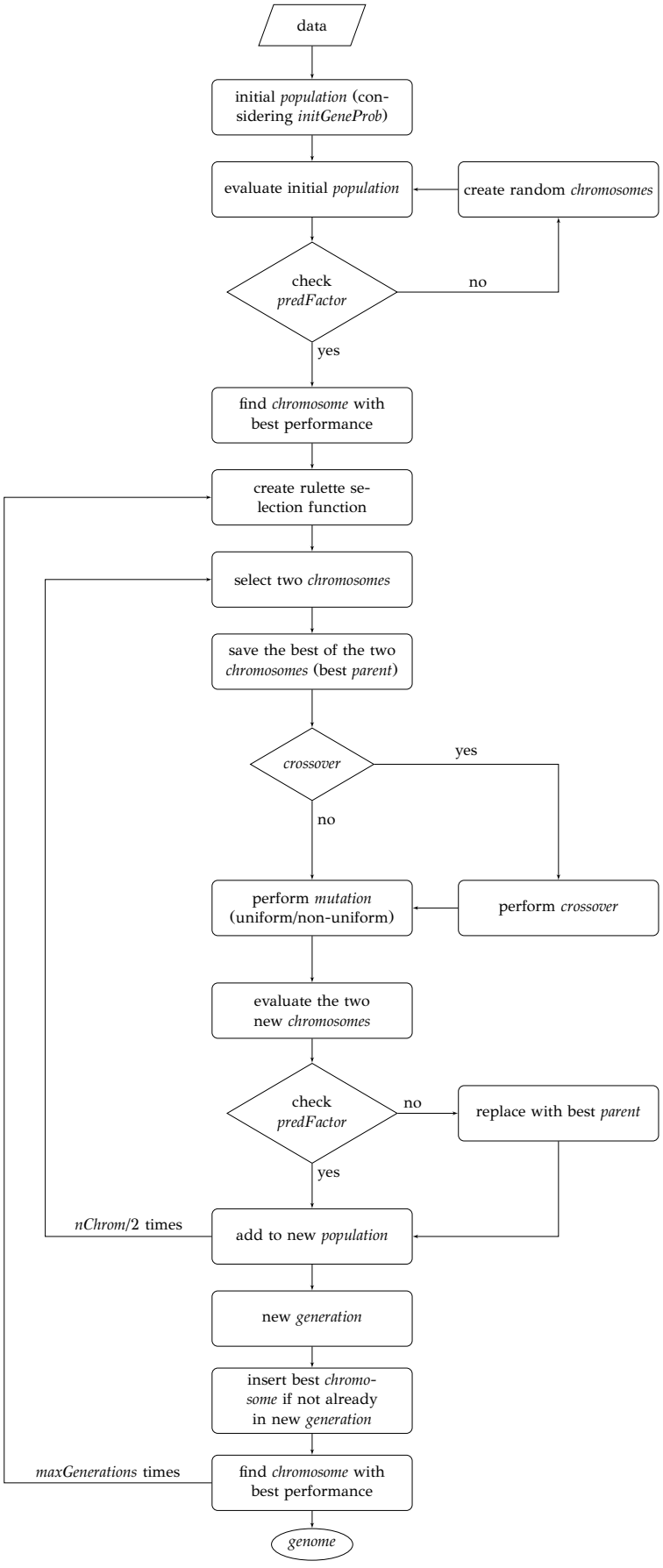


Figure 4.1: Schematic description of the proposed *genetic algorithm* workflow for the prediction of undesired ENMs properties.

Step I An initial *population* of *chromosomes* is created. The structure of the *chromosomes* is shown in Table 4.3. The *chromosome* is actually a vector, whose length is equal to the number of descriptors L plus the number of similarity criteria used for defining neighbours to a target ENM. Each descriptor and the threshold(s) are placed in specific positions in the *chromosome* representations. This creates hybrid *chromosomes* containing binary *genes* for descriptors and real *genes* for thresholds.

- The *genes* related to descriptors correspond to the $attr_\ell$ variables in the construction of the MINLP problem. A value of 1 means that the corresponding descriptor is selected for defining the distance matrix, while a value of 0 means that the descriptor has not been selected. The probability of a binary *gene* to be coded as 1 is denoted by *initGeneProb*.
- The real *genes* of the *chromosomes* contain the threshold values corresponding to the similarity criteria and their values are selected randomly from the distance matrices of all samples, considering all variables in each group. In case only one similarity criterion is used, the threshold is placed in the end of the *chromosome*, whereas if two criteria are used, the two thresholds are placed at the beginning and the end of the *chromosome* (Table 4.3).

Table 4.3: Examples of *chromosomes* with one and two thresholds.

1	0	0	1	0	...	1	1	2.718
2	1.772	1	0	0	...	0	1	1.618

Step II The fitness of each *chromosome* of the initial *population* (and later in every level of the evolutionary process) is calculated as follows:

- The Euclidean distances between all pairs of ENMs are computed using Eq. 4.3 for a single similarity criterion or Eqs. 4.13-4.14 for two similarity criteria.
- For each ENM, neighbour ENMs in the training set are identified as the ones whose distance from the reference ENM is equal or lower than the thr value (in case of two similarity criteria both distances should be equal or lower than the respective thresholds). During training it is ensured that an ENM will not be selected as neighbour of its own (Eqs. 4.6, 4.17, 4.20).
- The algorithm checks if Eq. 4.8 is satisfied, i.e. if ENMs with at least one neighbour are more than $predFactor$ multiplied by the total number of ENMs. If yes, the algorithm proceeds with next step. If not, the *chromosome* is rejected, and a new *chromosome* is generated as described in Step I. It is noted that a new *chromosome* is generated also in the case that no variables are selected.
- The read-across predictions are computed using the weighted average of the endpoints of the neighbouring ENMs in the training set -Eq. 4.25 (minor adaptations are needed for two or more similarity criteria) for ENMs with at least one neighbour. In case that no neighbours are found for a particular ENM ($pred_i = 0$), a prediction of its endpoint is not possible. A schematic representation of how the read-across prediction is computed is depicted in Figure 4.2.

$$\widehat{y}_i = \begin{cases} \frac{\sum_{j=1}^{N_{\text{tr}}} \frac{neib_{i,j}}{1+dist_{i,j}} \cdot y_j}{\sum_{j=1}^{N_{\text{tr}}} \frac{neib_{i,j}}{1+dist_{i,j}}}, \text{ if } pred_i \neq 0 \\ \\ NA, \text{ if } pred_i = 0 \end{cases} \qquad \forall i = 1, \ldots, N_{\text{tr}} \tag{4.25}$$

where, $\hat{y}_i$ is the predicted endpoint value for the ith ENM,
y_j is the actual endpoint value for the jth train ENM,
N_{tr} is the number of train ENMs,
$neib_{i,j}$ is a binary variable taking the value of 1 if ENMs i and j are neighbours and 0 if they are not and,
$dist_{i,j}$ is the Euclidean distance between ENMs i and j.

- Each *chromosome* is evaluated for its ability to lead to reliable predictions. This ability is encoded to a score (fitness value) that is used for *chromosomes* selection in the evolutionary process. To assess the reliability of the predictions, the mean squared error (MSE) over all ENMs with at least one neighbour is computed using Eq. 4.26 (similar to Eq. 2.8 adapted for ENMs with at least one neighbour).

$$MSE = \frac{1}{N_{\text{pred}}} \sum_{i=1}^{N_{\text{tr}}} pred_i \left(y_i - \hat{y}_i\right)^2 \tag{4.26}$$

where, N_{tr} is the number of ENMs in the training set,
y_i and $\hat{y}_i$ are the actual and predicted endpoint values for the ith ENMs,
$pred_i$ is a binary variable that becomes equal to 1, when a read-across prediction is achieved for the ith ENM, and 0, if no prediction is possible and,
N_{pred} is the total number of ENMs with a successful prediction, $N_{\text{pred}} = \sum_{i=1}^{N_{\text{tr}}} pred_i$.

The OF value of the *chromosome* is based on the MSE and the regularisation parameter, and it is computed by Eq. 4.27 (similar to Eq. 4.12). The fitness value of the *chromosome* is computed by just inverting the value of the OF (Eq. 4.28) (better predictions lead to higher scores). The constant term 10^{-5} in the denominator ensures that the $score$ value will not become infinite in the case of $OF = 0$.

$$OF = MSE + wf_{\text{OF}} \cdot \sum_{\ell=1}^{L} attr_\ell \tag{4.27}$$

where wf_{OF} is a positive user-defined regularisation parameter,
$attr_\ell$ is a binary variable indicating if the descriptor ℓ is selected and,
L is the total number of available descriptors.

$$score = 1/(OF + 10^{-5}) \tag{4.28}$$

- The *chromosome* with the highest (*best*) calculated fitness is saved for later analysis.

Step III A natural selection process takes place and it is iterated *maxGenerations* times. During each iteration, the following procedure is repeated *nChrom*/2 times and in total *nChrom* are selected that form the new *generation*.

- In order to assure the reproduction of the fittest *chromosomes*, a "roulette wheel" approach is used. The method selects a pair of *chromosomes* from the previous *population*, based on randomly generated numbers that indicate the "slots" corresponding to the different *chromosomes*[2]. The roulette wheel is constructed so that the size of each slot is proportional to the fitness of the corresponding *chromosome*. The roulette is "biased", thus *chromosomes* with a reproductive advantage (better fitness scores), have higher probability to be selected. For each pair of selected *chromosomes*, the one with the highest score is saved as the *bestParent* for later use.

[2]The "roulette wheel" function and the *crossover* operator are based on a previous UPCI team-work [112]

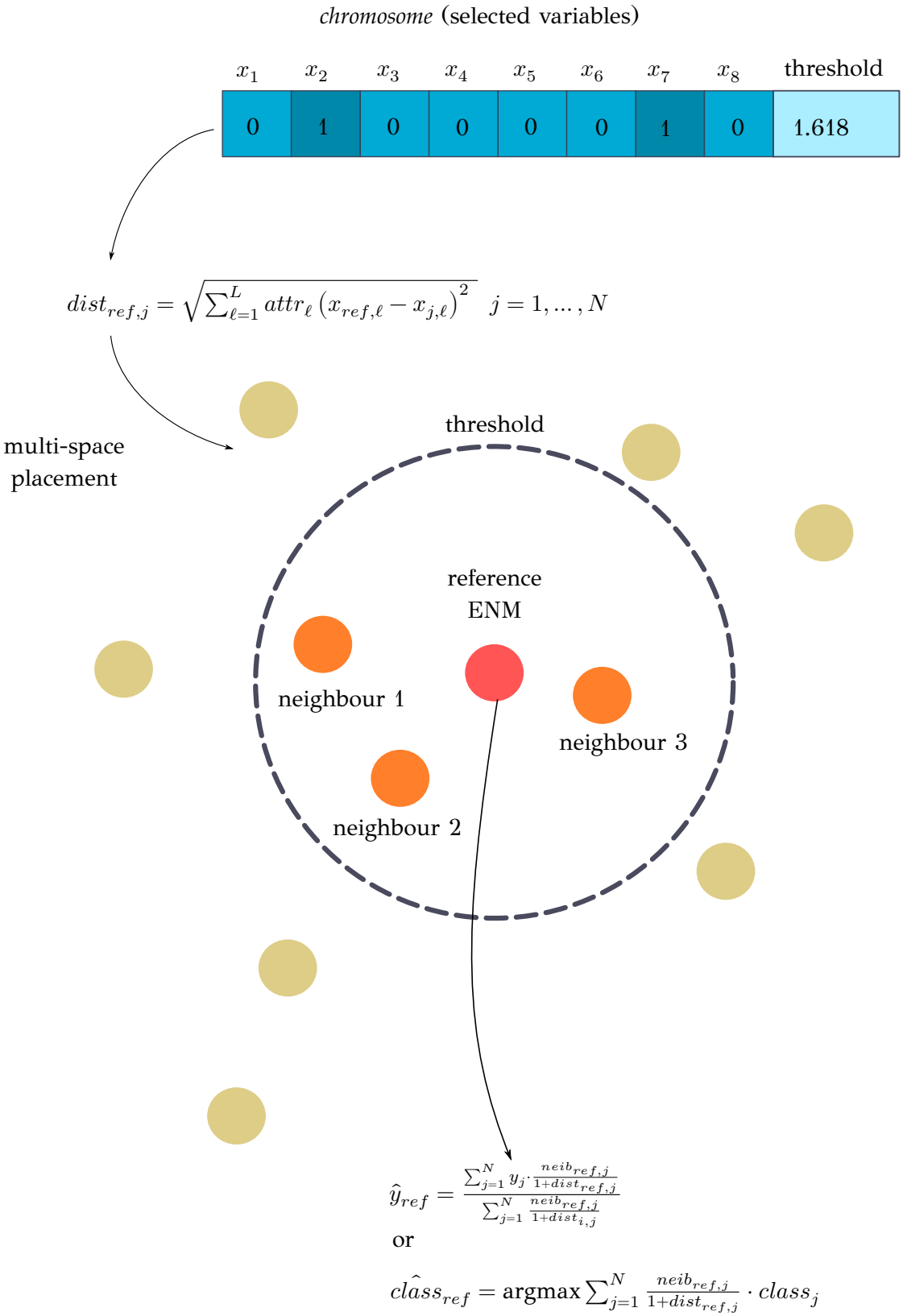


Figure 4.2: A schematic representation of the proposed read-across approach using the *genetic algorithms* optimisation scheme: the selected variables determine the neighbours in the multidimensional space and the optimal threshold value defines a circle around a reference ENM (red particle) and ENMs inside the circle are considered as neighbours (orange particles) whereas the rest ENMs (light yellow particles) do not belong to the reference ENM neighbourhood and are not involved in the read-across prediction.

- The genetic operators of *crossover* are applied. According to the *crossProb* value, it is decided if the *chromosomes* are going to exchange strings of *genes* or not, in a randomly selected point that indicates the position of *crossover*.
- The genetic operator of *mutation* is applied. With probability *mutProb*, binary *genes* that corresponds to a descriptor, invert their value from 0 to 1 and vice versa, while non-uniform *mutation* is always performed to the threshold values, according to Eq. 4.29.

$$thr_{\text{new}}^{\text{GA}} = \begin{cases} thr_{\text{old}}^{\text{GA}} + (thr_{\text{max}}^{\text{GA}} - thr_{\text{old}}^{\text{GA}}) \cdot (1 - r^{(1-g/maxGenerations)bGA}) \\ \text{if a random digit is 0} \\ \\ thr_{\text{old}}^{\text{GA}} - (thr_{\text{old}}^{\text{GA}} - thr_{\text{min}}^{\text{GA}}) \cdot (1 - r^{(1-g/maxGenerations)bGA}) \\ \text{if a random digit is 1} \end{cases} \tag{4.29}$$

where $thr_{\text{old}}^{\text{GA}}$ is the old threshold value,
$thr_{\text{new}}^{\text{GA}}$ the threshold value that results from the non-uniform *mutation*,
$thr_{\text{max}}^{\text{GA}}$ and $thr_{\text{min}}^{\text{GA}}$ are the upper and the lower bounds of the threshold values,
r is a random number between 0 and 1, g is the number of the current *generation* and,
bGA is a parameter which determines the degree of dependency on the *generation* number.

The non-uniform *mutation* process, searches the space uniformly in the first place avoiding stagnating, and as the number of iterations approximates the maximum number of *generations*, convergence is achieved. [105]

- The two new chromosomes are evaluated with the procedure described in Step II and in case a chromosome does not meet constraint 4.8, it is replaced by its *bestParent*. This replacement also takes place in case that after *crossover* and *mutation* all descriptor *genes* of the *child chromosome* have value of 0 (no variables are selected).

In case the best *chromosome* of the previous generation is not included in the new *generation*, the algorithm places it in the position of the *chromosome* with the minimum score, in order to ensure that the *chromosome* with the best performance will always survive in the evolutionary procedure.

The best chromosome of the last *generation* is the result of the algorithm (*genome*). The selected variables and threshold(s) corresponding to the *genome* will be used subsequently for read-across predictions of unknown ENMs. For evaluating the method, all the training examples are passed through Step II described above to produce the read-across predictions. The correlation coefficient among actual experimental values and read-across predictions (R^2, similar to Eq. 2.6 adapted for ENMs with at least one neighbour) is calculated as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{N_{\text{pred}}} (y_i - \bar{y})(\widehat{y_i} - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N_{\text{pred}}} (y_i - \bar{y})^2 \left(\widehat{y_i} - \bar{\hat{y}}\right)^2}} \right)^2 \tag{4.30}$$

where, y_i and $\hat{y}_i$ are the experimental and predicted endpoint values over the train set, $\bar{y}$ and $\bar{\hat{y}}$ are the averages over the experimental values and the read-across predictions respectively and,

N_{pred} is the number of training ENMs for which predictions are available (ENMs with at least one neighbour).

4.3.3 [m] Validation of the produced read-across models

An external validation approach is used to test the proposed read-across methodology, by dividing the full dataset into training and test subsets (see §2.3.2). This data partitioning can be achieved either by applying a random partition or a partition method (e.g Kennard-Stone).

The training set is used in the GA workflow described above and determines the optimal set of descriptors and threshold(s) values. For the test set, predictions are made using the workflow described in Step II of the algorithm, but now the selected descriptors, and the threshold(s) are fixed to their optimal values. Eventually, the read-across predictions are compared with the experimental endpoint values using the Q^2_{ext} statistic (Eq. 4.31, similar to Eq. 2.10 adapted for test ENMs with at least one neighbour), the MAE, (Eq. 4.32, similar to Eq. 2.7 adapted for test ENMs with at least one neighbour) and the MSE metric (Eq. 4.33), which is adapted to the test samples. [64]

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{N_{\text{pred}}} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{N_{\text{pred}}} (y_i - \bar{y}_{tr})^2} \tag{4.31}$$

$$MAE = \frac{1}{N_{\text{pred}}} \sum_{i=1}^{N_{\text{test}}} |pred_i \, (y_i - \ \hat{y}_i)| \tag{4.32}$$

$$MSE = \frac{1}{N_{\text{pred}}} \sum_{i=1}^{N_{\text{test}}} pred_i \, (y_i - \ \hat{y}_i)^2 \tag{4.33}$$

where, y_i and $\hat{y}_i$ are the actual and predicted endpoint values over the test set,
$\bar{y}_{tr}$ is the averaged value of the endpoint over the N_{tr} training ENMs and,
N_{test}, is the number of ENMs in the test set ,
N_{pred} is the number of test ENMs with $pred_i \neq 0$.

Finally, in order to eliminate the possibility of chance correlation between the descriptors and the endpoint, the Y-scrambling test (§2.3.4) is applied. Due to the stochastic nature of the proposed read-across strategy, it is not possible to fully repeat the exact same modelling steps. Still, we consider that the produced results from the response permutation test provide important conclusions regarding the reliability of the produced models.

4.3.4 [m] Use of a read-across model to predict the endpoint values of untested ENMs

Fully validated read-across models can be used for the endpoint estimation of untested ENMs, when the *genome*-indicated descriptors are known values. The endpoint estimation process is similar to the training-validation process: for each untested ENM its neighbours are located in the previously used training set, according to the Euclidean distances and the optimised threshold(s). The endpoint estimation is performed according to the endpoint values of the selected neighbours (Eqs. 4.9 or 4.10 depending on the similarity criteria used in training). If no neighbours are identified for an untested ENM, this ENM is considered to be located outside the domain of applicability of the model and a read-across prediction is not possible.

4.3.5 [m] Implementation

The implementation of the GA workflow was performed in the MATLAB® programming language (§A′.1.1). The source code is available at GitHub (`https://github.com/DemetraDanae/optimized-read-across`) considering a single threshold (extension to two or more criteria is trivial) released under GNU General Public License v.3. Minor modifications are needed to make the code compatible with GNU Octave and these are marked as comments. The interested users can also comment in/out the commands concerning the dataset partitioning and external validation. In the same repository, users can find the files of the dataset [45], [76] used in the following case study.

The source code of the GA workflow with the regularisation is available at GitHub (`https://github.com/DemetraDanae/optimized-read-across/tree/master/regularisation`) considering a single threshold. The difference between the two algorithms is the insertion of wf_{OF} in the OF thus, for $wf_{\mathrm{OF}} = 0$ the code will run with implicit variable selection.

4.3.6 [r] Results and discussion

The proposed read-across method is demonstrated on the *Gold ENMs* dataset which includes 84 gold anionic and cationic ENMs, 40 physicochemical descriptors (PDs) and 63 statistically significant proteins (biological descriptors, BDs). The values of each descriptor were scaled in the range [0,1] (Eq. 2.1), in order to be comparable and contribute impartially to the read-across predictions. The availability of two different types of descriptors, renders this dataset suitable for testing the proposed method with one or two similarity criteria.

The developed methodology was validated both internally using the entire dataset for training and predictions assessment (for reasons of comparison with our previous work [45]), and externally using a training and a test subset. In the internal validation the regularisation is not taken into account.

4.3.6.1 [r] Internal validation

The GA method was applied on the entire dataset with the operational parameters shown in Table 4.4. Due to the stochastic nature of the proposed GA strategy, different runs of the algorithm may produce different output results, even if the starting conditions are exactly the same. We selected three levels of the $predFactor$, and we executed the complete workflow 10 times in the following three variations of the method:

- Considering a single threshold, corresponding to the full set of descriptors.
- Assuming two different thresholds, one for the group of PDs and one for the group of BDs and obtaining the read-across predictions using the distances between PDs.
- Assuming two different thresholds, one for the group of PDs and one for the group of BDs and obtaining the read-across predictions using the distances between BDs.

Figures 4.3, 4.4 present in a sorted manner the R^2 values produced by individual runs of the GA workflow using one threshold and two thresholds respectively. The results are summarized in Table 4.5[3]. As expected, by increasing the value of the $predFactor$ parameter, the optimal threshold values determined by the GA are larger (Figure 4.5), which means that read-across predictions are obtained for more ENMs, because there are more ENMs having at least one neighbour (Figure 4.6). On the other hand, the accuracy of the read-across predictions measured by the R^2 statistic is decreased because additional ENMs with

[3]Summarized results of 10 runs of the GA workflow are depicted. It is noted that the minimum and maximum values for the different result values are not necessarily produced at the same run.

Table 4.4: Values for the user-defined operational parameters of the proposed read-across GA workflow applied on the *Gold ENMs* dataset.

Parameter	Value
nChrom	100
maxGenerations	1000
initGeneProb	0.6
crossProb	0.7
mutProb	0.01
nonUnf	0.1
$thr_{\min}^{\mathrm{GA}}$	0.1
$thr_{\max}^{\mathrm{GA}}$	mean value of the maximum distances between samples
bGA	1
predFactor	0.3-0.6-0.9
wf_{OF}	0

higher distances are considered as neighbours to the reference ENM and are involved in the calculation of the read-across prediction. An illustrative example is presented in Figure 4.7. By comparing the results between using one or two thresholds, we do not observe significant differences on the number of ENMs with read-across predictions, on the number of selected variables, or on the accuracy of the predictions expressed by R^2 statistic. The results obtained by using the PD and BD distances for computing the read-across predictions, are almost identical.

The prediction accuracy of the proposed method, using the 60% $predFactor$ level, is similar to the application of toxFlow [45] on the same dataset, in terms of the R^2 statistic (toxFlow produced a 0.973 R^2 value). However, the method proposed in this work was able to produce read-across predictions for significantly more ENMs (average 50 to 51 ENMs compared to 21 ENMs in toxFlow).

For the 60% $predFactor$ level, we also measured the frequency of appearance of the different descriptors in the selected sets of descriptors. It is clear that there exist descriptors which are selected in most runs, whereas some other descriptors are chosen very rarely. The descriptors appearing in more than 70% of the runs are considered as the most significant descriptors. The most frequently selected PDs and BDs are presented in Figures 4.8, 4.9 respectively.

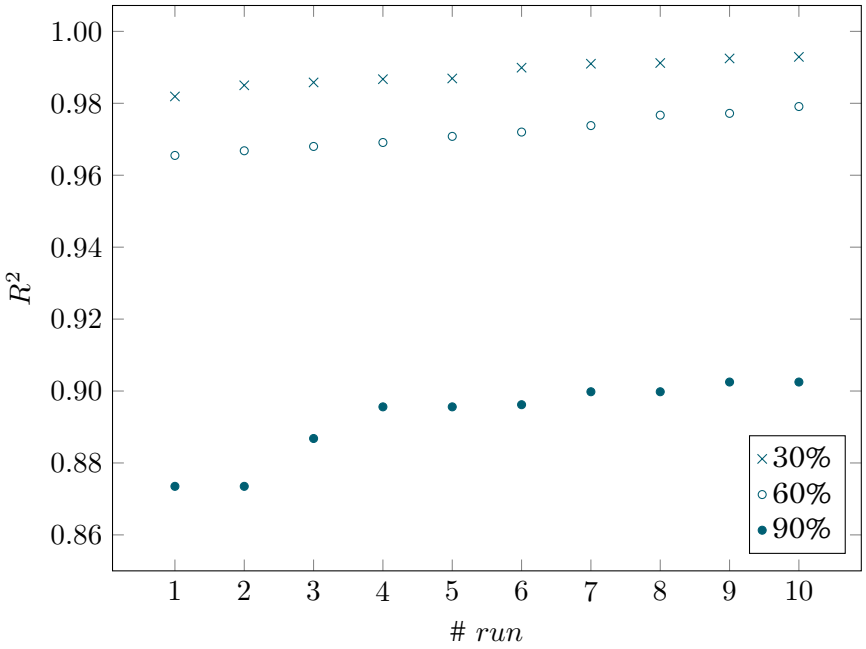


Figure 4.3: Sorted R^2 values for 10 runs of the *genetic algorithms* workflow and three levels of $predFactor$, using the *Gold ENMs* dataset and a single threshold.

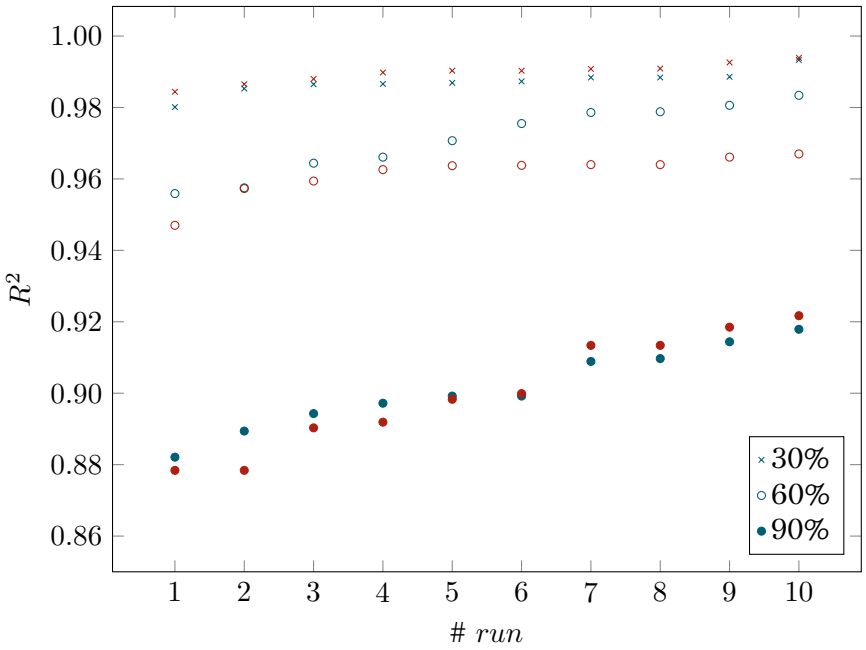


Figure 4.4: Sorted R^2 values for 10 runs of the *genetic algorithms* workflow and three levels of $predFactor$, using the *Gold ENMs* dataset and two thresholds. Black and red markers correspond to predictions using PD and BD distances respectively.

Table 4.5: Overview of the produced results and statistics from the GA workflow applied on the *Gold ENMs* dataset using a single or two thresholds in internal validation.

predFactor: 30%	Single threshold			Two thresholds					
				PD distances			BD distances		
	min	max	average	min	max	average	min	max	average
thresholds	0.5561	1.0134	0.8846	PD 0.4400	0.7579	0.5339	0.3436	0.7440	0.5803
				BD 0.5399	0.8806	0.7499	0.5373	0.8738	0.7455
# selected variables	46	61	53.6	43	56	49.5	48	59	52.1
# predicted samples	26	31	28.8	25	29	25.7	25	28	26.2
R^2	0.982	0.993	-	0.98	0.993	-	0.984	0.994	-
predFactor: 60%	Single threshold			Two thresholds					
				PD distances			BD distances		
	min	max	average	min	max	average	min	max	average
thresholds	0.9846	1.1728	1.0843	PD 0.2497	1.0554	0.7806	0.4841	0.8822	0.6844
				BD 0.6766	1.2691	0.9550	0.9780	1.1909	1.0801
# selected variables	46	59	52.3	39	62	50.6	50	62	53.7
# predicted samples	50	53	50.6	50	51	50.3	50	53	51
R^2	0.966	0.979	-	0.956	0.983	-	0.947	0.967	-
predFactor: 90%	Single threshold			Two thresholds					
				PD distances			BD distances		
	min	max	average	min	max	average	min	max	average
thresholds	1.5764	1.7488	1.6251	PD 0.9931	1.2318	1.1266	0.9834	1.3869	1.2383
				BD 1.1806	1.3546	1.2671	1.1084	1.4729	1.2399
# selected variables	55	65	60.3	47	64	54.0	48	58	55.2
# predicted samples	76	78	77.0	76	78	77.1	76	79	76.6
R^2	0.874	0.903	-	0.882	0.918	-	0.878	0.922	-

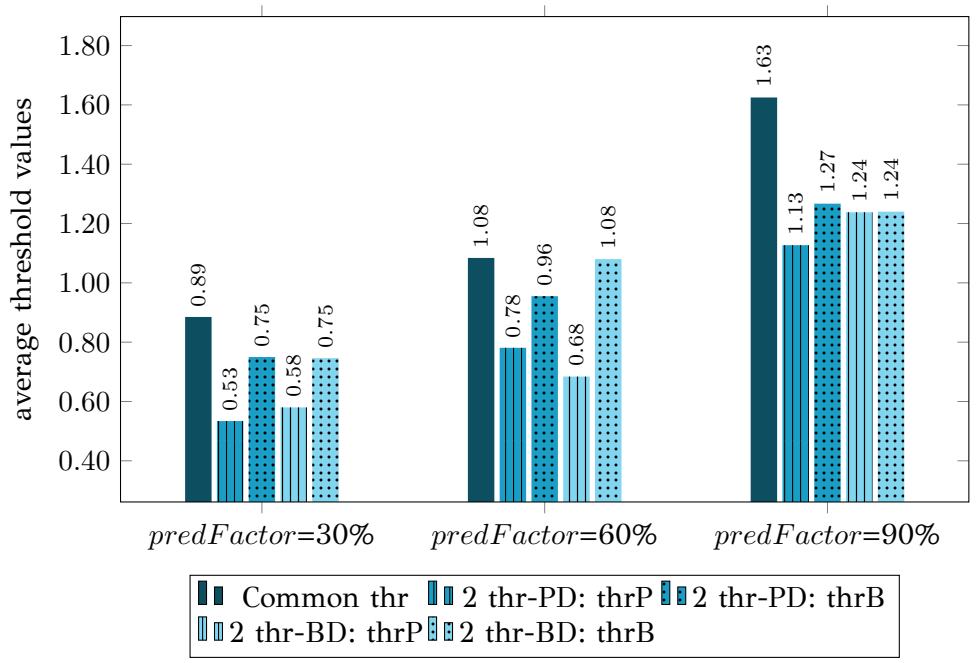


Figure 4.5: Average threshold values produced by the *genetic algorithms* workflow and three levels of $predFactor$, using the *Gold ENMs* dataset. Five columns are shown at each level. The first column shows the single threshold. The two next columns depict the thresholds corresponding to the groups of PDs and BDs respectively, when distances between PDs are used for the read-across predictions. The last two column present the two thresholds again, when read-across predictions are performed using the distances between BDs.

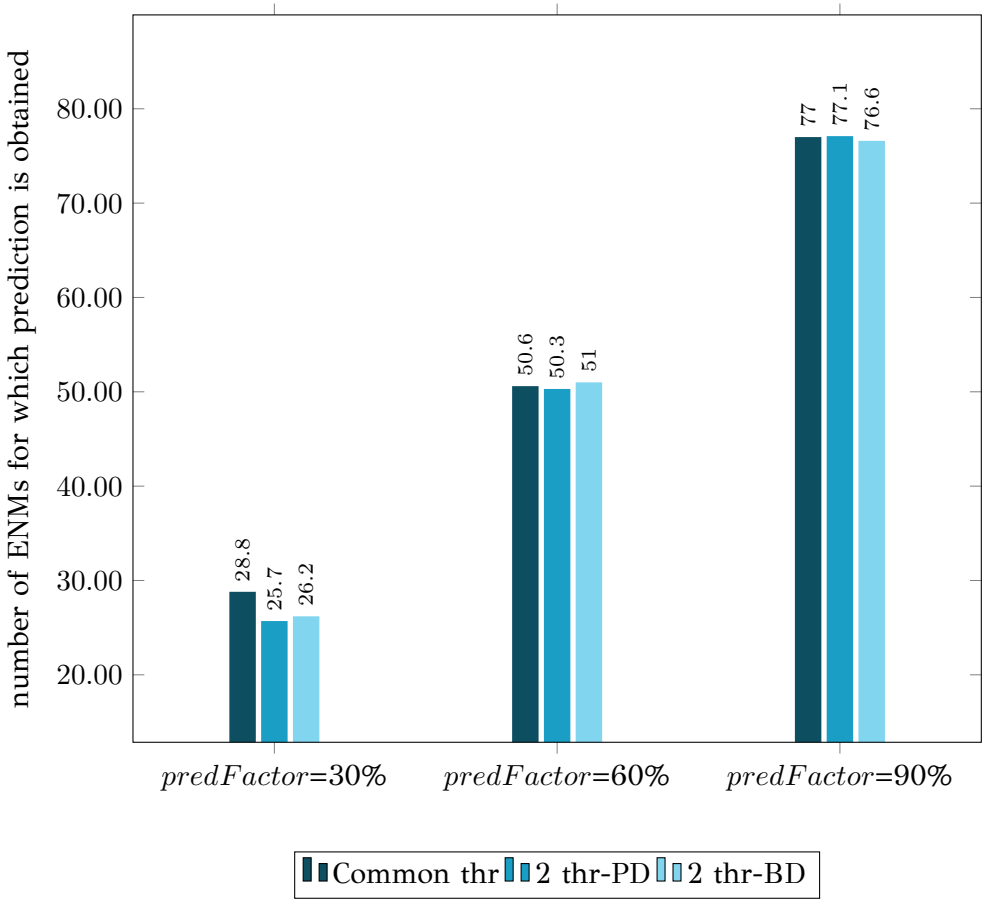


Figure 4.6: Average number of ENMs for which prediction is obtained from the *genetic algorithms* workflow and three levels of $predFactor$, using the *Gold ENMs* dataset.

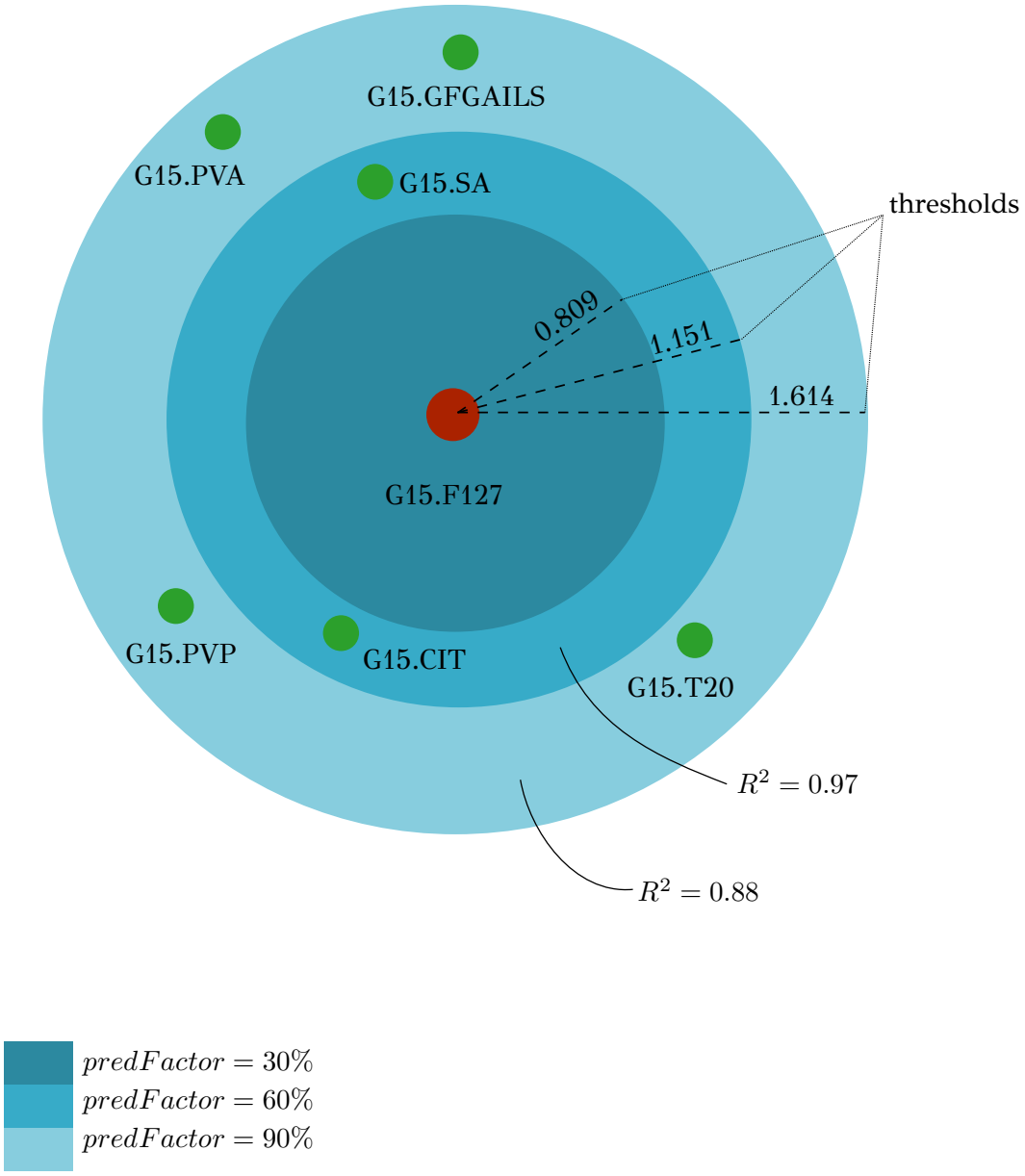


Figure 4.7: An example of the effect of the $predFactor$ of the *genetic algorithms* approach on the threshold, the number of neighbours and the predictive accuracy. The reference ENM is depicted with red color and the green ENMs are candidate neighbours. By increasing the $predFactor$ value, the threshold is increased, more ENMs with higher distances to the reference ENM are considered as neighbours and less accurate predictions are obtained. The single threshold variant is considered and a 2D projection of the multi-dimensional space is presented.

The presented descriptors in Figure 4.8 are extracted from ENM characterisation assays [76] and are further described next (see also Table Γ΄.2):

- lspri.rel.ch: ({LSPRi after serum exposure}-{LSPRi after synthesis})/{LSPRi after synthesis}
- zav.serum: Z-average hydrodynamic diameter (HD) after serum exposure
- vol.synth: Volume mean HD after synthesis
- num.serum: Number mean HD after serum exposure
- int.serum: Intensity mean HD after serum exposure

The localized surface plasmon resonance index (LSPRi) for each sample is computed from collected absorbance spectra, and is an empirical measure of the local dielectric environment surrounding plasmonic ENMs. The rest of the presented descriptors are measured by Dynamic Light Scattering (DLS) characterisation, using the available commercial software of the instrument (ZetaSizer Nano ZS, Malvern Instruments). [76] The HD parameter expresses "the size of a hypothetical hard sphere that diffuses in the same fashion as that of the particle being measured". [113] The HD is an important factor for ENM characterisation as it helps understand migration of ENMs into the (biological) media. Within a liquid (biological) medium, an electric dipole layer (in our case the protein corona) is formed around the dispersed ENM due to the surrounding macro-molecules and influences its brownian diffusion into the medium. [99], [114]–[116] Therefore, the HD encloses information of the ENM core along with any attached coating and formed solvent layer; a type of information that is based on resembled exposure conditions and cannot be estimated by other methods (e.g. size measured by TEM).

The hypergeometric test (Eq. 4.34) was applied, through *hygepdf* MATLAB® function, to the most frequently selected BDs shown in Figure 4.9, considering all genes (ENTREZ IDs) included in the molecular function category of the gene ontology (GO) at the time of writing (45633). [46]. Mapping of the significant PCF from Universal Protein Resource (UNIPROT) IDs to ENTREZ IDs was performed through the UniProt Consortium website (`www.uniprot.org/`). The most statistically significant GO terms (p-value$<$0.001) are depicted in Table 4.6.

$$p = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{4.34}$$

where K, the gene set size,
k, the selected biological variables (mapped in gene ENTREZ IDs) that belong to each gene set size,
N, the unique genes (ENTREZ IDs) in molecular function at the time of writing and,
n, the total number of biological variables (mapped in gene ENTREZ IDs) in frequency greater than 0.7, at $predFactor$ ratio equal to 0.6 (equal to 9).

Table 4.6: Significant GO terms containing the proteins of the *Gold ENMs* dataset selected by at least seven GA runs in the three variations. The size of the gene sets is placed in parenthesis next to their GO term name.

GO Term Name	GO Term ID	ENTREZ ID	UNIPROT	p-value
acrosin binding (4)	GO:0032190	5104	P05154	0.00079
acyl-L-homoserine-lactone lactonohydrolase activity (3)	GO:0102007	5444	P27169	0.00059
aryldialkylphosphatase activity (2)	GO:0004063	5444	P27169	0.00039
heparin binding (221)	GO:0008201	283; 5104	P03950; P05154	0.00081
phosphatidylcholine binding (26)	GO:0031210	341; 5104	P02654; P05154	0.00001
phosphatidylcholine-sterol O-acyltransferase activator activity (5)	GO:0060228	341	P02654	0.00099
protease binding (149)	GO:0002020	5104; 5265	P05154; P01009	0.00037
serine-type endopeptidase inhibitor activity (195)	GO:0004867	5104; 5265	P05154; P01009	0.00063

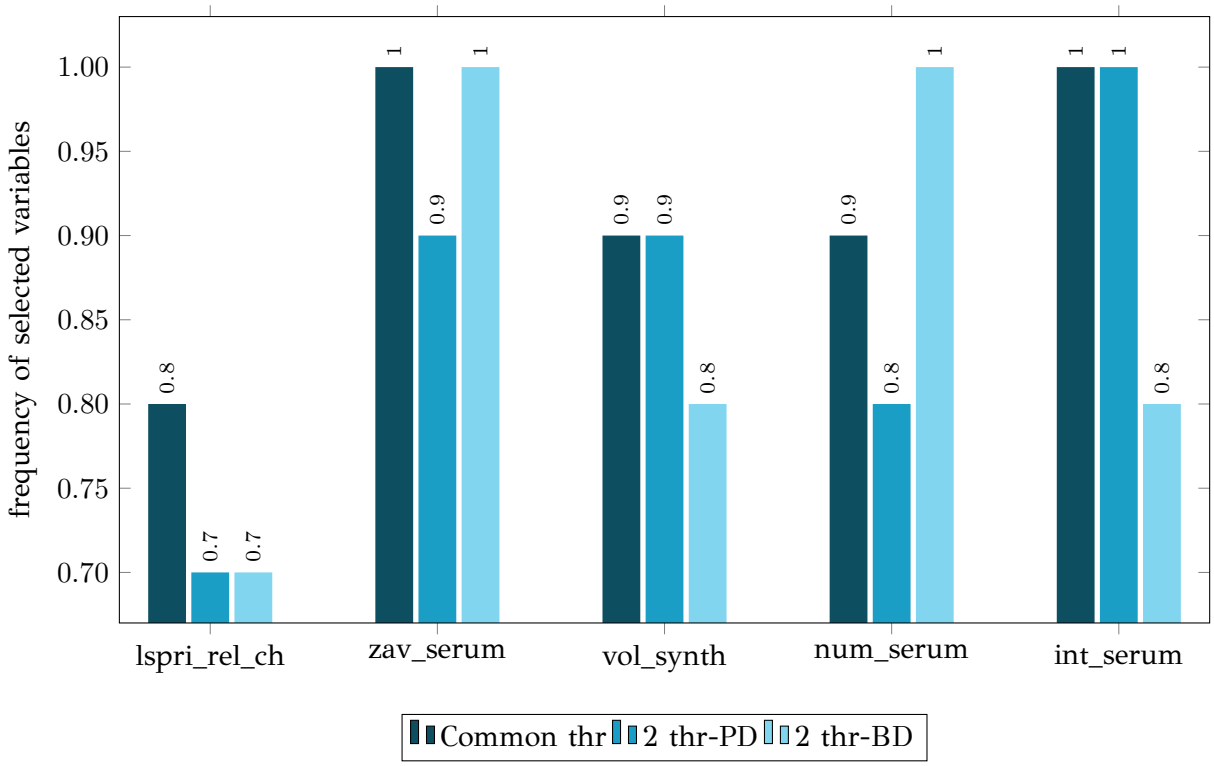


Figure 4.8: Selected physicochemical variables of the *Gold ENMs* dataset in frequency greater than 0.7 when using the *genetic algorithms* workflow at $predFactor$ ratio equal to 0.6.

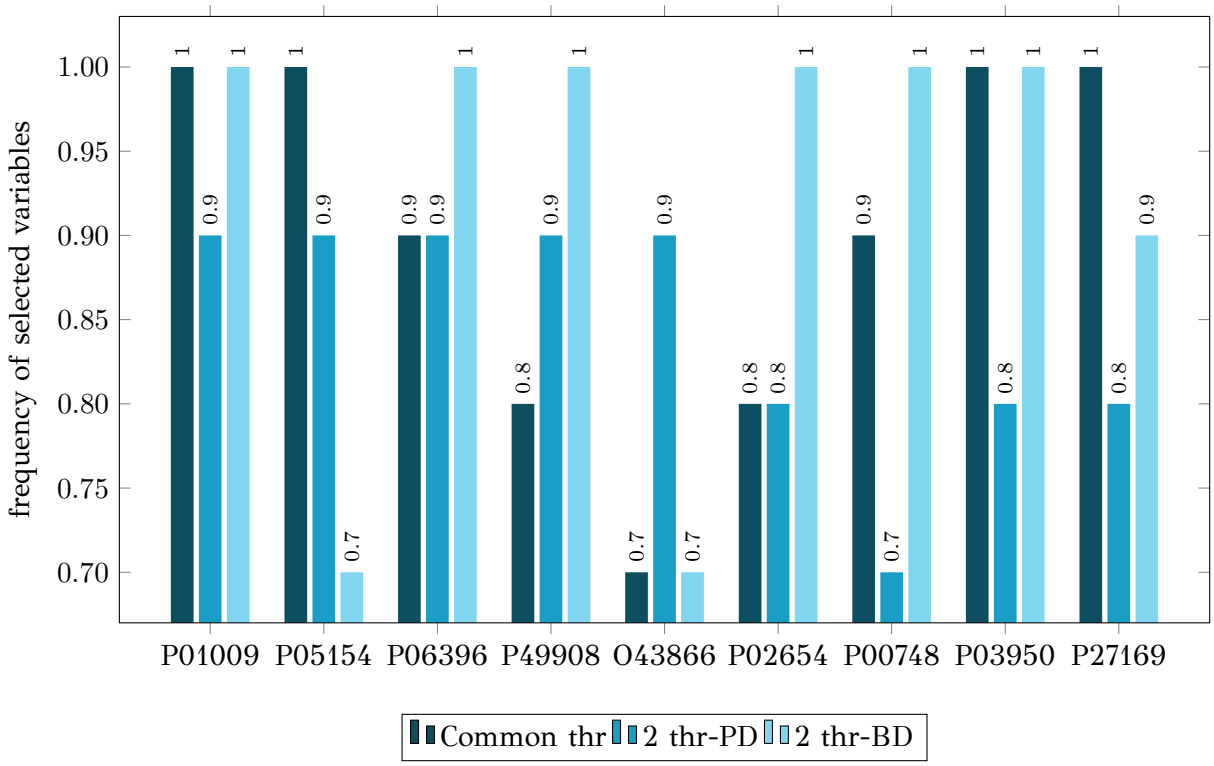


Figure 4.9: Selected biological variables of the *Gold ENMs* dataset in frequency greater than 0.7 when using the *genetic algorithms* workflow at $predFactor$ ratio equal to 0.6.

4.3.6.2 [r] External validation

Finally, the full dataset was divided into training and test sets in a ratio of 66:33 (55 training and 29 test ENMs) using the Kennard and Stone method. [66] We applied all three variations of the method described in the previous paragraph to the training data only and the parameters of Table 4.4. The selected variables and optimal threshold value(s) obtained by the solution of the optimisation problems were applied for computing read-across predictions for the test ENMs.

The results are summarized in Table 4.8[4] and in Figures 4.10, 4.11 for the single threshold and the two-threshold cases respectively. We observe that the $predFactor$ does not play a major role on the number of predicted samples, obviously because the Kennard-Stone algorithm forces the validation samples to be within the space defined by the training data. Even with the 30% $predFactor$ level, read-across predictions were obtained for most ENMs in the test set.

The best results in terms of the Q^2_{ext} statistic were produced with the 60% $predFactor$ level using two thresholds and the BD distances for calculating the read-across predictions. The prediction accuracy drops down significantly when applying the 90% $predFactor$ level, because in this case, as indicated before (Figures 4.5, 4.6, 4.7), the optimal threshold values are increased and the algorithm considers as neighbours, source ENMs with higher distances to the target ENM.

The Y-scrambling test was applied on the full dataset but using of scrambled endpoint values (only for the training set). Again, the dataset was divided into training and test sets in a ratio of 66:33 using the Kennard and Stone method and we applied the read-across workflow using a $predFactor$ of 0.6 and the rest training parameters of Table 4.4. The selected variables and optimal threshold value(s) obtained by the solution of the optimisation problems for scrambled endpoint values were applied for computing read-across predictions for the test ENMs. The results for 5 random scrambling of the training endpoints are summarized in Table 4.7. We can conclude, that the reliability of the produced models in terms of the Q^2_{ext} statistic, and is not due to chance correlation.

Table 4.7: Accuracy statistics of five different random shuffles in a Y-randomisation test for the GA workflow applied on the *Gold ENMs* dataset, using $predFactor = 0.6$.

	Q^2_{ext}
Y-random 1	-3.970
Y-random 2	-0.878
Y-random 3	-0.551
Y-random 4	-0.566
Y-random 5	-1.322

[4]Summarized results of 10 runs of the GA workflow are depicted.

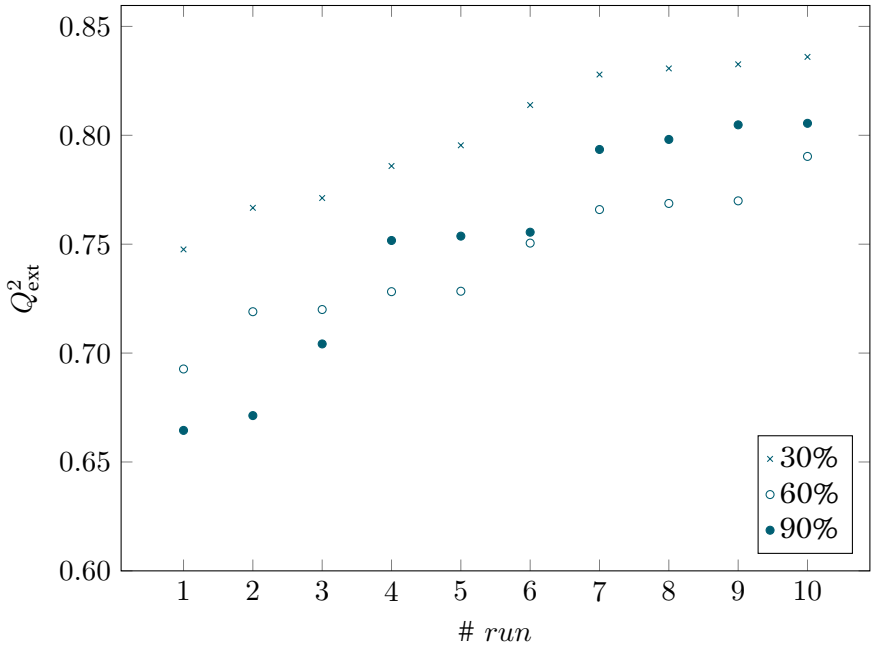


Figure 4.10: Sorted Q^2_{ext} values for 10 runs of the GA and three levels of $predFactor$, using a single threshold.

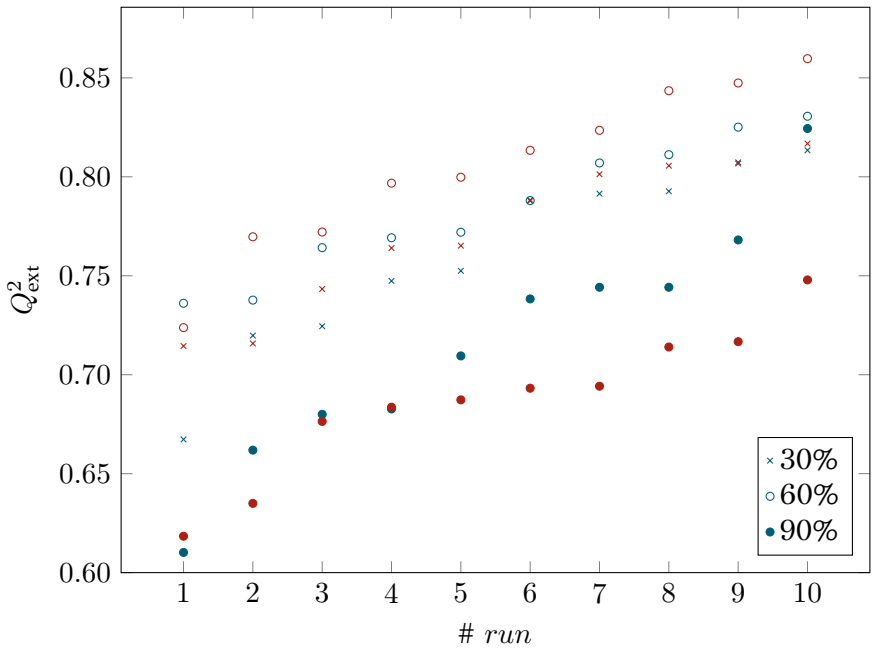


Figure 4.11: Sorted Q^2_{ext} values for 10 runs of the GA and three levels of $predFactor$, using two thresholds. Black and red markers correspond to predictions using PD and BD distances respectively.

Table 4.8: Overview of the produced results and statistics from the GA workflow applied on the *Gold ENMs* dataset using a single or two thresholds in external validation.

predFactor: 30%	Single threshold			Two thresholds					
				PD distances			BD distances		
	min	max	average	min	max	average	min	max	average
thresholds (PD)	1.0680	1.4009	1.2682	0.3336	0.7728	0.6158	0.4084	0.7793	0.5826
thresholds (BD)				1.0302	1.8066	1.3124	0.9279	1.8881	1.3770
selected variables	43	62	56.4	44	61	52.8	43	60	52.1
predicted samples	27	29	28.0	22	29	26.2	24	28	26.3
Q^2_{ext}	0.748	0.836	-	0.667	0.813	-	0.698	0.817	-
predFactor: 60%	Single threshold			Two thresholds					
				PD distances			BD distances		
	min	max	average	min	max	average	min	max	average
thresholds (PD)	1.3923	1.7148	1.4889	0.6779	1.3573	0.9772	0.8416	1.3832	1.0055
thresholds (BD)				1.1047	1.5270	1.3093	1.1309	1.4679	1.2805
selected variables	50	57	55	47	62	55.6	48	62	54.6
predicted samples	29	29	29	28	29	28.7	27	29	28.4
Q^2_{ext}	0.693	0.7903	-	0.736	0.831	-	0.724	0.860	-
predFactor: 90%	Single threshold			Two thresholds					
				PD distances			BD distances		
	min	max	average	min	max	average	min	max	average
thresholds (PD)	1.6594	1.9079	1.7494	1.1468	1.9579	1.4414	1.1214	1.5443	1.3407
thresholds (BD)				1.0865	1.7443	1.3872	1.1924	1.5521	1.3985
selected variables	52	63	56.8	45	64	55.0	50	67	58.5
predicted samples	29	29	29.0	29	29	29.0	29	29	29
Q^2_{ext}	0.664	0.806	-	0.610	0.824	-	0.618	0.748	-

4.3.6.3 [r] Controlling the influence of selected variables

The performance of the proposed read-across method with the regularisation term was studied by submitting the above dataset in external validation (using the same training and test sets). The workflow has been executed 10 times with the operational parameters of Table 4.4 in the three variations of the method (see §4.3.6.1), using this time two different wf_{OF} values (Tables 4.9 and 4.10). Figures 4.13 and 4.14 present the tracked Q^2_{ext} values, using one and two similarity thresholds respectively, and a wf_{OF} value of 0.005 from the 10 runs in a shorted manner. Similarly, Figures 4.15 and 4.16 present the tracked Q^2_{ext} values using a wf_{OF} value of 0.01.

Comparison with the implicit variable selection In Figure 4.12 a comparative bar-plot of the average numbers of selected variables from the 10 runs per different *predFactor* values and wf_{OF} values using one global similarity criterion are depicted. As expected the increase of the wf_{OF} value leads to the selection of less variables. However, the predictive power in external validation of the produced models is not largely affected by the selection of less variables, as presented in Figure 4.17, because read-across models have a similar predictive power compared to the case where wf_{OF} is set to 0. This is an important outcome from the use of a regularisation factor, as first the chance of over-fitting is eliminated from the use of less variables and second, in real-life applications the need of less predictive variables may contribute to the elimination of experiments needed for the risk assessment of ENMs. Similar results were produced from the use of two similarity criteria.

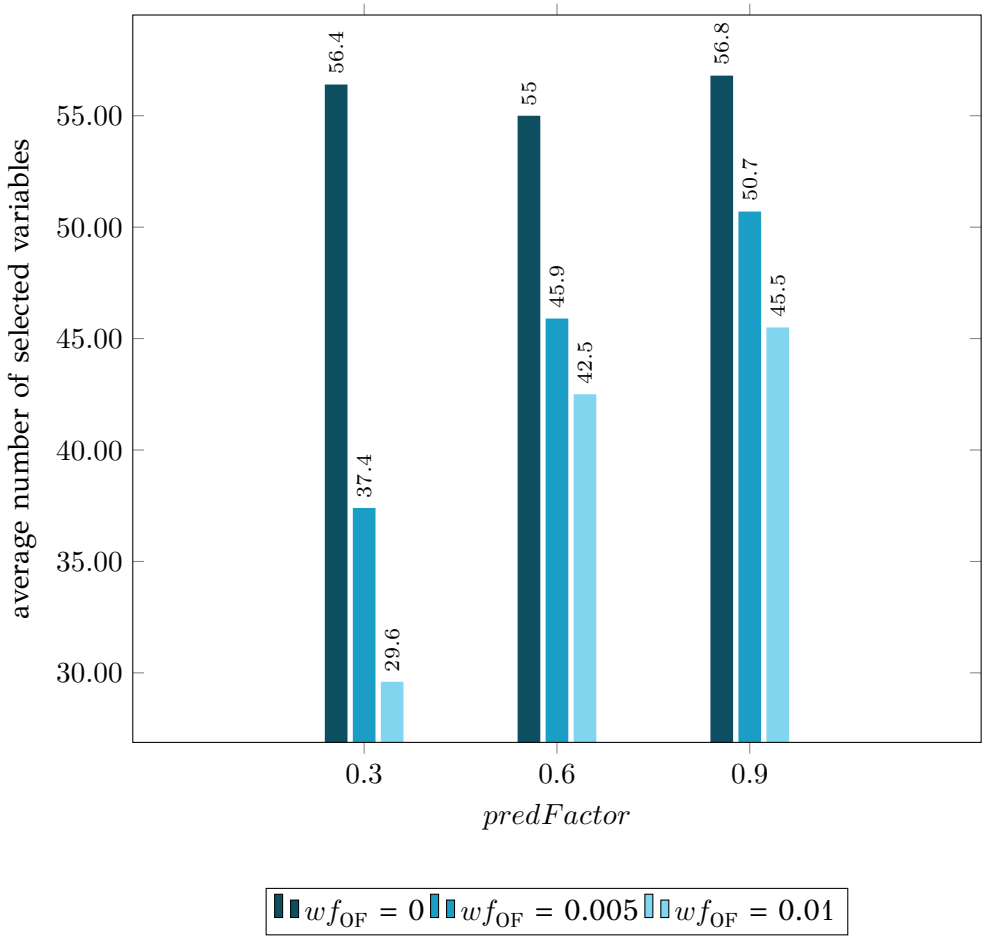


Figure 4.12: Average number of selected variables per different values of wf_{OF} and $predFactor$ using one similarity criterion in the GA scheme.

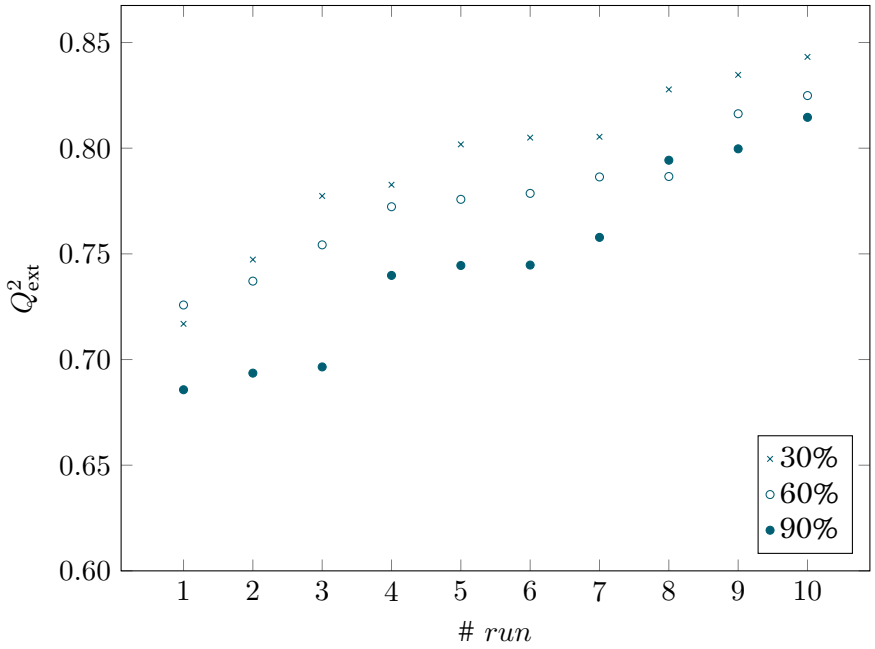


Figure 4.13: Sorted Q^2_{ext} values for 10 runs of the GA and three levels of $predFactor$, using a single threshold and a $wf_{\text{OF}} = 0.005$.

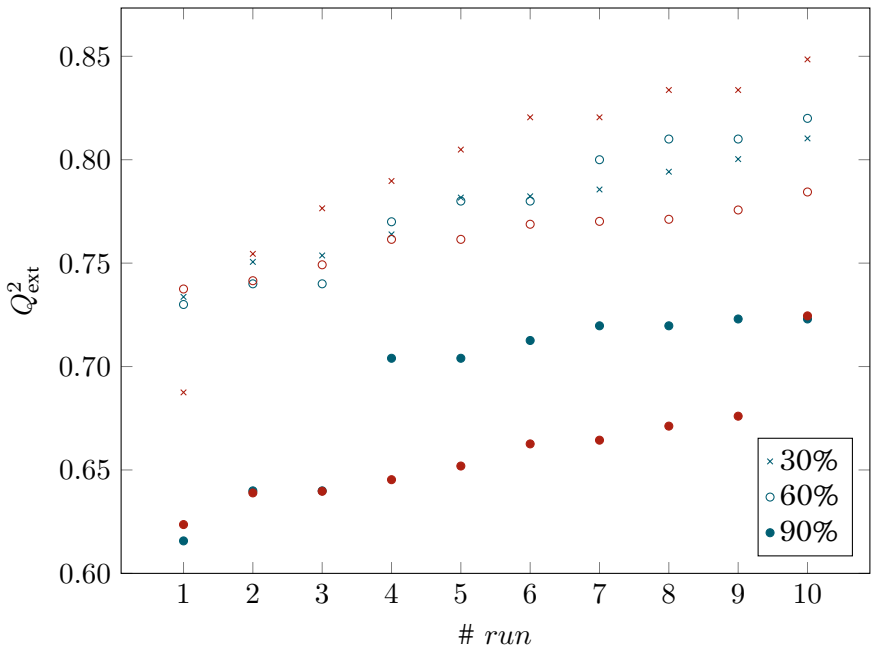


Figure 4.14: Sorted Q^2_{ext} values for 10 runs of the GA and three levels of $predFactor$, using two thresholds and a $wf_{\text{OF}} = 0.005$. Black and red markers correspond to predictions using PD and BD distances respectively.

Table 4.9: Overview of the produced results and statistics from the GA workflow applied on the *Gold ENMs* dataset using a single or two thresholds in external validation and a wf_{OF} = 0.005.

predFactor: 30%	**Single threshold**				**Two thresholds**					
					PD distances			BD distances		
	min	max	average		min	max	average	min	max	average
thresholds	0.9678	1.7088	1.1843	PD	0.3928	0.7883	0.6192	0.5422	0.7776	0.6611
				BD	0.6904	1.3024	1.0249	0.6197	1.2085	1.0158
selected variables	28	56	37.4		22	44	33.2	21	43	35.7
predicted samples	26	29	28.1		25	29	27.2	22	29	26.5
Q^2_{ext}	0.717	0.843	-		0.734	0.810	-	0.688	0.849	-
predFactor: 60%	**Single threshold**				**Two thresholds**					
					PD distances			BD distances		
	min	max	average		min	max	average	min	max	average
thresholds	1.1389	1.6079	1.4405	PD	0.7055	1.1715	0.9389	0.6547	1.0334	0.9296
				BD	1.0160	1.2377	1.1492	1.0100	1.3621	1.1812
selected variables	35	53	45.9		35	47	40.8	34	52	43.7
predicted samples	28	29	28.8		26	29	28.5	28	29	28.9
Q^2_{ext}	0.726	0.825	-		0.735	0.819	-	0.738	0.784	-
predFactor: 90%	**Single threshold**				**Two thresholds**					
					PD distances			BD distances		
	min	max	average		min	max	average	min	max	average
thresholds	1.5841	1.9669	1.7166	PD	1.1948	1.8513	1.5780	1.1176	1.8245	1.4298
				BD	1.1439	1.5270	1.3690	1.1556	1.4488	1.3510
selected variables	46	60	50.7		38	56	49.3	46	59	49.8
predicted samples	29	29	29.0		29	29	29.0	29	29	29.0
Q^2_{ext}	0.686	0.815	-		0.616	0.723	-	0.624	0.725	-

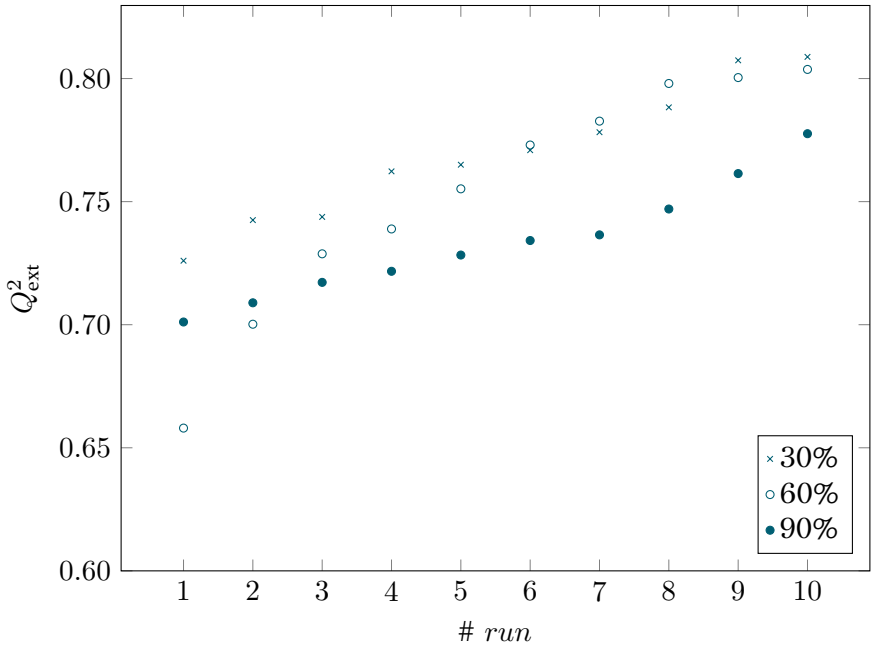


Figure 4.15: Sorted Q^2_{ext} values for 10 runs of the GA and three levels of $predFactor$, using a single threshold and a $wf_{\text{OF}} = 0.01$.

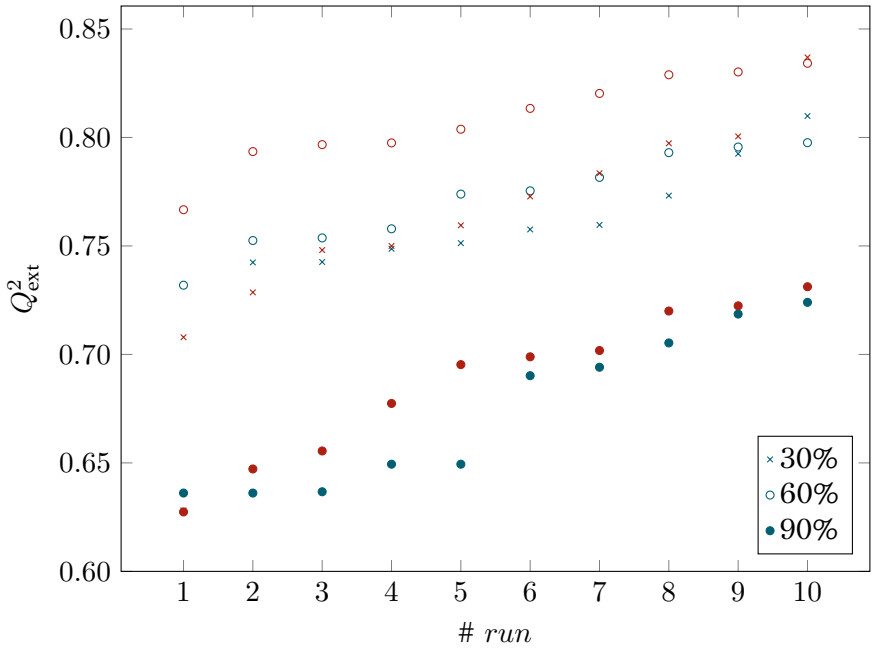


Figure 4.16: Sorted Q^2_{ext} values for 10 runs of the GA and three levels of $predFactor$, using two thresholds and a $wf_{\text{OF}} = 0.01$. Black and red markers correspond to predictions using PD and BD distances respectively.

Table 4.10: Overview of the produced results and statistics from the GA workflow applied on the *Gold ENMs* dataset using a single or two thresholds in external validation and a wf_{OF} = 0.01.

predFactor: 30%	**Single threshold**				**Two thresholds**					
					PD distances			BD distances		
	min	max	average		min	max	average	min	max	average
thresholds	0.8425	1.4398	0.9781	PD	0.4084	0.9218	0.5824	0.3555	0.7379	0.6038
				BD	0.6412	1.0835	0.8903	0.7336	1.2729	0.9533
selected variables	24	47	29.6		21	34	26.9	22	38	28.6
predicted samples	26	29	27.4		25	28	26.1	26	29	27.1
Q^2_{ext}	0.726	0.809	-		0.628	0.810	-	0.708	0.837	-
predFactor: 60%	**Single threshold**				**Two thresholds**					
					PD distances			BD distances		
	min	max	average		min	max	average	min	max	average
thresholds	1.2281	1.6079	1.4679	PD	0.7247	1.4246	0.8639	0.6551	1.2331	0.8479
				BD	0.8833	1.4547	1.1044	0.9125	1.1682	1.0790
selected variables	37	48	42.5		30	45	35.6	28	40	34.9
predicted samples	29	29	29		28	29	28.7	28	29	28.5
Q^2_{ext}	0.658	0.804	-		0.732	0.798	-	0.767	0.834	-
predFactor: 90%	**Single threshold**				**Two thresholds**					
					PD distances			BD distances		
	min	max	average		min	max	average	min	max	average
thresholds	1.4603	1.8511	1.6545	PD	1.1825	1.6958	1.3775	1.0651	1.6416	1.3125
				BD	1.2741	1.4683	1.3763	1.2083	1.5127	1.3684
selected variables	38	51	45.5		43	53	47.5	36	55	46.8
predicted samples	29	29	29.0		29	29	29.0	29	29	29.0
Q^2_{ext}	0.701	0.778	-		0.636	0.724	-	0.627	0.731	-

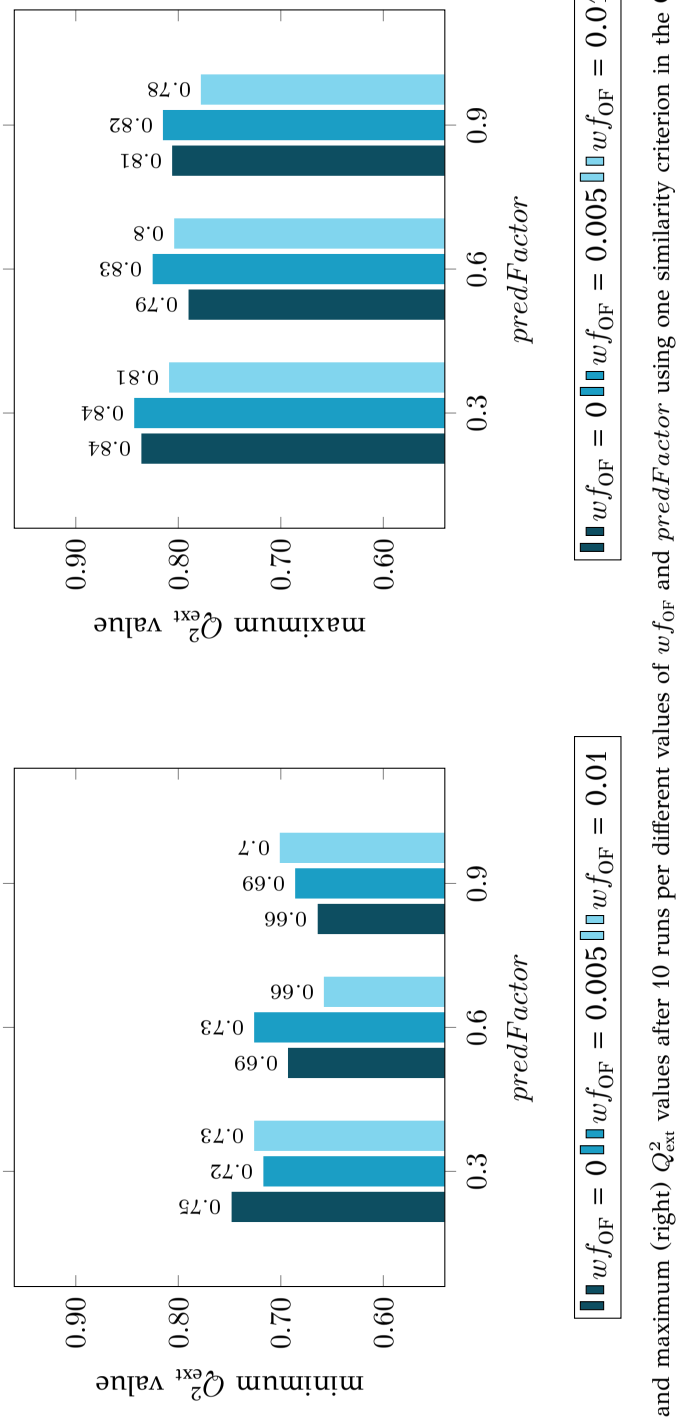


Figure 4.17: Minimum (left) and maximum (right) Q^2_{ext} values after 10 runs per different values of wf_{OF} and $predFactor$ using one similarity criterion in the GA scheme.

Table 4.11: GA workflow training parameters for the *Gold ENMs dataset* read-across model development.

Specifications	
Partitioning method	Kennard & Stone
Training ratio	0.66
# chromosomes	20
Generations	1000
Initial variable selection probability	0.6
Uniform *mutation* probability	0.01
Non-uniform *mutation* probability	0.1
Min threshold value	0.1
Max threshold value	mean(max(Dist))
bSA	1
Crossover probability	0.7
predFactor	0.6

4.3.6.4 [r] Additional case studies and comparison with other models and techniques

The developed read-across methodology was applied in two more case studies to further demonstrate its applicability and to discuss the produced results, using an external validation scheme.

Gold ENMs dataset The proposed read-across workflow has been previously applied on the *Gold ENMs* dataset presented by Walkey *et al.* (2014) [76] and filtered by Varsou *et al.* (2017) [2]. Here we present the results of two trained models using one similarity criterion and two regularisation factor values. Before training, all descriptor values were normalised. A detailed case study regarding this dataset can be found in §4.3.6.2.

Table 4.11 presents the training specifications, and Table 4.12 summarizes the produced results and statistics, with the presence or absence of the regularisation factor. Clearly, selection of a non-zero value for the regularisation factor wf_{OF} resulted to a read-across model that has a similar predictive power compared to the case where wf_{OF} is set to 0, but using significantly less input variables. This is an important factor when modelling is performed in order to address the needs of real-life applications; less experiments are needed and thus time and resources can be saved due to the necessity of less variables for the prediction calculation.

Next, analytical results for the PCF read-across model for $wf_{\text{OF}} = 0.05$ are presented. Table 4.13 presents the selected variables for the prediction of cell association and Table Δ′.2 depicts the training and test pairs of ENMs which are considered as neighbours by the algorithm. Pairs of neighbouring ENMs are marked with number 1. We can observe that in most cases neighbour ENMs share the same type of coating (either "anionic" or "cationic") thus, we can conclude that gold ENMs with the same surface charge modifications behave similarly as far as cell association levels are concerned. Future read-across studies could confirm the hypothesis of grouping the surface-modified ENMs based on their surface charge.

For comparison purposes we applied on the *Gold ENMs* dataset the *k*NN algorithm through KNIME and Waikato Environment for Knowledge Analysis (`WEKA`) suite nodes (§A′.1.4 and B′.2.4.1). Descriptor values were normalised between [0, 1] and the same train-test ratio was used in the Kennard and Stone partition method (0.66). Variable selection was applied before modelling using the `CsfSubsetEval - BestFirst` selection methodology included in `WEKA`. Eleven variables were used, the *k* parameter was tuned to the value of 3 and the Q^2_{ext} metric

Table 4.12: Results of the *Gold ENMs* read-across models built using the GA workflow. The number of predicted samples, and the MSE, MAE and Q^2_{ext} validation metrics refer to the test set.

Optimised parameters and validation metrics		
wf_{OF}	0	0.05
threshold	1.336	0.996
# variables	57	27
# predicted samples	29	29
MSE	0.692	0.662
MAE	0.647	0.612
Q^2_{ext}	0.769	0.779

Table 4.13: Selected variables of the *Gold ENMs dataset* read-across model built using the GA workflow for $wf_{\text{OF}} = 0.05$. Detailed explanation of the descriptors can be found in Tables Γ΄.1 and Γ΄.2.

Selected variables		
lspri.synth	zp.serum.mag	Q03591
zav.serum	P01024	O43866
num.synth	Q14624	P02749
int.serum	P01023	P02654
hdlayer.synth	P02656	Q99467
num.ch	P00739	P03952
int.ch	P19823	P00738
zp.synth.sign	P10720	P01011
zp.synth.mag	P49908	P00450

was equal to 0.723. This value is significantly lower than the prediction statistics produced by the optimised read-across models with or without a regularisation term.

MWCNTs [a] dataset Finally, our read-across methodology was applied in the *MWCNTs [a]* dataset as presented in §3.4. All descriptor values were scaled in range [0,1] prior to training.

Due to the small size of this dataset (only 5 descriptors), one read-across model was trained, using a regularisation factor wf_{OF} equal to zero. Table 4.14 summarizes the training specifications, and Table 4.15 the produced results and validation statistics.

Table 4.16 presents the selected variables of the generated model. Table Δ΄.3 presents the neighbour MWCNTs for this particular case study.

Using the same dataset, a *k*NN model was developed again in KNIME, using three neighbours and 4 features selected by the `CsfSubsetEval - BestFirst` selection methodology included in `WEKA`. The Q^2_{ext} metric was equal to 0.808; comparable to the accuracy of the optimised read-across models which use 3 features.

4.4 [m] Extension on classification problems

In the next part of this Chapter, we are presenting an extension of the above methodology (including the regularisation) to the prediction of categorical-class endpoints. The basic workflow steps in this case are similar to the original methodology, however due to the nature of classification data, prediction, solution evaluation and validation are performed in a different way.

Table 4.14: GA workflow training parameters for the *MWCNTs [a]* read-across model development.

Specifications	
Partitioning method	Kennard & Stone
Training ratio	0.75
# chromosomes	50
Generations	200
Initial variable selection probability	0.6
Uniform mutation probability	0.01
Non-uniform mutation probability	0.1
Min threshold value	0.1
Max threshold value	mean(max(Dist))
bSA	1
Crossover probability	0.7
predFactor	0.4

Table 4.15: Results of a *MWCNTs [a]* read-across model built using the GA workflow. The number of predicted samples, and the MSE, MAE and Q^2_{ext} validation metrics refer to the test set.

Optimised parameters and validation metrics	
threshold	0.302
# variables	3
# predicted samples	7
MSE	0.148
MAE	0.346
Q^2_{ext}	0.805

Table 4.16: Selected variables of the *MWCNTs [a]* read-across model built the GA workflow.

Selected variables	
R	Lone-pair electrons
α	Hydrogen-bond acidity
V	Lipophilicity interaction

4.4.1 [m] Development of a GA workflow

The three-step process described in §4.3.2 is still the core element of our GA workflow. In the next paragraphs we describe briefly the steps that lead from a *chromosome* to the calculation of the categorical read-across predictions and the assessment of their quality.

4.4.1.1 Variable selection

As described before, the key element in the GAs is the *chromosome* (Table 4.3). *Chromosomes* contain *genes* in a specific order. Each *gene* except the last one corresponds to a specific descriptor and takes a binary value that encodes the selection or not of the corresponding descriptor to the prediction workflow. Through the *generations*, *populations* of *chromosomes* are subject to the biological operators of *selection*, *crossover*, *mutation* and *elitism*, leading to the *genome* which contains the optimal combination of selected descriptors to produce reliable predictions. The total number of selected variables is controlled by a regularisation factor (wf_{OF}). *Mutation* for binary *genes* is performed by inverting the value of the selected *genes*: 0

becomes 1 and vice versa (uniform *mutation*).

4.4.1.2 Definition of neighbours

In order to define the neighbours of a query ENM (i), the Euclidean distances between the query and all training ENMs is calculated, considering only the selected descriptors of a particular *chromosome*. Next, from the pool of j training ENMs, the ones with distance $dist_{i,j}$ equal or lower than the threshold value - which is the last element of the *chromosome* - are selected as the neighbours of the query ENM (Figure 4.2). In this case the binary variables $neib_{i,j}$ take the value of 1, otherwise they take the value of 0. During training when $i = j$, the value of $neib_{i,j}$ is automatically set equal to zero. [2]

When different categories of descriptors are available (e.g. physicochemical, image, theoretical, bioinformatics descriptors), it is possible to use more than one similarity criteria (thresholds) for the selection of neighbours. In this case, Euclidean distances between ENMs are calculated independently for each type of descriptors (considering only the selected ones indicated by the *chromosome*). Two ENMs are considered similar, only if all distances are below the respective thresholds. [2], [45]

Threshold values are also subject to genetic operators and converge to their optimal value during the evolutionary process, as parts of the *chromosomes*. Unlike the rest of the *genes*, thresholds are continuous, and the non-uniform *mutation* operator described in (Eq. 4.29) is applied. In case that more than one similarity criteria are used for neighbours selection, equal number of thresholds that evolve independently through *generations*, is used.

4.4.1.3 Read-across prediction

Each *chromosome* generated through the GA process corresponds to a read-across model. The predicted read-across value for the ith ENM in the training set is calculated using only the neighbour ENMs ($neib_{i,j} = 1$). This calculation assumes the existence of at least one neighbour ENM. For categorical endpoints, the prediction is the distance-weighted majority vote of the closest training neighbours (Eq. 4.35). All $class$ arguments are of a binary type TRUE/FALSE.

$$\hat{class}_i = \begin{cases} \underset{class}{\text{argmax}} \left(\sum_{j=1}^{N_{\text{tr}}} \frac{neib_{i,j}}{1+dist_{i,j}} \cdot class_j \right), \text{ if } pred_i \neq 0 \\ \\ NA, \text{ if } pred_i = 0 \end{cases} \qquad \forall i = 1, \ldots, N_{\text{tr}} \tag{4.35}$$

where, N_{tr}, is the number of ENMs in the training set ,
$\hat{y}_i$, is the predicted endpoint value for the ith training ENM with at least one neighbour,
y_j is the actual endpoint value of the jth training ENM,
$\hat{class}_i$ is the predicted categorical endpoint value of the ith training ENM with at least one neighbour
$class_j$ is the actual categorical endpoint value of the jth training ENM
$neib_{i,j}$ is a binary variable taking the value of 1 if ENMs i and j are neighbours and 0 if they are not,
$dist_{i,j}$ is the Euclidean distance between ENMs i and j,
$pred_i$ is a binary variable that becomes equal to 1, when a read-across prediction is achieved for the ith ENM, and 0, if no prediction is possible.

In case that two types of descriptors *A* and *B* are available and neighbours can be selected based on two similarity thresholds, Eq. 4.35 can be adapted in order to predict the ENMs

class based on A (or B) distance metrics.

$$\hat{class}_i = \begin{cases} \underset{class}{\text{argmax}} \left(\sum_{j=1}^{N_{\text{tr}}} \frac{neib_{i,j}}{1+distA_{i,j}} \cdot class_j \right), \text{ if } pred_i \neq 0 \\ \\ NA, \text{ if } pred_i = 0 \end{cases} \qquad \forall i = 1, \ldots, N \tag{4.36}$$

where $distA_{i,j}$ is the Euclidean distance between ENMs based on A group of descriptors i and j.

A *chromosome* is accepted if the corresponding read-across model satisfies Eq. 4.8, which means that read-across predictions can be computed for at least a percentage of the training samples, defined by the parameter *predFactor*. In other words, this percentage of training samples have at least one neighbour. If a *chromosome* does not satisfy Eq. 4.8, it is rejected, and is substituted by its best *parent chromosome*.

4.4.1.4 Potential solution evaluation

Each *chromosome* is evaluated for its ability to produce accurate predictions using a score (fitness value). *Chromosomes* with higher fitness values are more probable to survive during the "roulette wheel" *selection* process of the evolutionary algorithm. To assess the reliability of the predictions, the MCC is calculated, according to Eq. 2.24.

The definition of the OF for categorical endpoints (Eq. 4.37) is based on the MCC (Eq. 2.24). The score value is computed as in the case of numerical endpoints (Eq. 4.28).

$$OF = |0.5 - 1/(1 + MCC)| + wf_{\text{OF}} \cdot \sum_{\ell=1}^{L} attr_\ell \tag{4.37}$$

where $attr_\ell$, is a binary variable indicating if the descriptor ℓ is selected and L is the total number of available descriptors.

In the OF definition for categorical endpoints, the first term is $|0.5 - 1/(1 + MCC)|$, which becomes zero when $MCC = 1$, i.e. in the case of perfect match between predicted and actual values. At this point we have to highlight that in case of $MCC = -1$ (total disagreement between predictions and actual endpoint values), *chromosome*'s score (Eq. 4.28) is automatically set to zero.

4.4.2 [m] Validation of the produced read-across model

In order to ascertain the reliability of the produced read-across predictions an external validation scheme [64] is used to test the performance of the produced read-across model in terms of predicting accurately the endpoint on ENMs that have not been used during the training process. To this end, the read-across model which is the final outcome of the training evolutionary workflow (*genome*), is used to compute endpoint predictions for the ENMs belonging to the test set. These calculations are performed using Eqs. 4.35 and 4.36, adapted for test samples. [2]

In case of categorical endpoints, validation results are displayed in a confusion matrix (Table 2.2), where TP, TN, FP, FN have been defined in Eq. 2.24. The proportion of actual TRUE-class ENMs that are correctly classified as "TRUE" is measured by sensitivity (Sn, Eq. 2.21) and the proportion of actual FALSE-class ENMs that are correctly classified as "FALSE" is measured by specificity (Sp, Eq. 2.22). The overall success rate is measured by accuracy (Ac, Eq . 2.23). [1]

The MCC metric is also calculated for the test set, according to the Eq. 2.24, which is adapted for the test samples.

4.4.3 [m] Use of a read-across model to predict the endpoints of untested ENMs

The use of a successfully trained model for the prediction of classes is similar to the use for numerical endpoints (§4.3.4) using the optimum variables and threshold(s), provided that the selected input descriptors indicated by the corresponding *genome* are available. The endpoint estimation is performed according to Eq. 4.35 (or Eq. 4.36). In case that no neighbours are located for an untested ENM in the train set, no estimation can be performed, and this ENM is considered out of model's domain of applicability.

4.4.4 [m] Implementation

The implementation of the GA workflow for the prediction of categorical endpoints was performed in R programming language. R is more flexible than MATLAB® in categorical data manipulation and in addition, due to our aspiration of creating a web application for our GA workflow (see §4.5) using R `shiny` package, there was a need of translation of our original MATLAB® code to R code. The source code is available at GitHub (`https://github.com/DemetraDanae/optimized-read-across/tree/master/categorical_endpoints`) considering a single threshold (extension to two or more criteria is trivial) released under GNU General Public License v.3.

4.4.5 [r] Results and discussion

The developed read-across GA workflow for the prediction of categorical endpoints is demonstrated on the *MeOx ENMs [a]* dataset (see §3.3) which includes 25 ENMs metal (hydr)oxide, 4 composition-related (non-nano) and 8 dimensional (nano) descriptors. The values of each descriptor were scaled in the range [0,1], in order to be comparable and contribute impartially to the read-across predictions (§2.1.1). The availability of two different types of descriptors, renders this dataset suitable for testing the proposed method with one or two similarity criteria. The developed methodology was validated both internally (following the same procedure as in §4.3.6.1) and externally using a training an a test subset.

4.4.5.1 [r] Internal validation

In this case the GA method was applied in the entire dataset with the operational parameters of Table 4.17. Again, due to the stochastic nature of the methodology, the whole workflow was executed 10 times and the accuracy statistics of the produced models were tracked. A *predFactor* level of 0.9 and a wf_{OF} of zero were selected in order to produce results comparable to the publication of Forest *et al.* [82] where the same dataset was used in a LOO cross-validation scheme -similar to this internal validation scheme. Three variations of the method were followed:

- Considering a single threshold, corresponding to the full set of descriptors.
- Assuming two different thresholds, one for the group of non-nano and one for the group of nano descriptors and obtaining the read-across predictions using the distances between non-nano descriptors.
- Assuming two different thresholds, one for the group of non-nano and one for the group of nano descriptors and obtaining the read-across predictions using the distances between nano descriptors.

Figures 4.18, 4.19 and 4.20 present in a shorted manner the accuracy, sensitivity and specificity values using one and two similarity thresholds for 10 different runs. The results

Table 4.17: Values for the user-defined operational parameters of the proposed read-across GA workflow applied on the *MeOx ENMs [a]* dataset in internal validation.

Parameter	Value
nChrom	50
maxGenerations	50
initGeneProb	0.6
crossProb	0.7
mutProb	0.01
nonUnf	0.1
$thr_{\min}^{GA}$	0.1
$thr_{\max}^{GA}$	mean value of the maximum distances between samples
bGA	1
predFactor	0.9

Table 4.18: Selected variables from the GA workflow applied on the *MeOx ENMs [a]* dataset in frequency greater than 0.7. The frequency ratio of each variable is presented in a parenthesis next to each name.

Scenario	Most significant variables			
Single threshold	solubility (1.0)	zeta (1.0)	SF (0.7)	CSF (0.7)
Two thresholds (non-nano prediction base)	solubility (0.9)	zeta (1.0)		
Two thresholds (nano prediction base)	solubility (1.0)	zeta (1.0)	$d_{\min}$ (0.7)	

are also summarized in Table 4.19[5]. All the applied scenarios lead to satisfactory results in terms of predictive accuracy, however the use of one instead of two thresholds shows higher probability of achieving a 100% accuracy and for all the dataset samples.

Any of the developed models with the proposed GA strategy using one similarity threshold, produced better accuracy results comparing to the decision tree strategy of Forest *et al*. [82], that presents in LOO cross-validation a global accuracy of 0.8, a sensitivity of 0.8 and a specificity also of 0.8. However, an average decision tree combining all the 25 developed trees of LOO cross-validation lead to a model of 100% accuracy on the whole dataset.

We also measured the frequency of appearance of the different descriptors in the selected sets of descriptors. It is clear that there exist descriptors which are selected in most runs, whereas some other descriptors are chosen very rarely. The descriptors appearing in more than 70% of the runs are considered as the most significant descriptors, and are presented in Table 4.18. Comparing to the initial study of Forest *et al*. [82], we can observe that our method leads to the selection of mainly dimensional descriptors instead of compositional. In fact, solubility, zeta-potential and $d_{\min}$ are the three descriptors excluded from the Forest *et al*. [82] analysis without any satisfactory explanation. Solubility and zeta-potential, which are dominant selected variables in all three scenarios are greatly related to nanotoxicity, especially due to their contribution in ENMs agglomeration phenomena (see §3.6).

[5]Summarized results of 10 runs of the GA workflow are depicted.

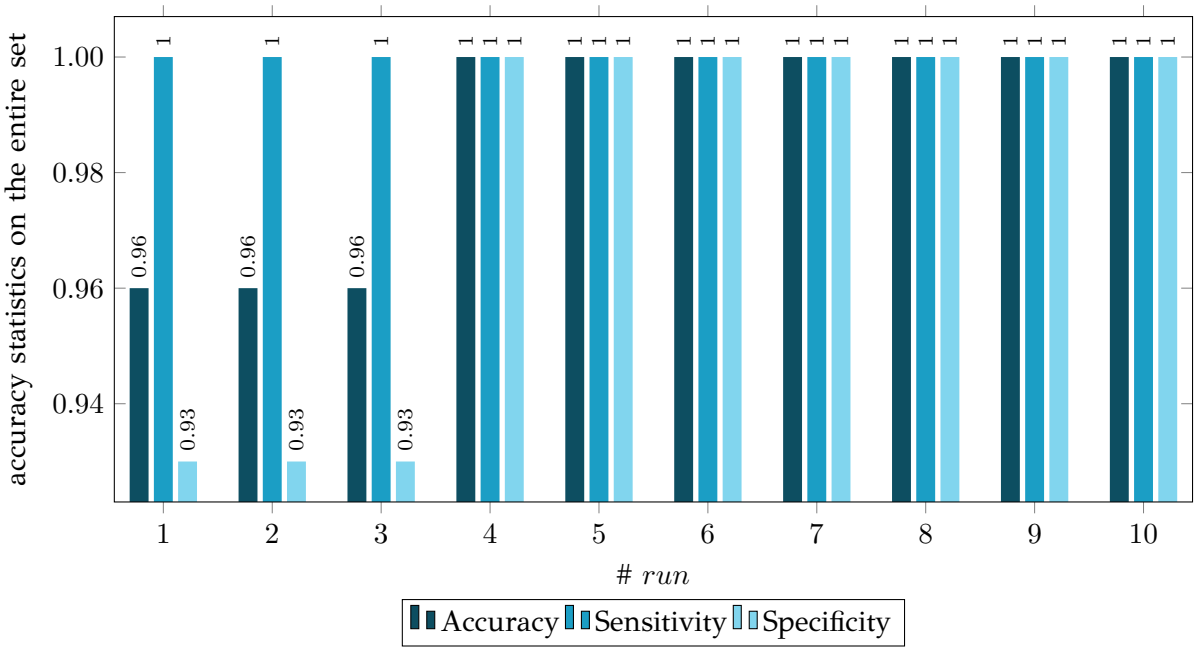


Figure 4.18: Accuracy statistics per 10 different runs using one similarity threshold for $predFactor$ = 0.9 and wf_{OF} = 0 in the GA scheme.

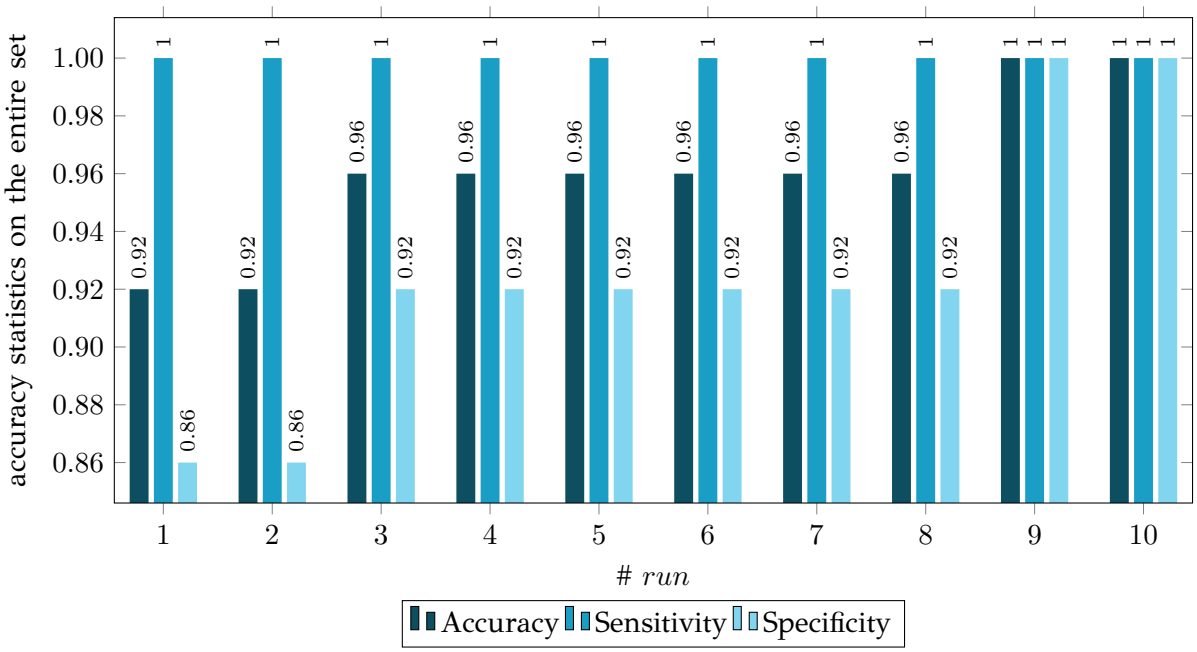


Figure 4.19: Accuracy statistics per 10 different runs using two similarity thresholds and non-nano prediction base for $predFactor$ = 0.9 and wf_{OF} = 0 in the GA scheme.

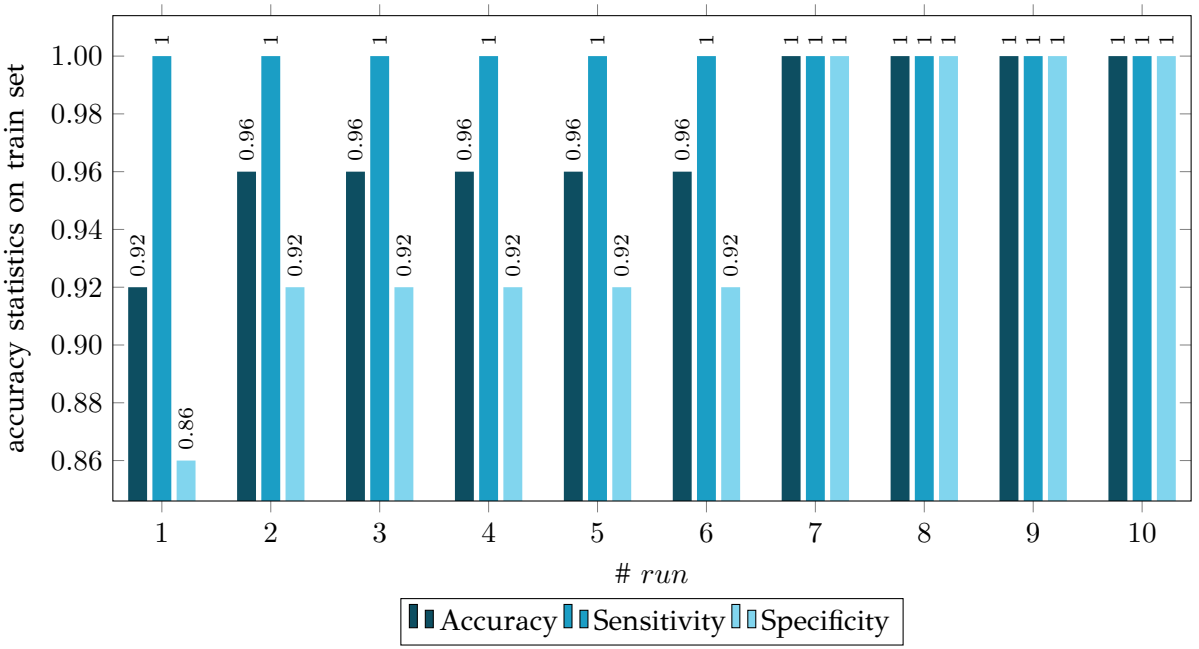


Figure 4.20: Accuracy statistics per 10 different runs using two similarity thresholds and nano prediction base for $predFactor$ = 0.9 and wf_{OF} = 0 in the GA scheme.

Table 4.19: Overview of the produced results and statistics from the GA workflow applied on the *MeOx ENMs [a]* dataset using a single or two thresholds in internal validation for *predFactor* = 0.9 and $wf_{\text{OF}} = 0$.

predFactor: 90%	Single threshold			Two thresholds (non-nano prediction base)			Two thresholds (nano prediction base)		
$wf_{\text{OF}} = 0$	min	max	average	min	max	average	min	max	average
selected non-nano variables	1	3	1.7	1	3	2.4	1	3	1.8
selected nano variables	3	6	4.7	2	6	4.3	3	6	4.8
predicted samples	25	25	25	22	25	24.1	22	25	24.3
Accuracy	0.960	1.000	-	0.920	1.000	-	0.920	1.000	-
Sensitivity	1.000	1.000	-	1.000	1.000	-	1.000	1.000	-
Specificity	0.929	1.000	-	0.857	1.000	-	0.857	1.000	-

4.4.5.2 [r] External validation

In continuation, the full dataset was divided into training and test sets in a ratio of 75:25 (19 training and 6 test ENMs) using the Kennard and Stone method. [67] We applied the described method for one similarity criteria (threshold), and with or without a regularisation factor (wf_{OF}) for controlling the number of selected variables. The training specifications are summarized in Table 4.20 The selected variables and optimal threshold value(s) obtained by the solution of the optimisation problems were applied for computing read-across predictions for the test set ENMs.

Table 4.20: Values for the user-defined operational parameters of the proposed read-across GA workflow applied on the *MeOx ENMs [a]* dataset in external validation using one similarity criterion.

Parameter	Value
nChrom	50
maxGenerations	50
initGeneProb	0.6
crossProb	0.7
mutProb	0.01
nonUnf	0.1
$thr_{\text{min}}^{\text{GA}}$	0.1
$thr_{\text{max}}^{\text{GA}}$	mean value of the maximum distances between samples
bGA	1
predFactor	0.3-0.6-0.9

The results are summarized in Table 4.21[6] and in Figures 4.21, 4.22 for the single threshold. The best results in terms of predictive accuracy (including sensitivity and specificity) were produced with the 60% *predFactor* level for all 10 runs either using a weighting factor of zero or a non-zero weighing factor. Especially for wf_{OF} = 0.001, a balance between predictive accuracy, number of used variables and number of ENMs with a successful prediction is achieved eliminating at the same time the chance of over-fitting. The prediction accuracy drops down when applying the 90% *predFactor* level, because in this case the optimal threshold values are increased and the algorithm considers as neighbours, source MeOx ENMs with higher distances to the target MeOx ENM.

Due to the availability of two types of descriptors (nano and non-nano) the GA workflow was also applied on the *MeOx ENMs [a]* dataset using two similarity criteria and the parameters of Table 4.22. Due to the rapid convergence of the method in optimal solutions, a total number of 10 generations was selected in order to expedite model development. In addition, due to the repeatability of the results, in terms of accuracy statistics, 5 runs of the GA workflow were performed. Table 4.23 presents the summarized results of 5 runs of the GA workflow for the developed models using different *predFactor* values, and either the compositional (non-nano) or dimensional (nano) distances as weighting factors (Eq. 4.36). In this case study, the use of two different thresholds does not contribute to the development of better models in terms of predictive power. Thus, as mentioned before the best models are produced using one similarity criterion, $predFactor = 0.6$ and wf_{OF} = 0.001.

[6]Summarized results of 10 runs of the GA workflow are depicted. It is noted that the minimum and maximum values for the different result values are not necessarily produced at the same run.

Table 4.21: Overview of the produced results and statistics from the GA workflow for categorical endpoints applied on the *MeOx ENMs [a]* dataset using one threshold in external validation and wf_{OF} = 0 or wf_{OF} = 0.001.

predFactor: 30%	$wf_{\mathrm{OF}} = 0$		
	min	max	average
thresholds	0.367	0.712	0.589
selected variables	7	10	9
predicted samples	6	6	6
Accuracy	1.000	1.000	-
Sensitivity	1.000	1.000	-
Specificity	1.000	1.000	-
predFactor: 60%			
	min	max	average
thresholds	0.239	0.690	0.557
selected variables	7	9	8.4
predicted samples	3	6	5.7
Accuracy	1.000	1.000	-
Sensitivity	1.000	1.000	-
Specificity	1.000	1.000	-
predFactor: 90%			
	min	max	average
thresholds	0.602	1.090	0.737
selected variables	5	9	7.4
predicted samples	6	6	6
Accuracy	0.667	1.000	-
Sensitivity	0.333	1.000	-
Specificity	1.000	1.000	-

predFactor: 30%	$wf_{\mathrm{OF}} = 0.001$		
	min	max	average
thresholds	0.144	0.476	0.296
selected variables	2	5	3.2
predicted samples	4	6	5.5
Accuracy	0.833	1.000	-
Sensitivity	1.000	1.000	-
Specificity	0.667	1.000	-
predFactor: 60%			
	min	max	average
thresholds	0.218	0.712	0.494
selected variables	2	9	5.3
predicted samples	5	6	5.9
Accuracy	1.000	1.000	-
Sensitivity	1.000	1.000	-
Specificity	1.000	1.000	-
predFactor: 90%			
	min	max	average
thresholds	0.225	0.853	0.658
selected variables	3	8	4.6
predicted samples	5	6	5.9
Accuracy	0.833	1.000	-
Sensitivity	0.667	1.000	-
Specificity	0.667	1.000	-

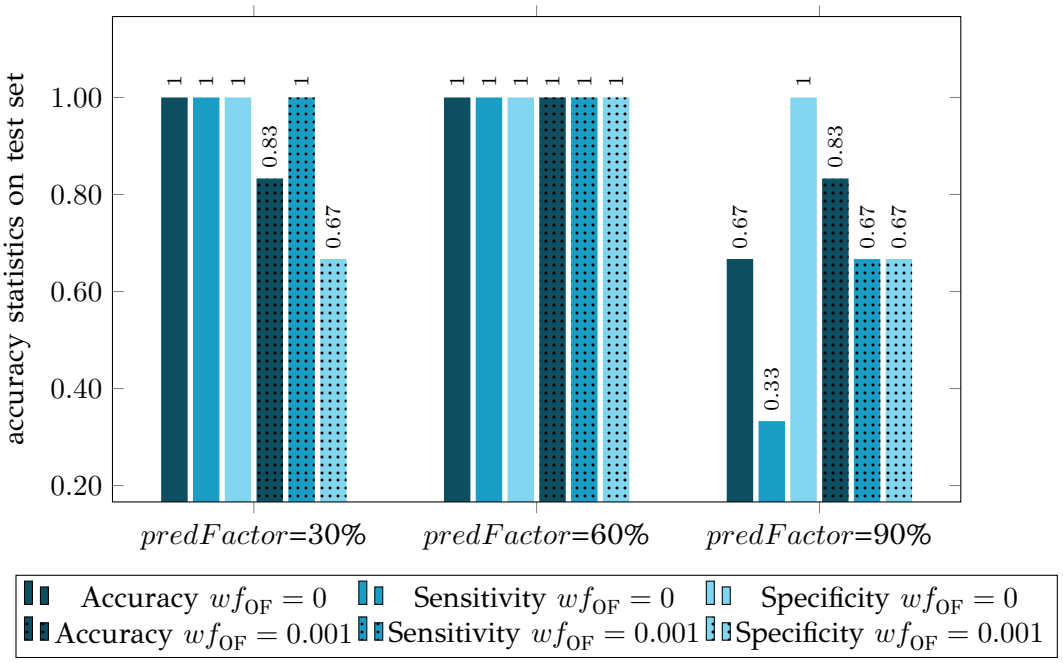


Figure 4.21: Accuracy statistics (on worst-case scenario) per different $predFactor$ levels for wf_{OF} = 0 and wf_{OF} = 0.001 values in the GA scheme.

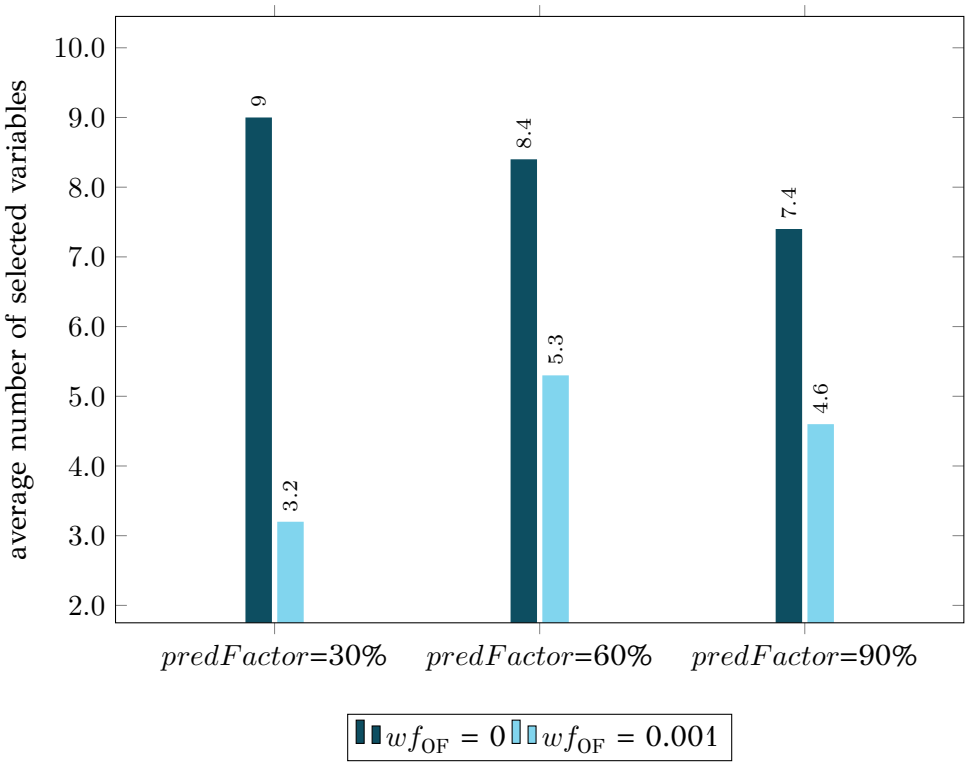


Figure 4.22: Average number of selected variables per different $predFactor$ levels for wf_{OF} = 0 and wf_{OF} = 0.001 values in the GA scheme.

For comparison purposes we applied on the *MeOx ENMs [b]* dataset the *k*NN algorithm through KNIME and `WEKA` suite nodes. Descriptor values were normalised between [0, 1] and the same train-test ratio was used in the Kennard and Stone partition method (0.75). Variable selection was applied before modelling using the `CsfSubsetEval - BestFirst` selection methodology included in `WEKA`. Two variables were used and *k* was tuned to the value of 3.

Table 4.22: Values for the user-defined operational parameters of the proposed read-across GA workflow applied on the *MeOx ENMs [a]* dataset in external validation using two similarity criteria.

Parameter	Value
nChrom	50
maxGenerations	10
initGeneProb	0.6
crossProb	0.7
mutProb	0.01
nonUnf	0.1
$thr_{\min}^{GA}$	0.1
$thr_{\max}^{GA}$	mean value of the maximum distances between samples
bGA	1
predFactor	0.3-0.6-0.9

The prediction accuracy on the test set was equal to 0.833, which is lower to the accuracy of the best read-across models (equal to 1) using one similarity criterion, a *predFactor* of 0.6 and wf_{OF} = 0.001.

4.4.5.3 [r] Additional case studies and comparison with other models and techniques

The developed read-across methodology extended in classification was applied in two more case studies to further demonstrate its applicability, using an external validation scheme.

MeOx ENMs [b] dataset The proposed read-across workflow is demonstrated on the *MeOx ENMs [b]* dataset, which was extracted by the publications of Zhang *et al*. (2012) [83] and Liu *et al*. (2013). [84] The descriptor values were scaled in range [0,1] prior to training.

Table 4.24 summarizes the training specifications, and Table 4.25 the produced results and statistics of the developed models using different regularisation factor values. The metrics are similar, however the inclusion of the regularisation factor has produced a model that uses only 2 descriptors and produces predictions for all test samples, in contrast to the model developed by setting the regularisation factor to zero, which requires 8 descriptors and fails to produce predictions for all ENMs.

Table 4.26 presents the two selected variables for toxicity prediction using the read-across model trained with wf_{OF}=0.05. The "atomic mass of metal" can be easily derived from the Periodic Table, whereas computations of the "chemical hardness" (η, Eq. 4.38) requires only the values of energy of conduction band (E_{c}) and energy of valence band (E_{v}). Table Δ′.4 presents the neighbour ENMs in this case study. In most pairs of neighbour MeOx, metals belongs to the same period.

$$\eta = \frac{E_{\text{c}} + E_{\text{v}}}{2} \tag{4.38}$$

For comparison purposes we applied on the *MeOx ENMs [b]* dataset the *k*NN algorithm. Descriptor values were normalised between [0, 1] and the same train-test ratio was used in the Kennard and Stone partition method (0.66). Variable selection was applied before modelling using the `InfoGainAttributeEval - Ranker` selection methodology included in `WEKA`. Two variables were used (Chemical hardness and MeOx atomization energy) and *k* was tuned to the value of 3. The prediction accuracy on the test set was equal to 0.875, which is equal to the accuracy of the optimised read-across model for $wf_{\text{OF}} = 0.05$.

Table 4.23: Overview of the produced results and statistics from the GA workflow for categorical endpoints applied on the *MeOx ENMs [a]* dataset using two thresholds in external validation and $wf_{OF} = 0$.

		Prediction base					
	predFactor: 30%	non-nano distances			nano distances		
		min	max	average	min	max	average
thresholds	non-nano	0.431	1.683	1.004	0.143	0.718	0.475
	nano	0.261	0.595	0.376	0.210	0.670	0.424
selected variables		5	10	7.6	7	12	9.4
predicted samples		4	6	5	4	6	5.2
Accuracy		0.833	1.000	-	0.800	1.000	-
Sensitivity		1.000	1.000	-	1.000	1.000	-
Specificity		0.667	1.000	-	0.500	1.000	-

		Prediction base					
	predFactor: 60%	non-nano distances			nano distances		
		min	max	average	min	max	average
thresholds	non-nano	0.137	0.637	0.463	0.299	0.670	0.492
	nano	0.435	1.631	1.164	0.529	2.096	1.154
selected variables		8	11	9.0	7	10	8.8
predicted samples		6	6	6	6	6	6
Accuracy		0.833	0.833	-	0.833	0.833	-
Sensitivity		1.000	1.000	-	1.000	1.000	-
Specificity		0.667	0.667	-	0.667	0.667	-

		Prediction base					
	predFactor: 90%	non-nano distances			nano distances		
		min	max	average	min	max	average
thresholds	non-nano	0.114	0.734	0.558	0.558	2.104	0.951
	nano	0.742	1.019	0.887	0.337	1.005	0.829
selected variables		5	11	8.2	5	9	7.4
predicted samples		6	6	6	6	6	6
Accuracy		0.833	0.833	-	1.000	1.000	-
Sensitivity		1.000	1.000	-	1.000	1.000	-
Specificity		0.667	0.667	-	1.000	1.000	-

Table 4.24: GA workflow training parameters for the *MeOx ENMs [b]* read-across model development.

Specifications	
Partitioning method	Kennard & Stone
Training ratio	0.66
# *chromosomes*	50
Generations	1000
Initial variable selection probability	0.6
Uniform *mutation* probability	0.01
Non-uniform *mutation* probability	0.1
Min threshold value	0.1
Max threshold value	mean(max(Dist))
bSA	1
Crossover probability	0.7
predFactor	0.6

Table 4.25: Results of the *MeOx ENMs [b]* read-across models built using the GA workflow. The number of predicted samples and the validation metrics refer to the test set.

Optimised parameters and validation metics		
wf_{OF}	0	0.05
threshold	0.688	0.377
# variables	8	2
# predicted samples	7	8
MCC	0.645	0.745
Accuracy	0.857	0.875
Specificity	0.833	0.833
Sensitivity	1.000	1.000

Table 4.26: Selected variables of the *MeOx ENMs [b]* read-across model using $wf_{\text{OF}} = 0.05$.

Selected variables
Atomic mass of metal
Chemical hardness

SPIONs dataset Finally, our read-across methodology was applied in the *SPIONs* dataset [101] as presented in §3.7. All descriptor values were scaled in range [0,1] prior to training. Due to the small size of the *SPIONs* dataset, one read-across model was trained, using a regularisation factor wf_{OF} equal to zero. Table 4.27 summarizes the training specifications, and Table 4.28 the produced results and validation statistics of the developed model.

Table 4.29 presents the selected variables of the SPIONs model. It is observed that two out of the three selected variables, namely the "Magnetic core" and the "Overall size", are also considered as the two important variables in the study of Kotzabasaki *et al*. Table Δ′.1 presents the neighbour ENMs for this particular case study. It is observed that all neighbouring SPIONs share the same core (either magnetite or maghemite). It addition, in most cases, neighbour ENMs belong to the same hierarchical cluster according to the clustering analysis of Kotzabasaki *et al*. Therefore we can conclude that there is an agreement between the two studies.

Table 4.27: GA workflow training parameters for the *SPIONs* read-across model development.

Specifications	
Partitioning method	Kennard & Stone
Training ratio	0.66
# *chromosomes*	50
Generations	100
Initial variable selection probability	0.6
Uniform *mutation* probability	0.01
Non-uniform *mutation* probability	0.1
Min threshold value	0.1
Max threshold value	mean(max(Dist))
bSA	1
Crossover probability	0.7
wf_{OF}	0
predFactor	0.6

Table 4.28: Results of the *SPIONs* read-across models built using the GA workflow. The number of predicted samples and the validation metrics refer to the test set.

Optimised parameters and validation metrics	
threshold	0.426
# variables	3
# predicted samples	5
MCC	0.667
Accuracy	0.800
Specificity	1.000
Sensitivity	0.667

For comparison purposes we applied on the *SPIONs* dataset the *k*NN algorithm. Descriptor values were normalised between [0, 1] and the same train-test ratio was used in the Kennard and Stone partition method (0.66). All available variables were used in modelling and *k* was tuned to the value of 3. The prediction accuracy on the test set was equal to 0.8, which is equal to the accuracy of the optimised read-across model, but it is achieved using more variables.

Table 4.29: Selected variables of the *SPIONs* read-across model built using the GA workflow.

Selected variables
Magnetic core
Magnetic field strength - B_0 (T)
Overall size (nm)

4.5 Apellis: an online tool for read-across model development

The grouping/read-across workflow described above, has been integrated in the Apellis R `shiny` web application (`https://apellis.jaqpot.org/`). The application is also available through a Docker Hub: `hub.docker.com/r/demetradanae/apellis`. Users of the Apellis tool can develop and validate predictive read-across models and can apply the produced models to reliably predict toxicity-related endpoints of ENMs or other chemicals.

The application consists of two main tabs for model development using one or two similarity criteria respectively. Each of the model development tabs consists of sub-tabs for training and use of the model:

Read-across training In this part, the users can train a toxicity-predictive model through a GA workflow as described in detail in §4.3.2 and 4.4. An external validation approach is used to test the proposed read-across methodology, by dividing the full dataset into training and test subsets. This data partitioning can be achieved either by applying a random partition or a partition method (e.g. Kennard-Stone). [67] The training set is used in the GA workflow and determines the optimal set of descriptors and threshold(s) values. Using read-across technique a table that contains all test ENMs with a successful prediction for the toxicity index is presented, a diagram of test ENMs with their neighbours, as well as diagrams and tables containing information for the optimised parameters and model's accuracy.

- Required specifications: This tab contains a series of operational parameters concerning the evolutionary process and the OF, that should be tuned by the users.
- Probabilities: This tab contains all the necessary probability values that should be tuned for the evolutionary process.

Obtain predictions In this part, the users can either use the trained model from the previous part right after its development, or they can upload and use a model that was trained and saved any-time, for the toxicity prediction of a set of ENMs with unknown toxicity index. After prediction process a table containing the predicted toxicity value for all the unknown ENMs is presented along with the ENMs neighbour diagram.

In order to support and facilitate the use of the Apellis application by the nanosafety community, particular emphasis has been given to the development of documentation material and tutorials. These include a detailed user-guide, a video tutorial, quick-start guides and informative images. The documentation and training material is freely available through the web application.

In this section we are firstly presenting the important technicalities of Apellis' development and in continuation, we are presenting the functionalities of the application in order to train and use a read-across model for both numerical and categorical endpoints using one similarity criterion. Same rules apply for model training using two similarity criteria.

4.5.1 [m] Web Implementation of the grouping/read-across workflow

The initial approach for the implementation of the GA workflow, described in detail in §4.3.2 and 4.4, and the corresponding training results collection were performed in MATLAB®. However, with our research we aspire to facilitate the advancement in the field of read-across toxicity assessment by providing our projects to all interested stakeholders -programming experts or not- in order to use them in real-life applications. To achieve this, we developed Apellis application that gives access to model development important features and results through a user-friendly environment, and we released it through Jaqpot under GNU General Public License v3. Stakeholders can access freely the application through `apellis.jaqpot.org`, train and use a read-across model without any need of advanced programming skills.

Apellis web application was developed in R, using `shiny` v.1.4.0 package. Necessary alterations have been made in order to adapt the original MATLAB® (GitHub repo: `github.com/DemetraDanae/optimized-read-across`) code to R particularities. The principal components of the R code and the important packages are presented in the next paragraphs.

Apart from the GA methodology itself, an element of great significance was the global user-experience. Therefore, through the `ui.R` file we incorporated all the necessary parameters that create an elegant and intuitive environment so that the user is "driven" by this exact same environment to easily built a model. The application consists of four tabs for different cases of modelling requirements (numerical/categorical endpoint, one or two similarity criteria) and a landing page (*Home*), using especially for this case the `shinyLP` package. In order to lead-help the user to insert the necessary data for the analysis, some features are enabled/disabled or hidden/shown, and pop-up informative/warning messages appear according to the provided information or the stage of the analysis. To do so, the `shinyjs` and the `shinyalert` packages were used. The aesthetic of Apellis was improved using `shinythemes` package for the overall appearance of the app, `shinycustomloader` and `shinyWidgets` for the loading graphics during execution and, `DT`, `plotly` and `ggplot2` were used to create data tables and graceful plots that facilitate results presentation.

The code that builds the read-across model in each of the four tabs as long as the auxiliary processes concerning the results presentation and deployment are included in the `server.R` file. The main processes concerning the evolutionary process, the neighbour selection, the prediction and the score calculation are incorporated as global functions. The particular processes concerning each tab are incorporated in controlled-reactive environments.

4.5.1.1 Deployment

In order to "package" Apellis application and deliver it in any server without any dependency impediments, we used Docker (see §A′.1.5) to develop a portable version of it. All the necessary actions for dockerising this `shiny` application were performed on Linux operating system (OS).

The fundamental element for the creation of a docker *image* is the *Dockerfile*, which is a simple text file including all the necessary instructions for the creation of the *image*. In order to build the Docker *image* from the *Dockerfile* all the necessary script files (e.g. `ui.R` and `server.R`), the necessary sub-folders containing the dependencies of the application (e.g. `www`), the *Dockerfile* and other auxiliary files must be included in the same directory, as in Listing 4.1.

The basic structure of the Apellis *Dockerfile* is presented in Listing 4.2. *Images* are built in "layers" thus in this case, the *Dockerfile* was based on the `rocker/shiny-verse:latest` source *image*. In continuation the required R packages in order to run the application were installed along with the necessary Linux and external R dependencies. The port of the app deployment could be also declared in the *Dockerfile*.

```
.
|-- ui.R
|-- server.R
|-- www
| |-- User guide.pdf
| |-- logo.png
| |-- demo.csv
|-- Dockerfile
```

Listing 4.1: Directory structure containing all the necessary files for an applications's *dockerimage* building.

Once the *Dockerfile* was ready, from terminal we navigated to the directory containing the application files, we created the *image* using `docker build -t apellis .` and later we created an executable container using `docker run --rm -p 3838:3838 apellis`. Finally, we *pushed* the Apellis docker *image* to Docker Hub.

```
# get shiny serves plus tidyverse packages image
FROM rocker/shiny-verse:latest

# system libraries of general use
RUN apt-get update && apt-get install -y \
    sudo \
    pandoc \
    pandoc-citeproc \
    libcurl4-gnutls-dev \
    libcairo2-dev \
    libxt-dev \
    libssl-dev \
    libssh2-1-dev

# install required R packages
RUN R -e "install.packages(c('shiny', 'shinyjs', 'DT', 'rhandsontable','prospectr',
    'shinythemes', 'extrafont', 'ggplot2', 'RColorBrewer', 'plotly', 'shinyWidgets',
     'shinyalert', 'rapportools', 'shinyLP', 'shinycustomloader', 'caret', 'e1071',
    'mltools'), repos='http://cran.rstudio.com/')"

# copy the app to the image
COPY ui.R /srv/shiny-server/
COPY server.R /srv/shiny-server/
COPY www /srv/shiny-server/www

# select port
EXPOSE 3838

# allow permission
RUN sudo chown -R shiny:shiny /srv/shiny-server

# run app
CMD ["/usr/bin/shiny-server.sh"]
```

Listing 4.2: *Dockerfile* basic structure for a `shiny` application deployment.

Interested users, that already have installed Docker in their system, can *pull* the *image* from the following link: `hub.docker.com/r/demetradanae/apellis` (Listing 4.3). Later, as presented in Listing 4.4 users can run Apellis application locally.

```
docker pull demetradanae/apellis
```

Listing 4.3: Command for Apellis download through Docker.

```
docker run --rm -p 3838:3838 demetradanae/apellis
```

Listing 4.4: Command for local execution of Apellis through Docker.

4.5.2 [m] The Apellis application

Apellis offers two alternatives for model training depending on the available data. In case that only one type of descriptors is available *Numerical single criterion* tab (*Class single criterion* tab for categorical endpoints) must be used, whereas in case of two different types of available descriptors *Numerical multiple criteria* tab (*Class multiple criteria* tab for categorical endpoints) should be used however, single criterion tabs can also be used and all descriptors will be treated as one category. Apellis' homepage (Figure 4.23) includes all the necessary information about its functionality (quick-use guides, a detailed user guide, a video tutorial, information about maintenance and license).

4.5.2.1 Numerical endpoints

The *Numerical single criterion* part (Figure 4.24) can be used for training and use of a read-across model for the prediction of toxicity endpoints and other properties of ENMs, when only one type of descriptors is available (or in case that users aspire to treat the descriptors as one category).

Read-across training In this section training of a read-across model is performed according to the provided dataset and a set of specifications defined by the user. A sequential workflow, as described in §4.3.2, is followed in order to find the optimal grouping hypothesis for the endpoint estimation (selected descriptors and threshold(s) values).

Specifications Users must upload one CSV file containing the dataset of interest by clicking on the *Browse* button in the *Upload a dataset* field. The file must contain the values of available descriptors (in columns separated by commas "," -otherwise an internal error may occur) and the values of the toxicity index (2nd column), which will be predicted by the model. In the 1st column the ENMs names should be listed. Missing values cannot be handled by this approach. In addition, columns containing the same value in all rows cannot be handled and must be deleted by the input file. An exemplary input file can be seen in Figure 4.25. The *Gold ENMs* dataset (see Chapter 3) is provided as demo for demonstration reasons.

The user has the option to scale the data in the range of [0,1] (Eq. 2.1) by clicking on *Scaling of raw data*.

For validation purposes (formation of training and test sets), users must choose a the partitioning method (Kennard-Stone as implemented in *prospectr* R library or random) and the corresponding training:test ratio. The users must initialize the hyper-parameters for the evolution of the algorithm presented in Table 4.30.

Table 4.30: Apellis' user-defined specifications.

Apellis **specification**	**Corresponding GA parameter**	**Accepted values**
Number of *chromosomes*	*nChrom*	positive even numbers
Number of *generations*	*maxGenerations*	positive numbers
Number of training samples with a prediction	*predFactor*	0-1
regularisation parameter for variable selection	reg_{GA}	0-1

Probabilities In continuation the users must tune a series of operational parameters of the GAs (Figure 4.26) that are directly linked to the biological processes of *selection*, *crossover* and *mutation* of *genes* (Table 4.31).

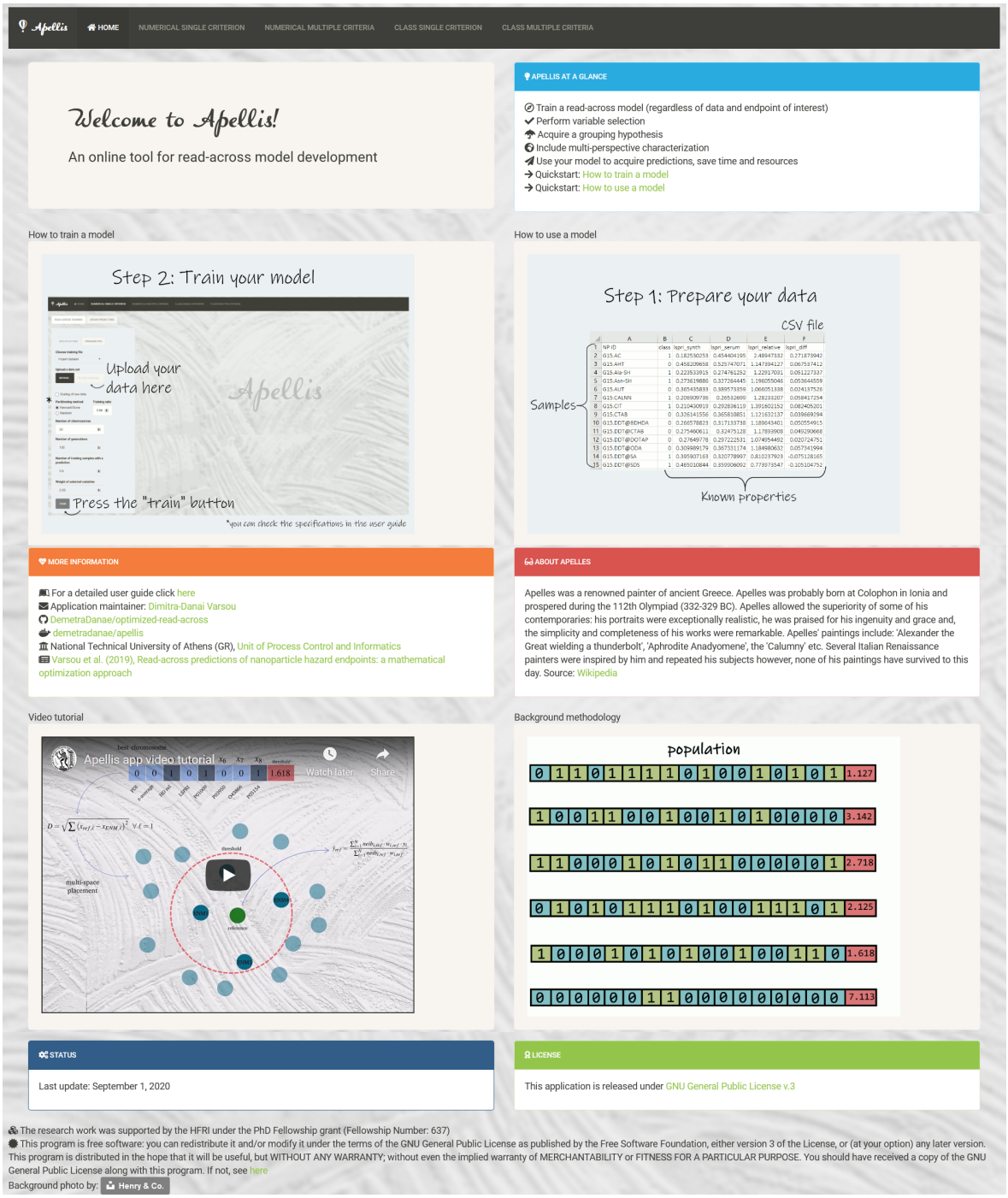


Figure 4.23: Apellis' landing page. From top ribbon users can chose numerical/class single criterion or multiple criteria model development according to the available data. The landing page includes a summary of the application functionalities and two quick-start guides in *Apellis at a glance* box and useful information about its use in *Help* box. The *Status* box includes the last update date and the *License* box the link towards the license file. In addition, two gifs explain in a few steps how to train and use a read-across model through the app. Finally, a video tutorial for Apellis' use and a gif explaining the background methodology are presented in the landing page.

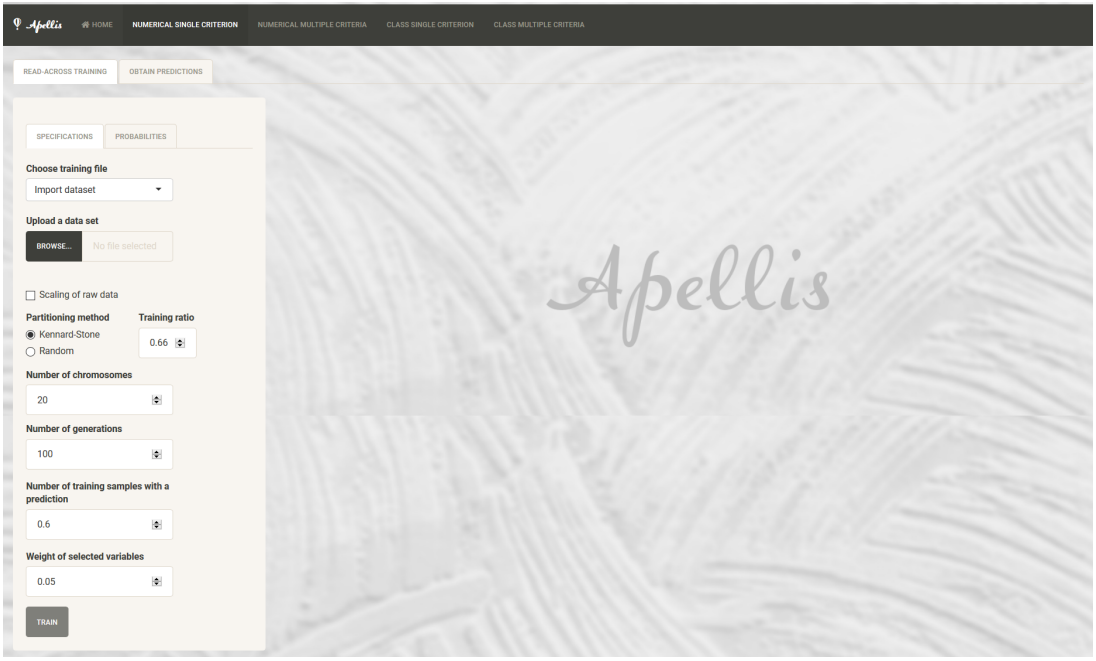


Figure 4.24: Apellis *Numerical single criterion tab*. On the right part of the interface the specifications that must be tuned are displayed and in the main part the training results are going to be displayed.

Table 4.31: Apellis' user-defined probability values.

Apellis **specification**	**Corresponding GA parameter**	**Accepted values**
Favor variable selection in 1st *generation*	*initGeneProb*	0-1
Crossover between *chromosomes*	*crossProb*	0-1
Uniform *mutation*	*mutProb*	0-1
Non-uniform *mutation*	*nonUnf*	0-1
Freezing parameter	*bGA*	positive numbers

Concerning the non-uniform *mutation*, the lower threshold value is always equal to 0.1, while the upper threshold value is equal to the mean value of the maximum distances between the provided samples.

Users are advised to select the default values for the hyper-parameters and only if the results are not satisfactory, apply different combinations. The number of *generations* affects the computational time required by the algorithm and should be increased if the algorithm is not converging using the default value.

If a dataset is uploaded or the demo dataset is selected, by pressing the *Train* button, the model training starts, according to the described workflow, otherwise the corresponding button remains disabled till all necessary files are provided. During training, a process bar is presented, indicating the number of processed *generation*.

Training results After the training procedure has been completed, the application returns the results to the user in the form of automatically created plots and tables. A scatter-plot depicting the actual and the predicted toxicity endpoint for all training and test ENMs is presented. Users can zoom in specific areas of the plot, hover over the depicted points and compare the actual (experimental) and the predicted endpoint values.

A table containing the actual and the predicted endpoint values for the test set is also presented (*Experimental vs. predicted endpoint values for the test set*). The prediction values have the same units as the actual values.

	A	B	C	D	E	F	G	H	I	J	K
1	NP ID	net.c	class	lspri_synth	lspri_serur	lspri_relati	lspri_diff	lspri_rel_c	zav_synth	zav_serum	pdi_synth
2	G15.AC	-5.1839	1	0.18253	0.454404	2.489473	0.271874	1.489473	22.36	57.53	0.084
3	G15.AHT	-1.00854	0	0.45821	0.525747	1.147394	0.067537	0.147394	30.95	90.06	0.399
4	G15.Ala-SH	-5.50439	1	0.223534	0.274761	1.22917	0.051227	0.22917	22.64	44.43	0.147
5	G15.Asn-SH	-5.67669	1	0.27362	0.327264	1.196055	0.053645	0.196055	23.09	37.75	0.15
6	G15.AUT	-1.31567	0	0.365436	0.389573	1.066051	0.024138	0.066051	23.8	55.98	0.326
7	G15.CALNN	-7.13797	1	0.20691	0.265327	1.282332	0.058417	0.282332	25.22	38.8	0.144
8	G15.CIT	-5.41982	1	0.210431	0.292836	1.391602	0.082405	0.391602	18.65	54.03	0.138
9	G15.CTAB	-5.86229	0	0.326142	0.365811	1.121632	0.039669	0.121632	15.6	59.7	0.465
10	G15.DDT@BDHDA	-7.29449	0	0.266579	0.317134	1.189643	0.050555	0.189643	23.15	47.03	0.187
11	G15.DDT@CTAB	-7.59005	0	0.275461	0.324751	1.178939	0.049291	0.178939	20.53	48.4	0.191
12	G15.DDT@DOTAP	-1.12756	0	0.276498	0.297223	1.074954	0.020725	0.074954	28.17	47.34	0.193
13	G15.DDT@ODA	-6.1218	0	0.309989	0.367331	1.184981	0.057342	0.184981	33.6	58.95	0.268
14	G15.DDT@SA	-6.8039	1	0.395907	0.320779	0.810238	-0.07513	-0.18976	82.41	59.93	0.249
15	G15.DDT@SDS	-7.67595	1	0.465011	0.359906	0.773974	-0.1051	-0.22603	27.94	100.13	0.302
16	G15.DTNB	-6.08314	1	0.241281	0.413864	1.715281	0.172583	0.715281	23.58	60.27	0.108
17	G15.F127	-5.36112	1	0.175809	0.244867	1.392798	0.069057	0.392798	42.05	46.63	0.216
18	G15.Gly-SH	-4.97528	1	0.261762	0.402932	1.539304	0.141169	0.539304	77.02	55.39	0.211
19	G15.HDA	-0.27033	0	0.243839	0.275642	1.130427	0.031803	0.130427	28.04	54.89	0.194
20	G15.LA	-5.96398	1	0.236782	0.30131	1.272521	0.064528	0.272521	22.45	48.09	0.23
21	G15.MAA	-6.14203	1	0.281334	0.448587	1.594501	0.167253	0.594501	30.74	60.99	0.273
22	G15.MBA	-5.38142	1	0.293044	0.504305	1.720917	0.211261	0.720917	29.87	67.7	0.321
23	G15.MES	-3.19932	1	0.226324	0.476291	2.104465	0.249967	1.104465	49.22	61.86	0.297
24	G15.Met-SH	-5.9284	1	0.215656	0.28201	1.307686	0.066354	0.307686	23.19	52.42	0.167
25	G15.MHA	-5.73543	1	0.278335	0.389208	1.398344	0.110873	0.398344	25.15	58.99	0.239
26	G15.MHDA	-5.77833	1	0.242059	0.294179	1.215319	0.05212	0.215319	23.12	43.62	0.076
27	G15.MPA	-5.39593	1	0.303644	0.451142	1.485761	0.147498	0.485761	25.73	55.65	0.284
28	G15.MSA	-6.10482	1	0.255577	0.424225	1.659871	0.168648	0.659871	34.33	52	0.203

Figure 4.25: An exemplary template file for Apellis model training (numerical endpoint). The first column contains the ENMs ID, the second the endpoint that is going to be predicted, followed by the rest of descriptors in the rest of the columns.

In addition, the app produces two tables one containing the optimal set of descriptors selected by the model during the evolutionary process, and one containing information about the trained model (optimised threshold, number of the *generation* that produced the model, the score calculated over the test set -Eq. 4.28, the total number of test samples with a prediction, the total number of selected variables, the MSE, the MAE and the Q^2_{ext} statistic -Eqs. 4.26, 4.32 and 4.31 respectively).

Finally, the app produces the ENMs' *Neighbours heatmap*, where the neighbours of the test ENMs in the training set are depicted in color code. In this heatmap for each pair of training-test ENMs, a value of 1 (red) is used to denote that two ENMs are neighbours, or a value of 0 (beige) is used in the opposite case. The neighbours heat map actually gives a graphical representation of the ENMs grouping.

Users can download all training results and the trained model for future use in the following tab (*Obtain predictions*) without any need for constant training with the same data. An example of training is presented in Figure 4.27.

Using the model for obtaining predictions This part of the application allows users to apply already developed read-across models to compute endpoint predictions for one or more untested ENMs (Figure 4.28).

If a model is already available from a previous training, the users can click on *Upload model* and import an adequate model file by clicking on *Browse* button of the *Upload an adequate model file* field. When the model is uploaded a *Template input file* can be downloaded in order to be filled-in by the users with the necessary descriptor/variables values that will be used for the endpoint prediction. The necessary variables are also presented, as well as the endpoint that will be predicted, the model title and the Q^2_{ext} statistic from external validation.

This part can be also used right after the training of the model. In that case the *Upload*

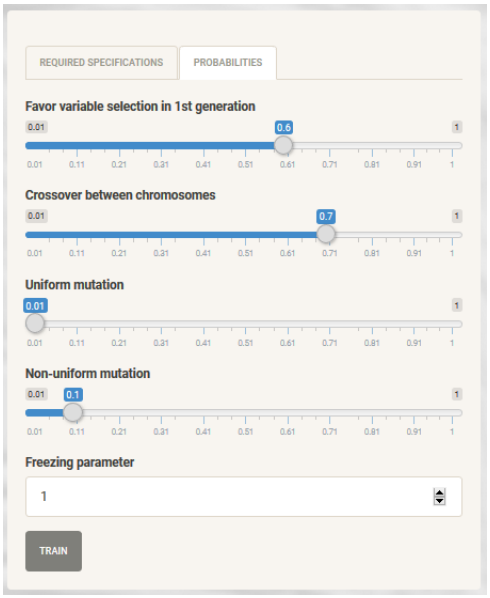


Figure 4.26: Apellis read-across training sub-tab for the probabilities tuning presented in Table 4.31. The probabilities' tab is the same in training using single or multiple criteria for numerical or categorical endpoints.

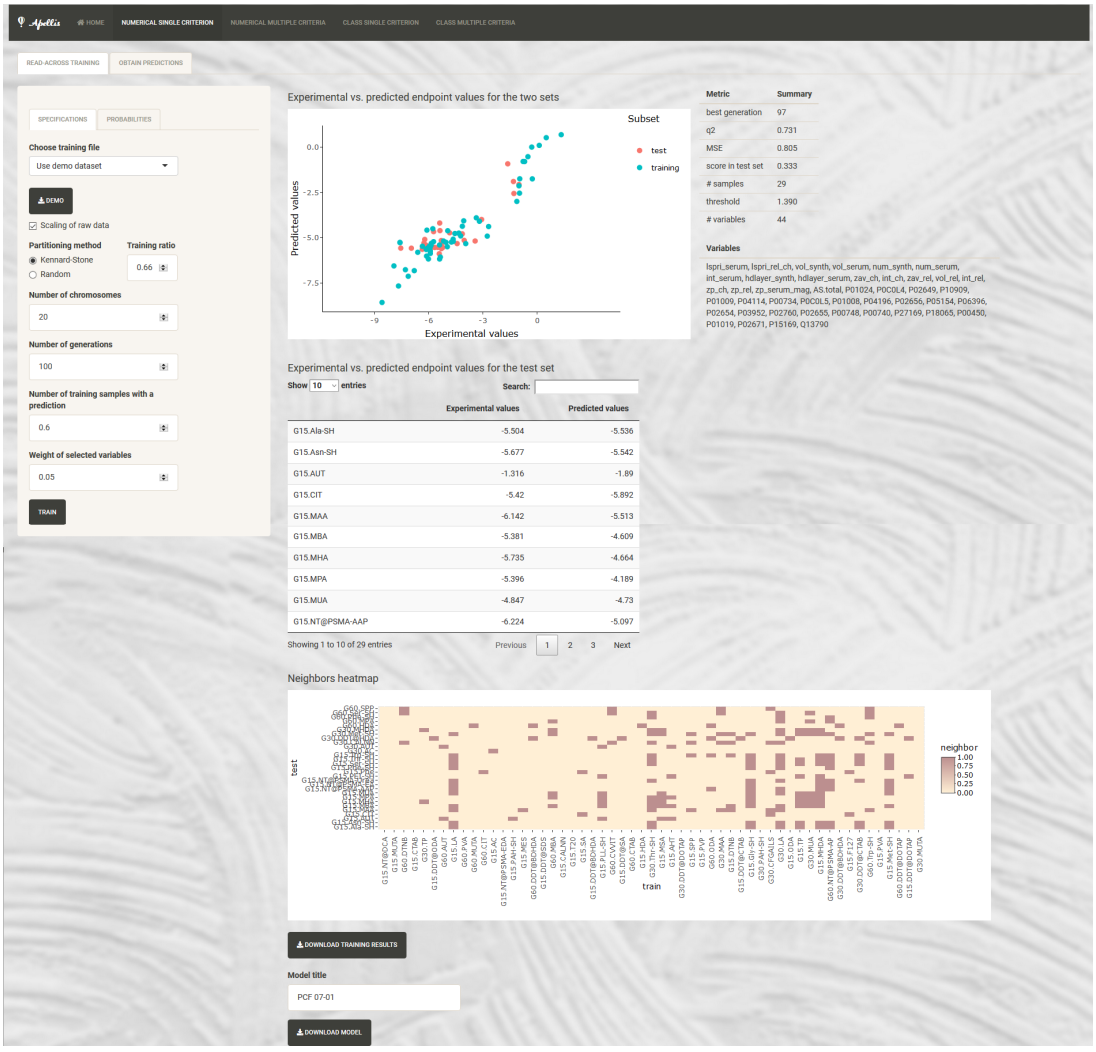


Figure 4.27: Apellis *Numerical single criterion* tab training results.

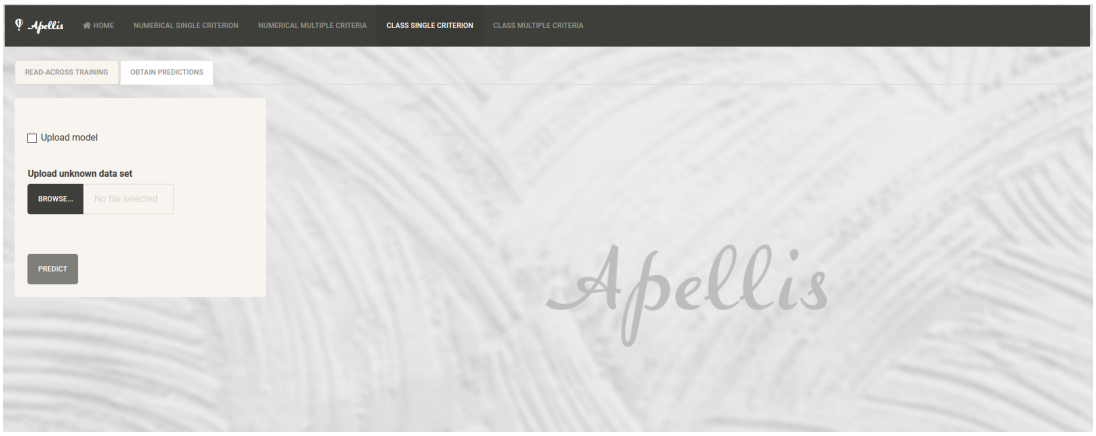


Figure 4.28: Apellis *Numerical single criterion* tab for model use. This interface is similar for the *Numerical multiple criteria*, *Class single criterion* and *Class multiple criteria* tabs.

model field remains disabled. The user just needs to upload a CSV file containing the input descriptors to the read-across model for the untested ENMs, either according to the template file (in case that a model is uploaded) or according to the training file (omitting the 2nd column). It is necessary to provide values only for the selected descriptors, however the columns of the non-selected variables must not be deleted; it is advisable to be filed with miscellaneous values.

By clicking on *Predict* button, the prediction process begins. The analysis produces that contains the predicted toxicity index value for all untested ENMs. In case that no neighbours are found in the training set, no predictions are produced. The ENMs *Neighbours heatmap*, depicting the neighbours of the untested ENMs in the training set is also produced. All results can be downloaded in ZIP format by clicking on *Download prediction results*. An example of model use can be found in Figure 4.29.

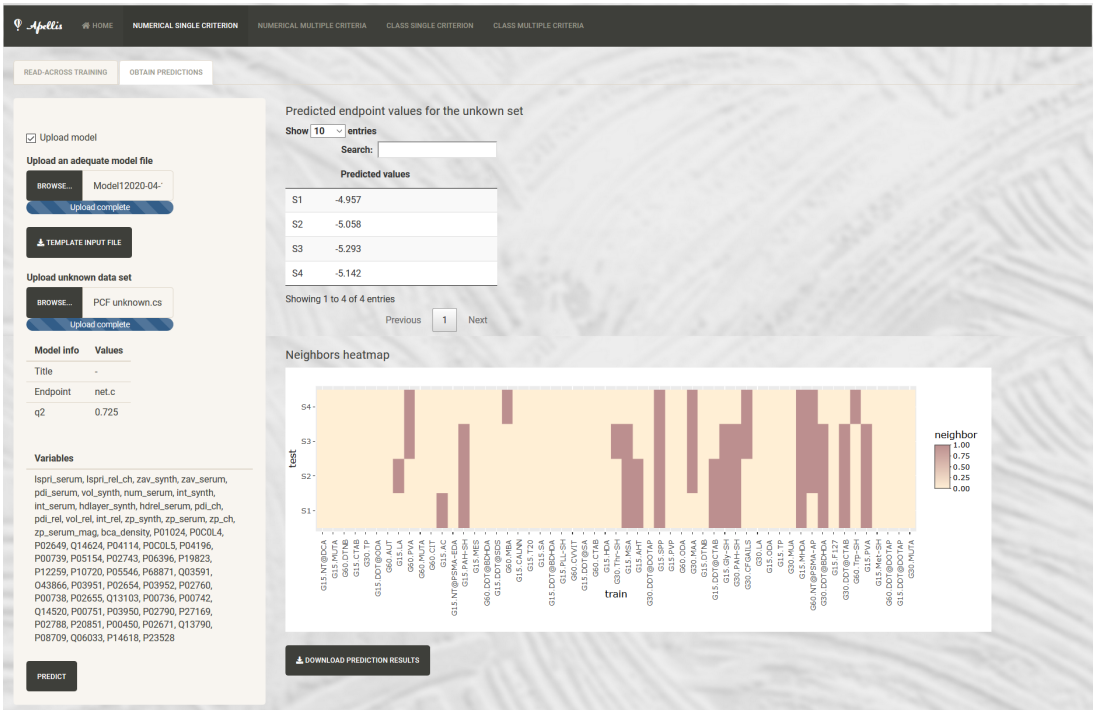


Figure 4.29: Apellis *Numerical single criterion* tab where a developed model is used to study untested ENMs. In this example an already trained model is uploaded to the app. Its details can be found in the right side of the interface. The predictions for an artificial dataset of gold NPs are presented on the left side of the interface.

4.5.2.2 Categorical endpoints

Similarly to numerical endpoints, users can train and use a categorical-endpoint read-across model from the adequate tabs. The *Class single criterion* part can be used for training and using a read-across model for the prediction of toxicity class of ENMs, when only one type of descriptors is available (or users aspire to treat the descriptors as one category).

Read-across training In this section training of a read-across model is performed according to the provided dataset and the set of specifications defined by the user. The interface is the same as for numerical endpoints training. The *MeOx ENMs [b]* dataset is provided as demo for demonstration reasons.

Specifications The *Specifications* part is similar to the *Numerical single criterion* part (Figure 4.24). Users must upload one CSV file containing the dataset of interest by clicking on the *Browse* button in the *Upload a dataset* field. The file must contain the values of available descriptors (in columns separated by commas ",") and the class (TRUE/FALSE) of the toxicity index (2nd column), which will be predicted by the model. In the 1st column the ENMs names should be listed. Missing values cannot be handled by this approach. In addition, columns containing the same value in all rows cannot be handled and must be deleted by the input file. An exemplary input file can be seen in Figure 4.30.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	NP	Toxicity	dp	EC	EV	EAmz	χMeO	ΔHsub	ΔHIE	ΔHsf	ΔHLat	ΔHIE	Z2/r	IEP	ZP
2	ZnO	TRUE	22.6	-3.891	-7.198	7.546	5.674	1.351	28.71	-3.608	42.928	9.394	0.0667	9.6	28.8
3	CuO	TRUE	12.8	-5.174	-6.515	7.719	5.874	3.497	31.515	-1.609	42.856	7.726	0.0548	7.9	7.6
4	Mn2O3	TRUE	51.5	-4.647	-7.635	11.709	5.919	2.936	59.677	-9.917	156.975	7.434	0.1552	3.7	-46.1
5	CoO	TRUE	71.8	-4.424	-6.832	9.454	5.735	4.422	29.387	-2.476	39.767	7.881	0.0615	9.2	21.6
6	Co3O4	TRUE	10	-4.593	-7.025	10.755	5.927	4.422	46.137	-9.38	99.573	7.881	0.1329	9.4	24.6
7	Cr2O3	TRUE	193	-4.439	-7.524	13.92	5.858	4.12	58.331	-11.717	158.322	6.767	0.1452	5.3	-32.6
8	Ni2O3	TRUE	140.6	-4.309	-7.688	11.709	6.052	4.458	65.455	-5.073	164.157	7.639	0.1607	8.3	32.2
9	Gd2O3	FALSE	43.8	-2.825	-8.102	16.782	5.499	4.12	42.989	-18.82	134.692	6.15	0.0957	8	6.5
10	In2O3	FALSE	59.6	-3.632	-7.322	11.188	5.583	2.518	55.204	-9.606	144.351	5.786	0.1125	9.2	61.9
11	CeO2	FALSE	18.3	-3.803	-7.45	20.121	5.65	4.354	77.697	-11.284	99.775	5.539	0.1649	7.8	21.4
12	SiO2	FALSE	13.5	-2.018	-11.118	18.734	6.19	4.664	107.795	-9.41	136.029	8.151	0.6154	1	-31.8
13	Al2O3	FALSE	14.7	-1.515	-9.815	15.872	5.665	3.429	56.691	-17.345	164.955	5.985	0.1667	7.4	0
14	Y2O3	FALSE	32.7	-2.352	-8.201	17.433	5.406	4.402	43.362	-19.748	131.676	6.217	0.1	9.6	42.7
15	SnO2	FALSE	62.4	-4.013	-8.013	14.397	5.812	3.122	96.334	-5.986	122.369	7.344	0.2319	4	-38.8
16	TiO2	FALSE	12.6	-4.161	-7.491	19.775	5.767	4.902	96.063	-9.779	125.924	6.828	0.2623	6.4	-19.4
17	ZrO2	FALSE	40.1	-3.192	-8.233	22.723	5.618	6.322	83.379	-11.252	115.954	6.634	0.1905	5.8	-12.8
18	Fe2O3	FALSE	12.3	-4.993	-6.987	12.489	5.978	4.306	59.047	-8.512	148.3	7.903	0.1636	7.2	-2.1
19	Sb2O3	FALSE	11.8	-3.645	-8.138	10.408	5.514	2.74	53.278	-7.346	142.071	8.608	0.1184	1	-35.3
20	HfO2	FALSE	28.4	-2.956	-8.371	23.938	5.705	6.409	84.863	-1.17	104.812	6.825	0.1928	8.1	33.5
21	WO3	FALSE	16.6	-5.532	-8.586	24.978	6.64	8.82	213.421	-8.734	250.324	7.864	0.6	0.3	-61.3
22	Yb2O3	FALSE	61.7	-2.831	-7.933	15.091	5.429	1.613	45.092	-18.807	138.672	6.254	0.0909	8.2	9.9
23	La2O3	FALSE	24.6	-2.38	-8.147	17.433	5.378	4.467	40.28	-18.668	129.054	5.577	0.0776	9.4	54.3
24	NiO	FALSE	13.1	-3.57	-7.445	9.454	5.744	4.458	30.266	-2.494	40.503	7.639	0.058	11.4	27.6

Figure 4.30: An exemplary template file for Apellis model training (categorical endpoint). The first column contains the ENMs ID, the second the class (TRUE/FALSE) that is going to be predicted, followed by the rest of descriptors in the rest of the columns.

Again, the user can select if the provided data should be normalised or not, the partitioning method for external validation and the corresponding training:test ratio. The user must also initialize some parameters for the evolution of the algorithm as before (Table 4.30).

Probabilities The operational parameters of the GAs for the *selection*, *crossover* and *mutation* of genes must also be tuned (Table 4.31, Figure 4.26).

Concerning the non-uniform *mutation*, the lower threshold value is always equal to 0.1, while the upper threshold value is equal to the mean value of the maximum distances between provided samples.

If a dataset is uploaded or the demo dataset is selected, by pressing the *Train* button, the model training starts, according to the described workflow, otherwise the corresponding

button remains disabled. During training a process bar is presented, indicating the number of processed *generation*.

Training results After the completion of training, a confusion matrix (Table 2.2) is presented depicting the validation results (TP, TN, FP, FN frequencies for the test set). The experimental and the predicted endpoint values for the test ENMs, are also presented in the produced table entitled *Experimental vs. predicted endpoint class for the test set*.

In addition, the app produces two tables one containing the optimal selected variables and one containing information about the trained model (optimised threshold, number of the *generation* that produced the best solution, the score calculated over the test ENMs -Eq. 4.28, the total number of test samples with a successful prediction, the total number of selected variables and the accuracy, sensitivity and specificity statistics, and MCC -Eqs. 2.23, 2.21, 2.22 and 2.24 respectively). Finally, the app produces the ENMs' *Neighbours heatmap*.

All training results and the produced read-across model, can be downloaded for future use in the application. An example of training is presented in Figure 4.31.

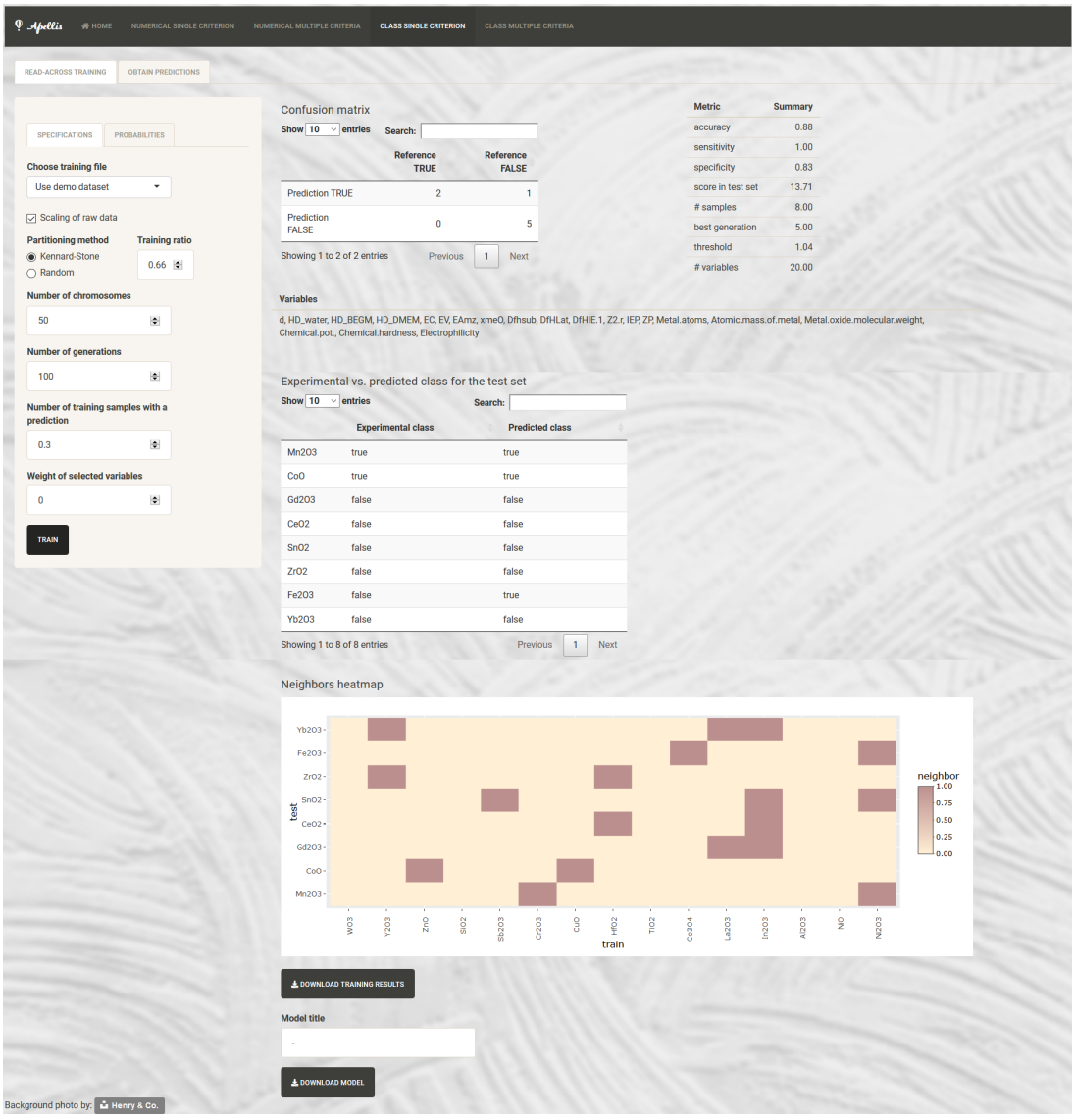


Figure 4.31: Apellis *Class single criterion* tab training results, including the confusion matrix for the test set, the corresponding accuracy metrics, the predicted class for each test sample and the neighbours heatmap.

Using the model for obtaining predictions This section of the application can be used right after the read-across training or after the input of an already trained model by Apellis (the interface is similar to the tab presented in Figure 4.28).

If a model is already available from a previous training, the users can click on *Upload model* and import an adequate model file by clicking on *Browse* button of the *Upload an adequate model file* field. When the model is uploaded a *Template input file* can be downloaded in order to be filled-in by the users with the necessary descriptor/variables values that will be used for the endpoint prediction. The necessary variables are also presented, as well as the endpoint that will be predicted, the model title and the accuracy statistic from external validation.

This part can be also used right after the training of the model. The user must upload a CSV file in the application containing the descriptor values for the untested ENMs dataset, either according to the template file (in case that a model is uploaded) or according to the training file (omitting the 2nd column). It is necessary to provide values only for the selected descriptors, however the columns of the non-selected variables must not be deleted; it is advisable to be filed with miscellaneous values.

By clicking on *Predict* button, the prediction process begins. The analysis produces a table that contains the predicted value of toxicity index for all the provided ENMs, and the ENMs *Neighbours heatmap*. All results can be downloaded in ZIP format by clicking on *Download prediction results*. An example of model use can be found in Figure 4.32.

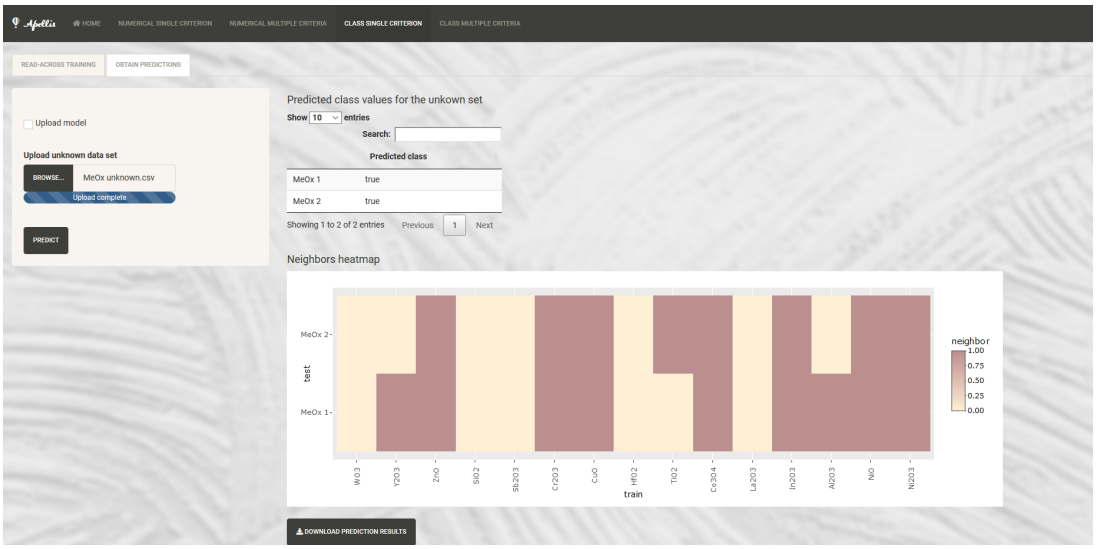


Figure 4.32: Apellis *Class single criterion* tab for model use. In this example predictions for a set of samples with unknown class endpoints, are acquired directly after the model train. The predictions for an artificial dataset of MeOx ENMs are presented on the left side of the interface.

4.6 Chapter summary

In this Chapter a robust read-across methodology is presented for the prediction of numerical or categorical toxicity-related endpoints of ENMs based on mathematical optimisation. With this method the following are achieved:

- Optimal read-across grouping based on similarities between ENMs
- Automation of the grouping procedure without any need of the formulation of a prior grouping hypothesis
- Automated prediction production of either numerical or categorical endpoints based on similar ENMs (Figure 4.33)

- Inclusion of the multi-perspective characterisation of ENMs in grouping and prediction procedures
- Automated variable selection
- Control of the total number of selected variables (in order to prevent over-fitting) by the inclusion of a regularisation parameter.

To demonstrate the efficiency of the method it was applied on two datasets, one including a numerical endpoint and one including a categorical endpoint. Both datasets contain two types of descriptors, thus it was possible to study the use of more than one similarity criteria during model development.

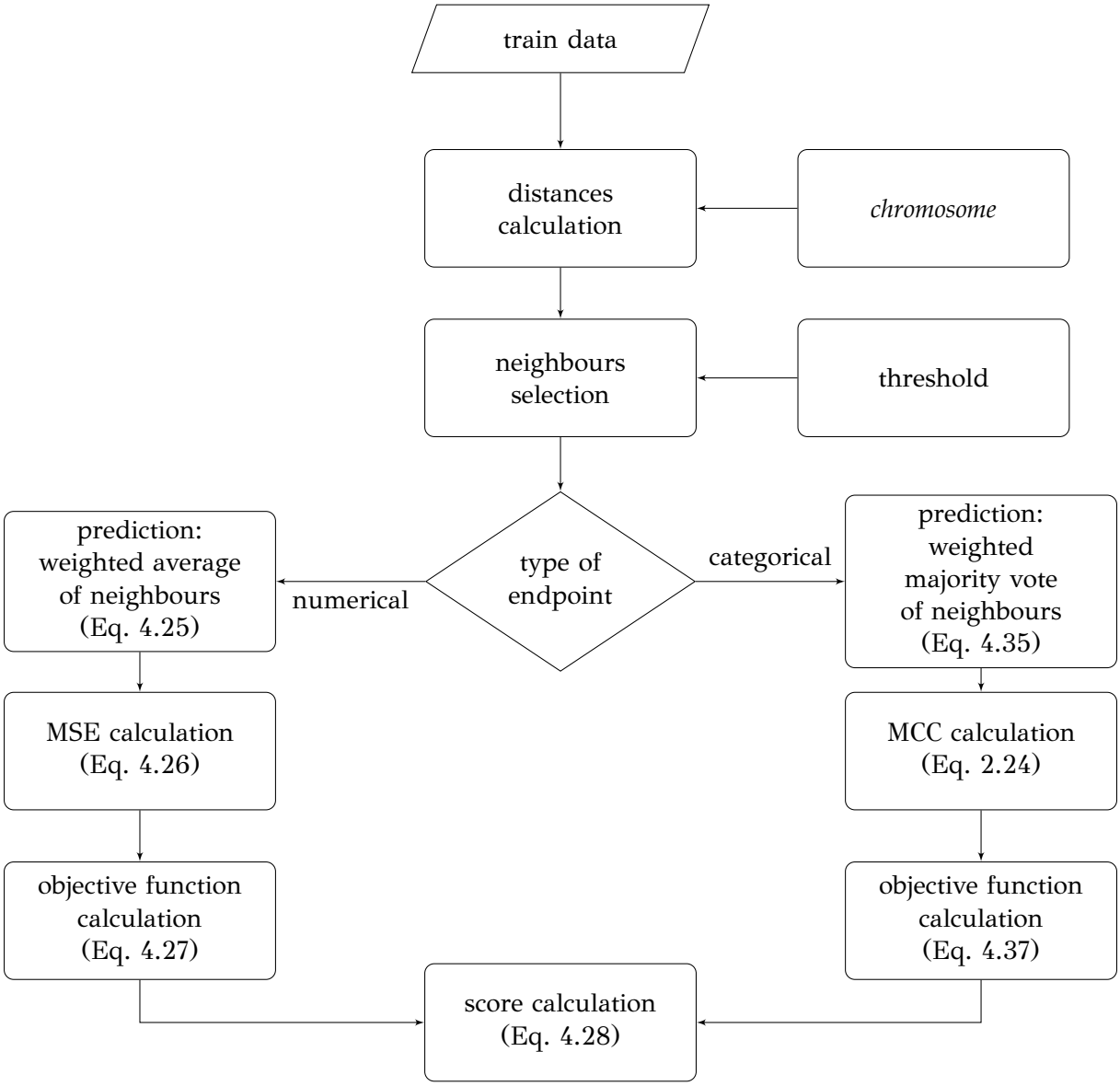


Figure 4.33: The main steps of a *chromosome*'s evaluation process during training in the *genetic algorithms* approach.

The proposed methodology is disseminated in order to be of use in real-life ENMs toxicity assessment applications. The developed methodology was used in order to built a user-friendly application that will facilitate its use from non-informatics experts. Apellis web ap-

plication is a useful tool for safety-by-design of novel structures and can also support the regulatory framework of nanosafety. In brief:

- Users can train and validate predictive read-across models using their own specific data even even if they do not have any prior computational skills
- The application applies a generic and novel computational workflow for estimating the endpoint of interest, which can be either categorical or numerical
- During training, the application selects the most important ENM properties of concern that affect their toxic behaviour
- In the process of grouping ENMs, the multi-perspective characterisation of ENMs can be taken into account, by defining more than one similarity criteria
- Visualisation tools -in the forms of easy-to-interpret plots and tables- are included in the application, which offer better and more clear understanding of grouping and similarities among ENMs
- The trained models can be saved in an electronic format, so that they can be easily retrieved, for calculating new predictions thereupon, developers can disseminate and share the produced models with the community
- Rich educational material and users support is provided along with the application
- Apellis is free to use and accessible at `https://apellis.jaqpot.org/`

4.6.1 Conclusions

The developed read-across method to predict toxicity related endpoints of ENMs is presented. This is an analogue read-across method where the analogues are defined as neighbours through an optimised threshold. This grouping boundary can be defined using one or more similarity criteria and is therefore expressed by one or more thresholds. The included variable selection and neighbours selection, is preformed and optimal solutions in terms of accuracy and predicted ENMs are found in an automated way.

The main outcomes of the method are: a reduced set of significant descriptors and a single or multiple threshold values which rigorously define the boundaries around a query ENM, where neighbouring ENMs are located.

Testing the proposed workflow with the *Gold ENMs* dataset used in the toxFlow [45] application, we confirmed the automation of the method concerning the selection of neighbours (the thresholds are not selected manually) and the optimisation in terms of accuracy (more predicted samples in the same level of produced R^2). The repetition of the executions with different starting chromosomes proved the stability of the method, given that the produced results per *predFactor* were similar.

The proposed methodology for ENM grouping and read-across has been extended, to include model development functionalities for the prediction of categorical toxicity-related endpoints. The methodology was demonstrated on the *MeOx ENMs [a]* dataset, which produced models that could predict in some cases the categorical endpoint of the dataset with accuracy up to 100%.

Chapter 5

Development of a grouping methodology based on the optimal piece-wise linear regression algorithm

To address the weakness of the formulation of a grouping hypothesis in a trial-and-error manner, in Chapter 4 we proposed an automated method that searches over the space of alternative hypotheses, and determines the one providing the most accurate read-across predictions. This novel grouping approach was based on the formulation and the solution of a MINLP problem. A specific GA scheme was developed to compute an approximate solution, because the problem could not be solved by standard optimisation methods. The GA that was designed for solving the problem is computationally intensive for large scale problems. To overcome this limitation, a different automated grouping/read-across methodology is presented in this Chapter, which is again based on the foundations of mathematical programming.

The proposed method overcomes the aforementioned limitations, by formulating a MILP problem that does not include non-linearities, and hence, can be solved efficiently to global optimality by standard very efficient optimisation techniques. Mixed-Integer Optimisation methodologies are commonly used to perform variable selection. [117] The proposed methodology is based on the work of Yang *et al*, [118] who presented the optimal piece-wise linear regression algorithm (OPLRA) for the development of reliable and interpretable QSAR models. OPLRA selects the feature that best separates the chemical domain into regions (partition feature) and produces predictive multi-variable linear equations in each region. Cardoso *et al*. [119] extended OPLRA with a regularisation parameter (OPLRAreg method) that has a twofold function; it controls over-fitting and simultaneously selects the most important features for the endpoint prediction. The OPLRA methodologies have been applied successfully to several cheminformatics datasets and have outperformed standard regression methodologies (support vector machines, random forest, *k*NN, LASSO etc.), in terms of predictive capability.

In this Chapter, an extension and adaptation of the OPLRAreg algorithm to the problem of grouping ENMs is presented, which includes optimal feature selection and exact definition of the grouping boundaries. The method can be considered as a category read-across approach, because it generates local multiple linear regression models, which establish a regular pattern in each group. [38], [44] When more than one type of descriptors are available (e.g. physicochemical, omics, quantum mechanical, image or biokinetics), it is possible to select multiple partition features and group ENMs in the multi-dimensional space (see §1.3.1). In the next paragraphs the formulation of the MILP problems with one (“1D MILP problem”) or two partition features (“2D MILP problem”) are presented. A workflow is developed for selecting the best partition feature in each dimension. The performance of the proposed method is

illustrated through the application to benchmark datasets and comparison with alternative predictive modelling approaches. The trained models using the above datasets were made publicly available through a user-friendly web service named vythos.

5.1 [m] Development of an automated grouping/read-across workflow

Similarly to the previous Chapter, during the development of a reliable grouping/read-across workflow for the prediction of ENMs toxicity, two separate objectives are pursued: the selection of the most important variables and the explicit definition of the grouping boundaries. These two different goals are achieved simultaneously in this work, by solving specifically formulated MILP problems, which allocate the available samples into regions, while the outcomes are predicted by local linear models in each region. The objective in the formulation of the MILP problems is to minimize the MAE between the experimental and the estimated values and to control over-fitting, through the elimination of the noisy features. The method is constructed around the development of the MILP formulations, which are presented next. This MILP problem -responsible for the selection of the partition feature in the first place, and the definition of the grouping boundaries- is the core-element in an iterative process that converges to the final number of regions.

5.1.1 [m] Development of the 1D MILP problem

The formulation of the 1D MILP problem is based on the publication of Cardoso-Silva *et al*, [119] where the full MILP model can be found. A minor modification is made on the error calculation equations: Simultaneous satisfaction of Eqs. 5.9, 5.10, and 5.11 means that when a sample s does not belong to region r, the error E_s^r becomes 0, otherwise it is larger or equal to the absolute value of the difference between the actual and the predicted value.

For the formulation of the problem, we assume that the partition feature (f^*) has been defined and also the number of groups (regions) where the available ENMs will be partitioned is fixed. The goal is to determine the grouping boundaries, select the variables of importance, place each ENM to a specific group and develop a multi-variable regression model in each region for computing the endpoint predictions. For convenience, we are using the same notation as in the original publication.

5.1.1.1 Data

A dataset comprising the values of F descriptors and the endpoint for S ENMs is needed. The dataset is represented by A_{sf}, a matrix containing the values of the F descriptors of the S ENMs and Y_s, the vector of the endpoint values of the S ENMs. The data are first scaled in the range [0,1], in order to be comparable and contribute impartially to the read-across predictions (§2.1.1).

5.1.1.2 Variables

The indices used in the formulation of the MILP problem are:

- s: samples, $s = 1, 2, \dots, S$
- f: features, $f = 1, 2, \dots, F$
- f^*: partition feature
- r: regions, $r = 1, 2, \dots, R$

The parameters used for the formulation of the MILP problem are:

- U, U': arbitrarily large positive numbers
- λ: the regularisation parameter
- ε: the smallest difference between complementary breakpoints

The continuous variables associated with the model are:

- $X_{f^*}^r$: breakpoint r on partition feature f^*, $r = 1, 2, \dots, R$
- $Pred_s^r$: predicted output for sample s in region r, $s = 1, 2, \dots, S$, $r = 1, 2, \dots, R$
- W_f^r: regression coefficient for feature f in region r, $f = 1, 2, \dots, F$, $r = 1, 2, \dots, R$
- B^r: intercept of regression function in region r, $r = 1, 2, \dots, R$
- E_s: training error in prediction for sample s, $s = 1, 2, \dots, S$
- E_s^r: training error in prediction for sample s in region r, $s = 1, 2, \dots, S$, $r = 1, 2, \dots, R$

The binary variable associated with the model is:

- F_s^r: equal to 1 if sample s belongs to region r, equal to 0 otherwise, $s = 1, 2, \dots, S$, $r = 1, 2, \dots, R$

5.1.1.3 Constraints

At every iteration, the number of regions r and the partition feature f^* are considered known. The breakpoints $X_{f^*}^r$ are consistent and arranged in an ordered way as shown in Eq. 5.1. Eq. 5.2 guarantees that the method will perform data partitioning and Eq. 5.3 sets the breakpoint of the last region equal to 1 given that the data are always normalised between $[0, 1]$.

$$X_{f^*}^r \geq X_{f^*}^{r-1} + \varepsilon \quad \forall r \in \{2, 3, \dots R\} \tag{5.1}$$

$$X_{f^*}^r \geq \varepsilon \quad r = 1 \tag{5.2}$$

$$X_{f^*}^r = 1 \quad r = R \tag{5.3}$$

The position of sample s in region r is encoded with binary variable F_s^r. If a sample s belongs to region r, then $F_s^r = 1$, otherwise $F_s^r = 0$. Eq. 5.4 guarantees that each sample s belongs to only one region.

$$\sum_{r=1}^{R} F_s^r = 1 \quad \forall s \in \{1, 2, \dots S\} \tag{5.4}$$

Eqs. 5.5 and 5.6 assign samples to the correct region according to the breakpoints $X_{f^*}^r$ and the A_{sf^*} numeric value of sample s on feature f^*. U is an arbitrarily large positive number (see also page 37).

$$A_{sf^*} \geq X_{f^*}^{r-1} + \varepsilon - U \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, \dots S\}, r \in \{2, 3, \dots, R\} \tag{5.5}$$

$$A_{sf^*} \leq X_{f^*}^r - \varepsilon + U \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, \dots S\}, r \in \{1, 2, \dots, R-1\} \tag{5.6}$$

Eq. 5.7 represents the multi-variable linear model which calculates the predicted value $Pred_s^r$ for sample s in region r, where W_f^r is the regression coefficient of feature f in region r and B^r is the intercept in region r.

$$Pred_s^r = \sum_{f=1}^{F} W_f^r \cdot A_{sf} + B^r \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{5.7}$$

Constraints 5.8 to 5.12 are used to compute the prediction error E_s for sample s, as the absolute difference between the actual endpoint value and the value predicted by the linear model corresponding to the region where samples s belongs, or equal to zero. Eq. 5.8 forces E_s to be a non-negative number.

The positive error in a specific region E_s^r takes values smaller than a large positive number if sample s belongs to region r, and is equal to zero if sample s does not belong to that region (Eq. 5.9). Eqs. 5.10 and 5.11 apply constraints on the absolute error in prediction E_s^r of sample s in region r.

$$E_s \geq 0 \quad \forall s \in \{1, 2, ... S\} \tag{5.8}$$

$$E_s^r \leq U' \cdot F_s^r \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{5.9}$$

$$E_s^r \geq Y_s - Pred_s^r - U' \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{5.10}$$

$$E_s^r \geq Pred_s^r - Y_s - U' \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{5.11}$$

$$E_s \geq E_s^r \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{5.12}$$

5.1.1.4 Objective function

The OF includes two terms; the mean absolute error (MAE) which penalizes the summation of prediction errors over all samples s, divided by the number of samples and a regularisation term (REG) that reduces the risk of over-fitting by penalizing the summation of the absolute values of all regression coefficients. MAE and REG are defined by the equations shown below:

$$MAE = \frac{\sum_{s=1}^{S} E_s}{|S|} \tag{5.13}$$

$$REG = \sum_{r=1}^{R} \sum_{f=1}^{F} W_f^{r+} \tag{5.14}$$

$$W_f^{r+} \geq W_f^r \quad \forall f \in \{1, 2, ... F\}, r \in \{1, 2, ..., R\} \tag{5.15}$$

$$W_f^{r+} \geq -W_f^r \quad \forall f \in \{1, 2, ... F\}, r \in \{1, 2, ..., R\} \tag{5.16}$$

where E_s is the absolute error for each sample s computed in Eqs. 5.8 to 5.12, $|S|$ is the number of samples in the training set and W_f^{r+} indicates the absolute value of regression coefficients W_f^r.

The OF to be minimized is shown in Eq. 5.17.

$$z = MAE + \lambda \cdot REG \tag{5.17}$$

where λ is a positive user-defined parameter that controls the influence of regularisation.

5.1.2 [m] Development of the 2D MILP problem

Due to the complex structure of ENMs, different types of descriptors are often used for their characterisation, such as physicochemical, biological or image descriptors. This multi-perspective characterisation of ENMs is addressed by this extension, where the objective is to define groups of ENMs in two or more dimensions.

Due to the complex structure of ENMs, different types of descriptors are often used for their characterisation, such as physicochemical, biological or image descriptors. This multi-perspective characterisation of ENMs is addressed by this extension, where the objective is to define groups of ENMs in two or more dimensions. The goal is again to determine the grouping boundaries in the 2D or higher-dimension space, select the variables of importance from each type of descriptors, place each ENM into a specific group and develop a multi-variable regression model in each region for computing the endpoint predictions. The extension is presented for two dimensions for clarity and brevity.

5.1.2.1 Data

A dataset comprising the values of F descriptors and the endpoint for S ENMs is needed. The descriptors are partitioned into sets F_{M} and F_{N}, which include the descriptors of types M and N respectively. Accordingly, the full descriptor dataset represented by A_{sf}, is now divided into two datasets, $A_{\mathrm{M}_{sf}}$, $A_{\mathrm{N}_{sf}}$. Y_s is the vector of the endpoint values of the S ENMs. The data are first scaled in the range [0,1]. The number of regions R_{M} and R_{N} and the partition features f^*_{M} and f^*_{N} are fixed for types M and N respectively.

5.1.2.2 Variables

The indices used in the formulation of the 2D MILP problem are:

- s: samples, $s = 1, 2, \dots, S$
- f: features, $f = 1, 2, \dots, F$
- f_{M}: features in category M, $f_{\mathrm{M}} = 1, 2, \dots, F_{\mathrm{M}}$
- f_{N}: features in category N, $f_{\mathrm{N}} = 1, 2, \dots, F_{\mathrm{N}}$
- f^*_{M}: partition feature in dimension M
- f^*_{N}: partition feature in dimension N
- r_{M}: regions in dimension M, $r_{\mathrm{M}} = 1, 2, \dots, R_{\mathrm{M}}$
- r_{N}: regions in dimension N, $r_{\mathrm{N}} = 1, 2, \dots, R_{\mathrm{N}}$

The parameters used in the data representation and MILP solution are:

- U, U': arbitrarily large positive numbers
- λ: the regularisation parameter
- ε: the smallest difference between complementary breakpoints

The continuous variables associated with the model are:

- $X_{\mathrm{M}}{}^{r_{\mathrm{M}}}_{f^*_{\mathrm{M}}}$: breakpoint r_{M} on partition feature f^*_{M}, $r_{\mathrm{M}} = 1, 2, \dots, R_{\mathrm{M}}$
- $X_{\mathrm{N}}{}^{r_{\mathrm{N}}}_{f^*_{\mathrm{N}}}$: breakpoint r_{N} on partition feature f^*_{N}, $r_{\mathrm{N}} = 1, 2, \dots, R_{\mathrm{N}}$

- $Pred_s^{r_{\mathrm{M}} r_{\mathrm{N}}}$: predicted output for sample s in region $r_{\mathrm{M}} r_{\mathrm{N}}$, $s = 1, 2, \dots, S$, $r_{\mathrm{M}} = 1, 2, \dots, R_{\mathrm{M}}$, $r_{\mathrm{N}} = 1, 2, \dots, R_{\mathrm{N}}$
- $W_f^{r_{\mathrm{M}} r_{\mathrm{N}}}$: regression coefficient for feature f in region $r_{\mathrm{M}} r_{\mathrm{N}}$, $f = 1, 2, \dots, F$, $r_{\mathrm{M}} = 1, 2, \dots, R_{\mathrm{M}}$, $r_{\mathrm{N}} = 1, 2, \dots, R_{\mathrm{N}}$
- $B^{r_{\mathrm{M}} r_{\mathrm{N}}}$: intercept of regression function in region $r_{\mathrm{M}} r_N$, $r_{\mathrm{M}} = 1, 2, \dots, R_{\mathrm{M}}$, $r_{\mathrm{N}} = 1, 2, \dots, R_{\mathrm{N}}$
- E_s: training error in prediction for sample s, $s = 1, 2, \dots, S$
- $E_s^{r_{\mathrm{M}}}$: training error in prediction for sample s in region r_{M}, $s = 1, 2, \dots, S$, $r_{\mathrm{M}} = 1, 2, \dots, R_{\mathrm{M}}$
- $E_s^{r_{\mathrm{N}}}$: training error in prediction for sample s in region r_{N}, $s = 1, 2, \dots, S$, $r_{\mathrm{N}} = 1, 2, \dots, R_{\mathrm{N}}$

The binary variables associated with the model are:

- ${F_{\mathrm{M}}}_s^{r_{\mathrm{M}}}$: equal to 1 if sample s belongs to region r_{M} in dimension M, equal to 0 otherwise
- ${F_{\mathrm{N}}}_s^{r_{\mathrm{N}}}$: equal to 1 if sample s belongs to region r_{N} in dimension N, equal to 0 otherwise

5.1.2.3 Constraints

Eqs. 5.18α′, 5.18β′, 5.19α′, 5.19β′ secure that dimensions M and N are partitioned into R_{M} and R_{N} non-zero intervals respectively and the breakpoints ${X_{\mathrm{M}}}_{f_{\mathrm{M}}^*}^{r_{\mathrm{M}}}$ and ${X_{\mathrm{N}}}_{f_{\mathrm{N}}^*}^{r_{\mathrm{N}}}$ are arranged in an ordered way.

$$ {X_{\mathrm{M}}}_{f_{\mathrm{M}}^*}^{r_{\mathrm{M}}} \geq \varepsilon \quad r_{\mathrm{M}} = 1 \tag{5.18α′} $$

$$ {X_{\mathrm{N}}}_{f_{\mathrm{N}}^*}^{r_{\mathrm{N}}} \geq \varepsilon \quad r_{\mathrm{N}} = 1 \tag{5.18β′} $$

$$ {X_{\mathrm{M}}}_{f_{\mathrm{M}}^*}^{r_{\mathrm{M}}} \geq {X_{\mathrm{M}}}_{f_{\mathrm{M}}^*}^{r_{\mathrm{M}}-1} + \varepsilon \quad \forall r_{\mathrm{M}} \in \{2, 3, \dots R_{\mathrm{M}}\} \tag{5.19α′} $$

$$ {X_{\mathrm{N}}}_{f_{\mathrm{N}}^*}^{r_{\mathrm{N}}} \geq {X_{\mathrm{N}}}_{f_{\mathrm{N}}^*}^{r_{\mathrm{N}}-1} + \varepsilon \quad \forall r_{\mathrm{N}} \in \{2, 3, \dots R_{\mathrm{N}}\} \tag{5.19β′} $$

Eqs. 5.20α′ and 5.20β′ require that the right limit point of the last interval in each dimension is equal to 1, given that data are scaled in [0,1].

$$ {X_{\mathrm{M}}}_{f_{\mathrm{M}}^*}^{r_{\mathrm{M}}} = 1 \quad r_{\mathrm{M}} = R_{\mathrm{M}} \tag{5.20α′} $$

$$ {X_{\mathrm{N}}}_{f_{\mathrm{N}}^*}^{r_{\mathrm{N}}} = 1 \quad r_{\mathrm{N}} = R_{\mathrm{N}} \tag{5.20β′} $$

The combination of $F_{\mathrm{M}_s}^{r_{\mathrm{M}}}$ and $F_{\mathrm{N}_s}^{r_{\mathrm{N}}}$ binary variables defines the position of sample s in the two-dimensional space. Eqs. 5.21α′ and 5.21β′ guarantee that a sample belongs to one interval only in each dimension. Eqs. 5.22α′ and 5.22β′ assign the samples to the correct intervals r_{M} in dimension M, according to the breakpoints and the $A_{\mathrm{M}_{s f_{\mathrm{M}}^*}}$ numeric value of sample s on feature f_{M}^*. Similarly, Eqs. 5.23α′ and 5.23β′ assign the samples to the correct intervals r_{N} in dimension N.

$$ \sum_{r_{\mathrm{M}}=1}^{R_{\mathrm{M}}} {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}} = 1 \quad \forall s \in \{1, 2, \dots S\} \tag{5.21α′} $$

$$\sum_{r_{\mathrm{N}}=1}^{R_{\mathrm{N}}} {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}} = 1 \quad \forall s \in \{1, 2, \dots S\} \tag{5.21β′}$$

$$A_{\mathrm{M}sf_{\mathrm{M}}^*} \leq {X_{\mathrm{M}}}_{f_{\mathrm{M}}^*}^{r_{\mathrm{M}}} - \varepsilon + U \cdot (1 - {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}}) \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}} - 1\} \tag{5.22α′}$$

$$A_{\mathrm{M}sf_{\mathrm{M}}^*} \geq {X_{\mathrm{M}}}_{f_{\mathrm{M}}^*}^{r_{\mathrm{M}}-1} + \varepsilon - U \cdot (1 - {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}}) \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{2, 3, \dots, R_{\mathrm{M}}\} \tag{5.22β′}$$

$$A_{\mathrm{N}sf_{\mathrm{N}}^*} \leq {X_{\mathrm{N}}}_{f_{\mathrm{N}}^*}^{r_{\mathrm{N}}} - \varepsilon + U \cdot (1 - {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}}) \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}} - 1\} \tag{5.23α′}$$

$$A_{\mathrm{N}sf_{\mathrm{N}}^*} \geq {X_{\mathrm{N}}}_{f_{\mathrm{N}}^*}^{r_{\mathrm{N}}-1} + \varepsilon - U \cdot (1 - {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}}) \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{N}} \in \{2, 3, \dots, R_{\mathrm{N}}\} \tag{5.23β′}$$

Eq. 5.24 is the multi-variable regression model which predicts the endpoint value $Pred_s^{r_{\mathrm{M}}r_{\mathrm{N}}}$ of sample s in region $r_{\mathrm{M}}r_{\mathrm{N}}$ in the two-dimensional space: $W_f^{r_{\mathrm{M}}r_{\mathrm{N}}}$ are the coefficients and $B^{r_{\mathrm{M}}r_{\mathrm{N}}}$ is the intercept.

$$\begin{gathered} Pred_s^{r_{\mathrm{M}}r_{\mathrm{N}}} = \sum_{f=1}^{F} W_f^{r_{\mathrm{M}}r_{\mathrm{N}}} \cdot A_{sf} + B^{r_{\mathrm{M}}r_{\mathrm{N}}} \\ \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \end{gathered} \tag{5.24}$$

Eq. 5.25 forces E_s to be a non-negative number. Eqs. 5.26α′ to 5.29β′ guarantee that the error E_s is equal to the absolute value of the difference between the predicted and the actual value, in the region where sample s is assigned. Simultaneous satisfaction of Eqs. 5.26α′, 5.27α′, and 5.28α′ means that when a sample s does not belong to region r_{M}, the error $E_s^{r_{\mathrm{M}}}$ becomes 0, otherwise it is larger or equal to the absolute value of the difference between the actual and the predicted value. The same rules apply for error in dimension N.

$$E_s \geq 0 \quad \forall s \in \{1, 2, \dots S\} \tag{5.25}$$

$$E_s^{r_{\mathrm{M}}} \leq U' \cdot {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}} \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\} \tag{5.26α′}$$

$$E_s^{r_{\mathrm{N}}} \leq U' \cdot {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}} \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \tag{5.26β′}$$

$$\begin{gathered} E_s^{r_{\mathrm{M}}} \geq Y_s - Pred_s^{r_{\mathrm{M}}r_{\mathrm{N}}} - U' \cdot (1 - {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}}) - U' \cdot (1 - {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}}) \\ \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \end{gathered} \tag{5.27α′}$$

$$\begin{gathered} E_s^{r_{\mathrm{N}}} \geq Y_s - Pred_s^{r_{\mathrm{M}}r_{\mathrm{N}}} - U' \cdot (1 - {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}}) - U' \cdot (1 - {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}}) \\ \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \end{gathered} \tag{5.27β′}$$

$$\begin{gathered} E_s^{r_{\mathrm{M}}} \geq Pred_s^{r_{\mathrm{M}}r_{\mathrm{N}}} - Y_s - U' \cdot (1 - {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}}) - U' \cdot (1 - {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}}) \\ \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \end{gathered} \tag{5.28α′}$$

$$\begin{gathered} E_s^{r_{\mathrm{N}}} \geq Pred_s^{r_{\mathrm{M}}r_{\mathrm{N}}} - Y_s - U' \cdot (1 - {F_{\mathrm{M}}}_s^{r_{\mathrm{M}}}) - U' \cdot (1 - {F_{\mathrm{N}}}_s^{r_{\mathrm{N}}}) \\ \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \end{gathered} \tag{5.28β′}$$

$$E_s \geq E_s^{r_{\mathrm{M}}} \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\} \tag{5.29α′}$$

$$E_s \geq E_s^{r_{\mathrm{N}}} \quad \forall s \in \{1, 2, \dots S\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \tag{5.29β′}$$

5.1.2.4 Objective function

The OF includes two terms; the MAE which penalizes the summation of prediction errors over all samples S, divided by the number of samples and a regularisation term (REG) that reduces the risk of over-fitting by penalizing the summation of the absolute values of all regression coefficients. The MAE and REG terms are defined in Eq. 5.30 and Eq. 5.31:

$$MAE = \frac{\sum_{s=1}^{S} E_s}{|S|} \tag{5.30}$$

$$REG = \sum_{r_{\mathrm{M}}=1}^{R_{\mathrm{M}}} \sum_{r_{\mathrm{N}}=1}^{R_{\mathrm{N}}} \sum_{f=1}^{F} W_f^{r_{\mathrm{M}} r_{\mathrm{N}}{}^+} \tag{5.31}$$

where $|S|$ is the number of samples in the training set and $W_f^{r_{\mathrm{M}} r_{\mathrm{N}}{}^+}$ indicates the absolute value of regression coefficients of $W_f^{r_{\mathrm{M}} r_{\mathrm{N}}}$, defined in Eqs. 5.32 and 5.33:

$$W_f^{r_{\mathrm{M}} r_{\mathrm{N}}{}^+} \geq W_f^{r_{\mathrm{M}} r_{\mathrm{N}}} \quad \forall f \in \{1, 2, \dots F\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \tag{5.32}$$

$$W_f^{r_{\mathrm{M}} r_{\mathrm{N}}{}^+} \geq -W_f^{r_{\mathrm{M}} r_{\mathrm{N}}} \quad \forall f \in \{1, 2, \dots F\}, r_{\mathrm{M}} \in \{1, 2, \dots, R_{\mathrm{M}}\}, r_{\mathrm{N}} \in \{1, 2, \dots, R_{\mathrm{N}}\} \tag{5.33}$$

The OF to be minimized is the same as in the 1D MILP formulation (Eq. 5.17).

5.2 [m] Proposed workflow

The full read-across workflow followed in this paper for the 1D problem, is presented in Figure 5.1. A similar workflow is used in the case of 2D problems. The algorithm starts by solving a standard multi-variable linear regression (MLR) problem on the training data using Eq. 5.17 as the OF, without defining multiple regions in any dimension. The model is applied on the test dataset and the values of the OF z for the training and test sets, $z_{\mathrm{R=1,train}}$ and $z_{\mathrm{R=1,test}}$, are recorded.

Subsequently, the MILP problem considering two regions is solved multiple times. For the 1D MILP problem, a different independent feature is used each time as the partition feature. The feature that produces the best model in terms of minimizing the OF is selected as the partition feature f^*. The best two-region model is applied to the test set and the value of the OF $z_{\mathrm{R=2,test}}$ is compared to $z_{\mathrm{R=1,test}}$. Eq. 5.34 is used for evaluating the level of improvement. If Eq. 5.34 is not satisfied, the algorithm stops, and the two-region MILP is the final read-across model. Otherwise, the algorithm proceeds with defining more regions. This process is iterated until the improvement in the value of z_{test} in consecutive iterations does not satisfy Eq. 5.34.

$$z_{\mathrm{R,test}} \leq (1 - \beta) \cdot z_{\mathrm{R-1,test}} \tag{5.34}$$

where β is the difference of absolute errors between consecutive iterations (user-defined) and, $z_{\mathrm{R-1,test}}$ and $z_{\mathrm{R,test}}$ are the OF values for the test set between two consecutive iterations.

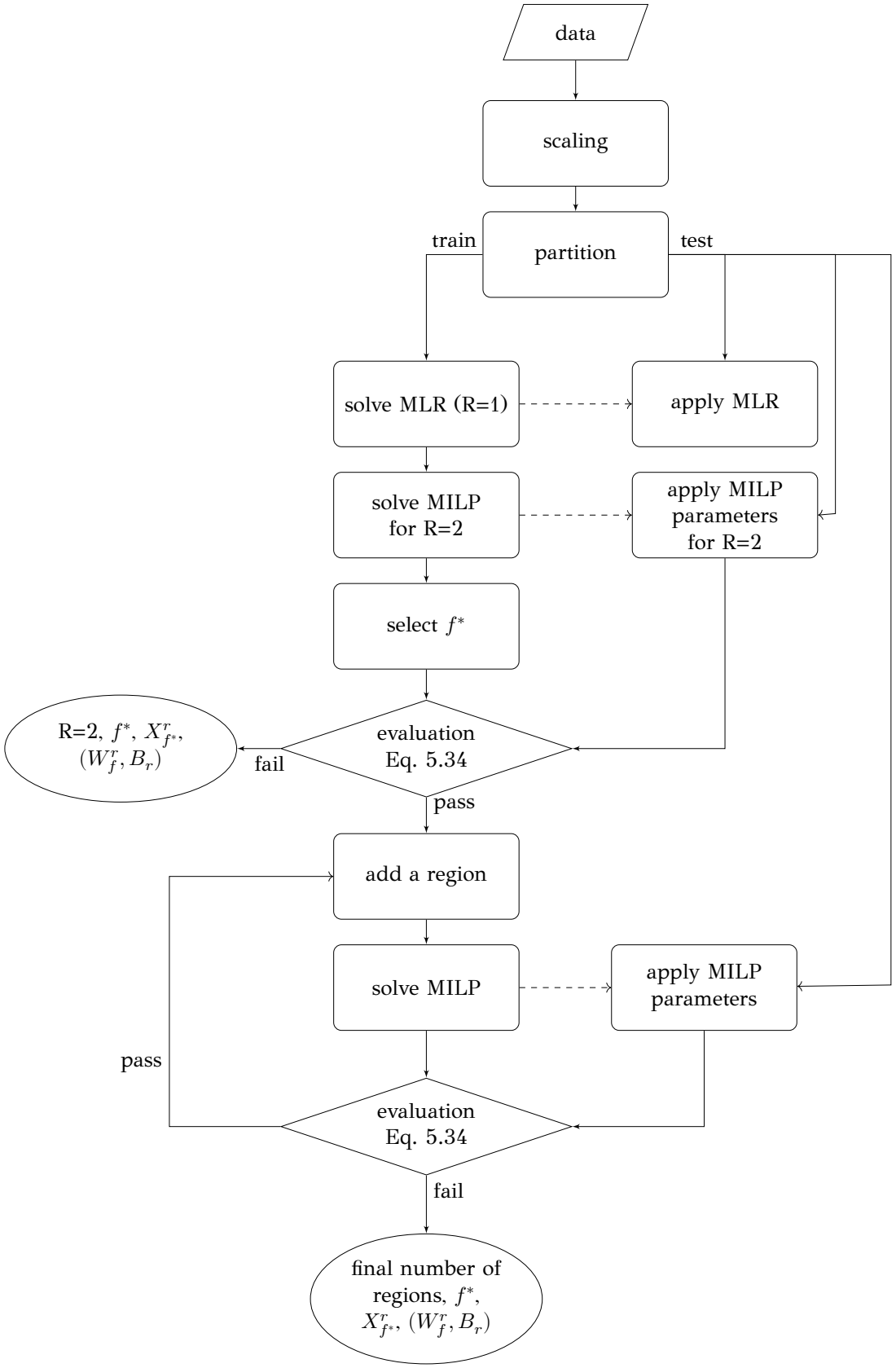


Figure 5.1: Schematic description of the grouping/read-across proposed methodology based on the input properties. The workflow describes the 1D MILP problem however, it is similar when dealing with the 2D MILP problem.

In the 2D MILP problem, the selection of the partition features f_{M}^{*} and f_{N}^{*} is performed simultaneously, sequentially or independently for each feature type considering two regions in each dimension (in total four regions):

- The simultaneous approach examines all combinations of features from types M and N by formulating and solving the corresponding 2D MILP problems. The combination of partition features f_{M}^{*} and f_{N}^{*} that produces the best value of the OF for the test set is considered as the best model.
- In the sequential approach, the 1D MILP problem is solved first for each feature of type M. The feature that produces the minimum value of the OF is selected as partition feature f_{M}^{*}. In 2D, f_{M}^{*} is fixed and each feature of category N is examined as partition feature f_{N}^{*}.
- If the selection is accomplished independently, the algorithm solves the 1D MILP problem for every feature of category F_{M} and category F_{N} separately to determine the partition features in each dimension.

After the selection of the partition features, the algorithm examines the addition of regions as described above for the 1D MILP problem in a similar iterative fashion. Since the regions are now defined in two dimensions, there are three scenarios for the addition of regions: increase one region in dimension M, one region in dimension N or one region in each dimension. All scenarios are tested and the scenario that minimizes the value of the OF is selected. Eq. 5.34 is used again to decide if execution of the workflow is stopped or the workflow continues with considering additional regions.

5.3 [m] Validation

Many alternative procedures were applied for the validation of the proposed methodology, starting with an external validation method that divides the full dataset into training and test sets. The Kennard and Stone method was used for data partitioning, as it allows the selection of representative samples in both the training and test sets that cover uniformly the chemical space of the full dataset. [66] Samples from the training set are used for developing the model, i.e. defining the grouping boundaries, selecting the variables of importance and building the regression model in each group.

The quality of fit between the predicted and experimental values of the training set is expressed by the correlation coefficient R^2 (Eq. 2.6). The test dataset is used for validating the final model and controlling the number of regions defined in each dimension. We used the external explained variance metric, Q^2_{ext}, which compares the read-across predictions for the test samples with the actual endpoint values (Eq. 2.10). [54]

The RMSE index was computed on both the training and test sets to further assess the validity and accuracy of the produced models. Together, the MAE and RMSE indexes provide a complete and thorough validation of the prediction accuracy, independently of the distribution level of the training-test endpoint values. [71] Eq. 2.9 presents the formula that was used to compute the RMSE index on the training set.

It has been reported in the Literature that for datasets with small numbers of samples, external validations metrics may have high variation among different splits of the data. [120] Taking into account that many datasets in the field of nanoinformatics are small-sample datasets, we addressed this possible drawback by considering three additional validation procedures, namely multiple random splittings into training and external validation sets, the LOO cross validation method and the Y-randomisation test. [73] In the LOO method, the explained variance in prediction (Q^2_{LOO}, Eq. 2.11) is calculated. [54]

5.4 [m] Domain of applicability

In this methodology, where multiple linear equations constitute the read-across model, we consider a prediction as reliable only if all the independent descriptors are within the ranges defined by the training samples. Therefore, before using the model for predicting the endpoint of an external ENM, the input variables should be scaled first according to the original dataset's min-max values. Only if all scaled values are within the range [0,1], the read-across prediction can be accepted.

5.5 [r] Results and discussion

The above workflow was implemented in the MATLAB® programming language (§A′.1.1), using the `YALMIP` (`yalmip.github.io/`) toolbox, for formulating and solving the MILP problems. `YALMIP` offers a selection of solvers for robust optimisation. [121], [122] The results presented in the following paragraphs are produced using the `mosek` solver (`www.mosek.com/`) and a free academic license. The source code is available at GitHub (`github.com/NikiKou/MILP-read-across`).

The proposed workflow was demonstrated on two datasets: the *MWCNTs [a]* dataset (§3.4) and the *Gold ENMs* dataset (§3.1). The datasets were scaled between 0 and 1 previous to any modelling activities. In addition they were divided into train and test sets using a ratio of 75:25 for the *MWCNTs [a]* dataset and 66:33 for the *Gold ENMs* dataset.

5.5.1 [r] Results of the 1D models

The 1D workflow was applied first, considering that all independent variables belong to a single group. The MILP problems were solved with different values of the regularisation parameter λ and with $\beta = 0.05$. Tables 5.1 and 5.2 summarize the results of the optimisation problem for the two case studies, which include: the optimal values of the MAE and REG terms, the correlation coefficient R^2, the external explained variance Q^2_{ext}, the total number of selected variables, the number of regions, the number of selected variables per region r, and the ENMs assigned to each region r for both training and test sets.

By setting λ equal to 0, MAE is zero and R^2 is equal or approaches to 1, which means that an almost perfect fit is achieved between experimental and actual values in the training set. However, this is due to over-fitting, and the models fail completely when they are applied to the test sets. When the regularisation term is introduced, the external explained variance Q^2_{ext} is improved, and smaller or zero regression coefficients are obtained. Zero regression coefficients indicate the noisy variables which need to be removed.

***MWCNTs* [a] dataset** For the *MWCNTs [a]* dataset, the best model -in terms of the MAE and RMSE validation metrics on the test set- was produced by setting the regularisation parameter λ equal to 0.01. The model selected four independent features, which are presented in Table 5.3. The "hydrogen-bond acidity" (α) was chosen as the partition feature, which partitioned the data into two regions, which are shown in Figure 5.2. The full predictive model is presented in Eq. 5.35.

Figure 5.3 presents graphically the predicted vs. actual values for the training and test samples.

The training probe compounds that belong to regions A and B are presented in Table 5.4. For each unknown ENM belonging to region A or B the training ENMs of that region are its neighbours.

The results of the application of the LOO cross-validation procedure and the Y-randomisation method are summarized in Tables 5.5 and 5.6 respectively. The high Q^2_{LOO}

value and the large predictive errors obtained in five random shuffles of the Y-vector, indicate that the model development phase is reliable and that the predictive power of the model is not due to chance correlations. Five additional predictive models were developed, by splitting the full dataset into training and test sets, using random partitioning instead of the Kennard-Stone method. The results are presented in Table 5.7 and illustrate that random partition can still produce predictive models with high accuracy on both the training and test sets.

Table 5.1: Results of the 1D MILP workflow applied on the *MWCNTs [a]* dataset for different values of the regularisation parameter, λ.

		λ				
		0.000	**0.005**	**0.010**	**0.020**	**0.030**
	Regions	2	2	2	2	2
	REG	269.47	13.87	7.43	7.09	4.86
	Selected variables	5	5	4	4	5
	Selected variables per region$_r$	5|5	5|3	4|0	4|0	5|0
Training	**MAE**	0.09	0.09	0.13	0.14	0.19
set	**RMSE**	0.18	0.16	0.21	0.21	0.26
21	R^2	0.97	0.97	0.96	0.96	0.93
samples	**ENMs per region**$_r$	16|5	17|4	20|1	20|1	19|2
Test	**MAE**	3.94	0.19	0.22	0.22	0.27
set	**RMSE**	7.31	0.27	0.27	0.26	0.46
7	$Q^2_{\mathbf{ext}}$	-105.77	0.85	0.86	0.86	0.59
samples	**ENMs per region**$_r$	5|2	7|0	6|1	6|1	6|1

Table 5.2: Results of the 1D MILP workflow applied on the *Gold ENMs* dataset for different values of the regularisation parameter, λ.

		λ					
		0.000	**0.005**	**0.010**	**0.020**	**0.030**	**0.050**
	Regions	2	3	2	2	2	2
	REG	45242.00	56.86	37.89	15.53	12.33	8.17
	Selected variables	70	44	39	23	16	9
	Selected variables per region$_r$	54|41	33|18|0	33|8	20|4	15|1	9|0
Training	**MAE**	0.00	0.00	0.14	0.50	0.58	0.73
set	**RMSE**	0.00	0.01	0.40	1.10	1.14	1.31
55	R^2	1	1	0.97	0.82	0.81	0.73
samples	**ENMs per region**$_r$	50|5	35|19|1	45|10	45|10	45|10	45|10
Test	**MAE**	163.30	0.57	0.50	0.56	0.56	0.56
set	**RMSE**	245.81	0.85	0.60	0.70	0.70	0.71
29	$Q^2_{\mathbf{ext}}$	-20664.00	0.75	0.88	0.83	0.83	0.83
samples	**ENMs per region**$_r$	25|4	16|13|0	28|1	28|1	28|1	28|1

$$\log k = \begin{cases} 1.841 \cdot \pi - 0.016 \cdot \alpha - 2.463 \cdot \beta + 3.106 \cdot V + 2.431 & \text{if } \alpha \leq 0.907 \text{ region A} \\ 4.260 & \text{if } \alpha > 0.907 \text{ region B} \end{cases} \tag{5.35}$$

Table 5.3: Variables involved in the 1D *MWCNTs [a]* model.

Role	Symbol
Endpoint	$\log k$
Partition feature	α
MLR variables	π
	α
	β
	V

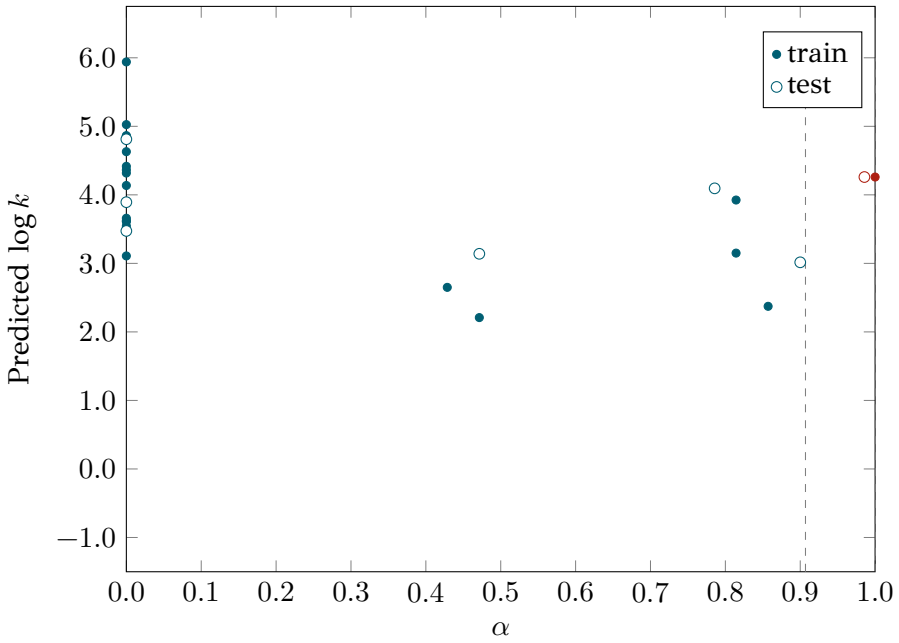


Figure 5.2: Breakpoints, and region distribution of the *MWCNTs [a]* samples, resulted by application of the grouping/read-across 1D MILP workflow. Blue markers indicate the samples belonging to region A and red markers indicate the samples belonging to region B. Training samples are represented by circular markers whereas, test samples are represented by triangular markers.

Table 5.4: Groups of *MWCNTs [a]* training samples as produced by the application of the 1D MILP workflow.

	Training probe compounds
Region A	4-chloroanisole, phenethyl alcohol, 1-methylnaphthalene, benzyl alcohol, phenol, benzonitrile, 3-methylphenol, chlorobenzene, p-xylene, bromobenzene, acetophenone, 3,5-dimethylphenol, methyl benzoate, iodobenzene, propylbenzene, 4-chlorotoluene, ethyl benzoate, 4-nitrotoluene, 4-chloroacetophenone, naphthalene
Region B	3-bromophenol

Table 5.5: LOO cross-validation results of the 1D MILP workflow applied on the *MWCNTs [a]* dataset for $\lambda = 0.01$.

λ	MAE	RMSE	$Q^2_{\mathbf{LOO}}$
0.010	0.27	0.33	0.87

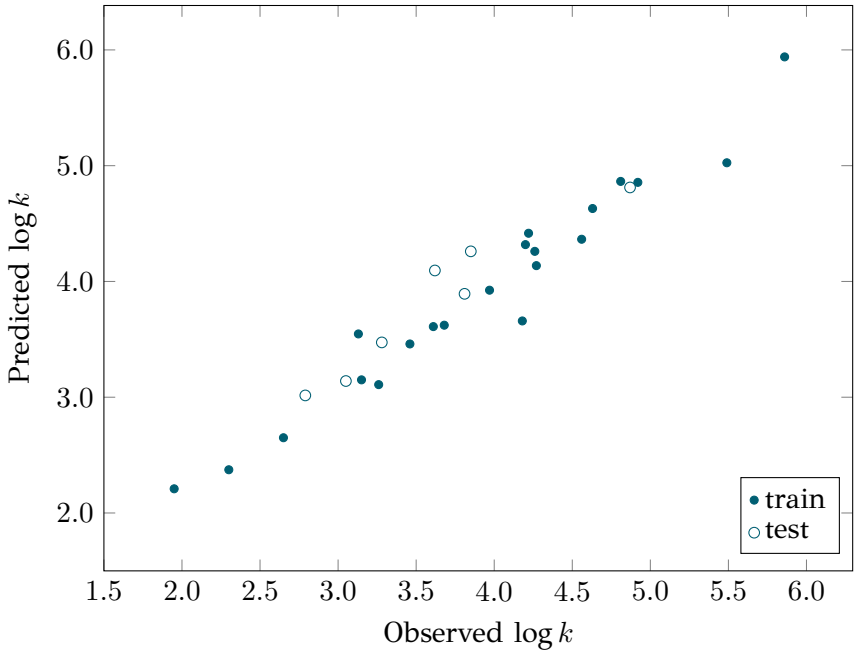


Figure 5.3: Plot of experimental versus predicted log k values produced by the application of the grouping/read-across 1D MILP problem to the *MWCNTs [a]* dataset. Training samples are represented by circular markers whereas, test are represented by triangular markers.

Table 5.6: Y-randomisation results of the 1D MILP workflow applied on the *MWCNTs [a]* dataset for $\lambda = 0.01$.

		# y-randomised endpoint				
		1	**2**	**3**	**4**	**5**
	Regions	2	2	2	2	2
	REG	11.26	19.15	14.93	13.50	6.31
	Selected variables	4	5	5	5	4
	Selected variables per region$_r$	4|3	5|3	2|4	5|4	4|2
Training	**MAE**	0.35	0.36	0.46	0.43	0.44
set	**RMSE**	0.64	0.72	0.82	0.72	0.70
21	R^2	0.57	0.47	0.29	0.45	0.53
samples	**ENMs per region**$_r$	15|6	15|6	5|16	15|6	17|4
Test	**MAE**	0.90	0.86	0.63	0.49	1.18
set	**RMSE**	1.18	0.93	0.83	0.60	1.45
7	$Q^2_{\mathbf{ext}}$	-1.78	-0.72	-0.37	0.28	-3.20
samples	**ENMs per region**$_r$	5|2	5|2	1|6	5|2	5|2

Table 5.7: Results of the 1D MILP workflow applied on the *MWCNTs [a]* dataset for $\lambda = 0.01$ and for different random train-test set partitions.

		# random partition				
λ = **0.01**		**1**	**2**	**3**	**4**	**5**
	Regions	2	2	2	2	2
	REG	5.68	7.03	8.32	7.76	7.10
	Selected variables	5	5	4	5	5
	Selected variables per region$_r$	5|0	0|5	4|1	5|1	5|1
Training	**MAE**	0.12	0.14	0.09	0.10	0.11
set	**RMSE**	0.17	0.19	0.12	0.17	0.17
21	R^2	0.97	0.95	0.98	0.96	0.97
samples	**ENMs per region**$_r$	19|2	1|20	19|2	19|2	18|3
Test	**MAE**	0.32	0.26	0.26	0.27	0.31
set	**RMSE**	0.38	0.32	0.31	0.32	0.32
7	$Q^2_{\mathbf{ext}}$	0.82	0.91	0.74	0.92	0.85
samples	**ENMs per region**$_r$	6|1	1|6	7|0	7|0	7|0

***Gold ENMs* dataset** For the *Gold ENMs* dataset, the regularisation parameter value that produced the best model -in terms of the most reliable predictions in external validation- was $\lambda = 0.01$. The model is given in Eq. 5.36. and the external explained variance Q^2_{ext} was 0.88. The domain was partitioned into two regions and the partition feature was the Apolipoprotein B-100 (P04114), as shown in Figure 5.4. This particular protein operates as a recognition signal for the cellular binding and integration of low-density lipoprotein (LDL) particles by the apoB/E receptor (`www.uniprot.org/uniprot/P04114`). The variables selected by the model are presented in Table 5.8 and further explained in Tables Γ′.1 and Γ′.2. Figure 5.5 depicts graphically the predicted vs. observed values for the training and test samples.

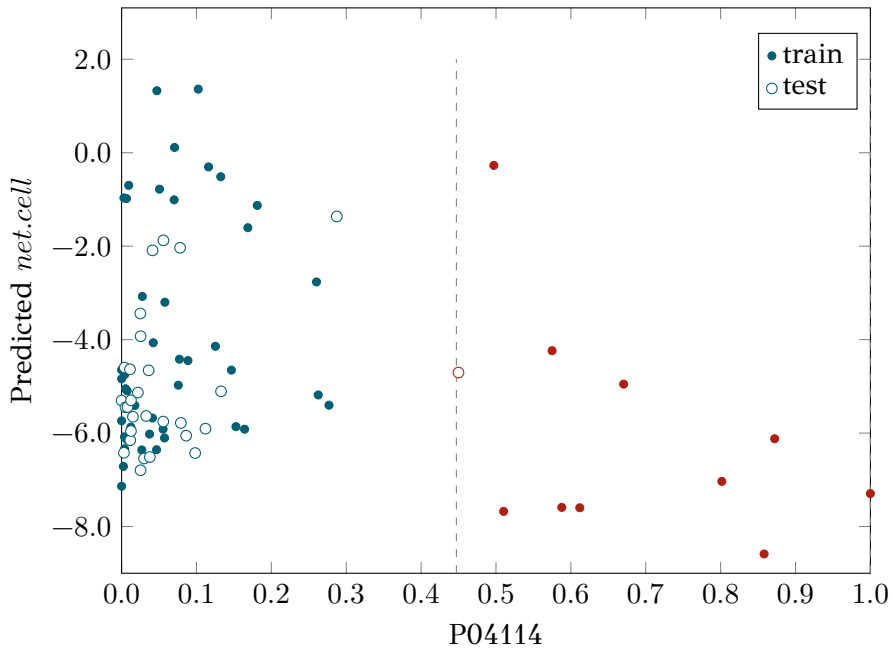


Figure 5.4: Breakpoints, and region distribution of the *Gold ENMs* samples, resulted by application of the grouping/read-across 1D MILP workflow. Blue markers indicate the samples belonging to region A and red markers indicate the samples belonging to region B. Training samples are represented by circular markers whereas, test samples are represented by triangular markers.

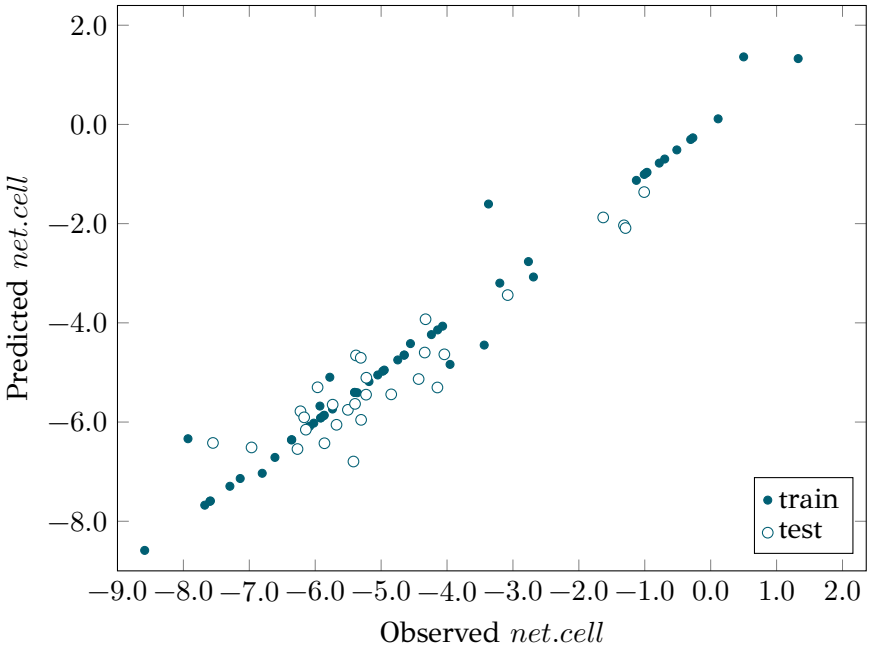


Figure 5.5: Plot of experimental versus predicted $net.cell$ values produced by the application of the grouping/read-across 1D MILP problem to the *Gold ENMs* dataset. Training samples are represented by circular markers whereas, test are represented by triangular markers.

$$
net.cell = \begin{cases}
0.625 \cdot lspri.serum + 0.523 \cdot lspri.relative + 0.333 \cdot zav.serum \\
+2.443 \cdot int.serum + 0.716 \cdot hdrel.serum - 0.337 \cdot vol.ch \\
+0.505 \cdot pdi.rel - 2.607 \cdot zp.rel + 1.610 \cdot zp.synt.sign \\
-0.793 \cdot zp.serum.mag - 1.131 \cdot AS.total - 0.508 \cdot P01024 \\
+0.151 \cdot P00734 - 0.009 \cdot P05154 + 1.577 \cdot P19823 + 0.358 \cdot P12259 \\
+0.384 \cdot P10720 - 0.041 \cdot P68871 + 0.756 \cdot O43866 + 2.653 \cdot P02654 \\
-0.718 \cdot P03952 + 0.063 \cdot P18428 - 0.177 \cdot P02655 - 0.356 \cdot P00751 \\
-0.158 \cdot P02790 + 0.251 \cdot P18065 + 0.588 \cdot P08567 + 0.172 \cdot P01019 \\
+0.576 \cdot P02671 + 0.444 \cdot P00451 + 0.095 \cdot P14618 - 0.373 \cdot P23528 \\
+0.467 \cdot Q99467 - 4.814 \quad \text{if } P04114 \leq 0.447 \text{ region A} \\
\\
-2.600 \cdot pdi.serum - 0.820 \cdot int.rel + 0.052 \cdot zp.synth.mag \\
+2.338 \cdot P01009 + 0.934 \cdot P02749 - 3.186 \cdot P02655 + 4.722 \cdot P27169 \\
+0.746 \cdot P01019 - 5.327 \quad \text{if } P04114 > 0.447 \text{ region B}
\end{cases} \tag{5.36}
$$

The training *Gold ENMs* that belong to regions A and B are presented in Table 5.9. For each unknown ENM belonging to region A or B, the training ENMs of that region are its neighbours. We can observe that there is a possible pattern between the ENMs of each group: the ENMs modified with 1-dodecanethiol (DDT) are assembled in Group B thus, we can assume that behave similarly as far as cell association levels are concerned.

The results of the application of the LOO cross-validation and the Y-randomisation methods on the best predictive model, are presented in Table 5.10 and Table 5.11 respectively. Similarly to the *MWCNTs [a]* case study, the high Q^2_{LOO} value and the large predictive errors obtained in the Y-randomisation process illustrate the reliability of the model. Five random partitions of the dataset into training and test sets were used to train additional predictive

models, which are presented in Table 5.12 and further illustrate the reliability of the proposed methodology.

Table 5.8: Variables involved in the 1D *Gold ENMs* model.

Role	**Symbol**		
Endpoint	*net.cell*		
Partition feature	P04114		
MLR variables	lspri.serum	P01024	Q99467
	lspri.relative	AS.total	P01009
	zav.serum	P00734	P02749
	pdi.serum	P05154	P00751
	int.serum	P19823	P01019
	hdrel.serum	P12259	P23528
	vol.ch	P68871	P08567
	pdi.rel	P10720	P18065
	int.rel	O43866	P00451
	zp.rel	P18428	P27169
	zp.synth.sign	P03952	P14618
	zp.synth.mag	P02655	P02790
	zp.serum.mag	P02654	P02671

Table 5.9: Groups of *Gold ENMs* training samples as produced by the application of the 1D MILP workflow

	Training *Gold ENMs*
Region A	G15.AC, G15.AHT, G15.CALNN, G15.CTAB, G15.DDT@DOTAP, G15.DTNB, G15.F127, G15.Gly-SH, G15.MES, G15.Met-SH, G15.MHDA, G15.MSA, G15.MUTA, G15.NT@PSMA-EDA, G15.NT@PSMA-Urea, G15.ODA, G15.T20, G15.SA, G15.PAH-SH, G15.PLL-SH, G15.PVA, G15.PVP, G15.SPP, G15.TP, G30.AC, G30.CFGAILS, G30.DDT@DOTAP, G30.LA, G30.MAA, G30.MUA, G30.MUTA, G30.PAH-SH, G30.Thr-SH, G30.TP, G60.AUT, G60.CIT, G60.CTAB, G60.CIT, G60.CVVIT, G60.DDT@DOTAP, G60.DTNB, G60.MBA, G60.MUTA, G60.ODA, G60.PVA, G60.Trp-SH
Region B	G15.DDT@BDHDA, G15.DDT@CTAB, G15.DDT@ODA, G15.DDT@SA, G15.DDT@SDS, G15.HDA, G15.NT@DCA, G30.DDT@BDHDA, G30.DDT@CTAB, G60.DDT@BDHDA

Table 5.10: LOO cross-validation results of the 1D MILP workflow applied on the *Gold ENMs* dataset for $\lambda = 0.01$.

λ	**MAE**	**RMSE**	$Q^2_{\mathbf{LOO}}$
0.01	0.88	1.30	0.66

Table 5.11: Y-randomisation results of the 1D MILP workflow applied on the *Gold ENMs* dataset for $\lambda = 0.01$.

		# y-randomised endpoint				
		1	**2**	**3**	**4**	**5**
	Regions	2	2	2	2	2
	REG	63.76	55.67	58.37	85.28	93.03
	Selected variables	31	42	37	37	39
	Selected variables per region$_r$	22|12	31|11	6|33	34|8	32|11
Training	**MAE**	0.43	0.32	0.59	0.27	0.17
set	**RMSE**	0.93	0.96	1.40	0.71	0.58
55	R^2	0.86	0.85	0.68	0.92	0.95
samples	**ENMs per region$_r$**	41|14	42|13	7|48	46|9	41|14
Test	**MAE**	1.63	2.21	3.30	2.00	1.22
set	**RMSE**	2.41	2.63	4.00	2.30	1.63
29	$Q^2_{\mathbf{ext}}$	-0.98	-1.37	-4.47	-0.81	0.10
samples	**ENMs per region$_r$**	22|7	27|2	2|27	27|2	19|10

Table 5.12: Results of the 1D MILP workflow applied on the *Gold ENMs* dataset for $\lambda = 0.01$ and for different random train-test set partitions.

		# random partition				
λ **= 0.01**		**1**	**2**	**3**	**4**	**5**
	Regions	2	2	2	2	2
	REG	36.20	28.11	34.66	36.07	28.11
	Selected variables	25	36	33	37	36
	Selected variables per region$_r$	15|15	35|2	26|9	32|8	35|2
Training	**MAE**	0.07	0.19	0.13	0.08	0.19
set	**RMSE**	0.17	0.49	0.29	0.24	0.49
55	R^2	0.99	0.95	0.99	0.99	0.95
samples	**ENMs per region$_r$**	29|26	52|3	44|11	46|9	52|3
Test	**MAE**	1.04	0.81	0.71	0.70	0.81
set	**RMSE**	1.35	1.21	0.83	0.96	1.21
29	$Q^2_{\mathbf{ext}}$	0.65	0.72	0.75	0.80	0.72
samples	**ENMs per region$_r$**	22|7	28|1	29|0	27|2	28|1

5.5.2 [r] Results of the 2D model

Between the two case studies, the *Gold ENMs* dataset contains two different types of descriptors and is suitable for the application of the 2D algorithm. More specifically, category M contains the physicochemical descriptors, and category N includes the biological descriptors. We applied all different approaches for the selection of partition features, and the impact of different λ values was tested, with $\beta = 0.05$. The sequential approach produced the best models in terms of predictive accuracy. The results for different values of the regularisation parameter are summarized in Table 5.13 and include the number of resulting regions, the total number of selected variables, the accuracy metrics for both training and test sets, the number of selected variables per region $r_{\mathrm{M}}r_{\mathrm{N}}$, and the ENMs belonging to each region $r_{\mathrm{M}}r_{\mathrm{N}}$ for both training and test sets.

The partition feature selected from the physicochemical descriptors was the difference between the Intensity mean HD after serum exposure and the Intensity mean HD after synthesis

Table 5.13: Results of the 2D MILP workflow using the sequential approach applied on the *Gold ENMs* dataset for different choices of the regularisation parameter, λ.

		λ					
		0.000	**0.005**	**0.010**	**0.020**	**0.030**	**0.050**
	Regions	4 (2×2)	4 (2×2)	4 (2×2)	4 (2×2)	4 (2×2)	4 (2×2)
	REG	65895.00	51.94	29.88	22.50	10.82	1.61
	Selected variables	100	43	36	23	14	3
	Selected variables per region$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	41|12	31|18	33|3	23|1	14|1	3|0
		47|48	2|0	2|0	0|0	0|0	0|0
Training	**MAE**	0.00	0.00	0.17	0.27	0.58	0.89
set	**RMSE**	0.00	0.00	0.53	0.57	1.11	1.33
55	R^2	1.00	1.00	0.96	0.95	0.81	0.73
samples	**ENMs per region**$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	21|1	32|19	47|4	47|4	44|7	13|7
		29|4	3|1	3|1	3|1	1|3	32|3
Test	**MAE**	99.35	0.63	0.53	0.60	0.52	0.66
set	**RMSE**	135.46	0.94	0.76	0.77	0.64	0.83
29	$Q^2_{\mathbf{ext}}$	-6275.00	0.70	0.80	0.80	0.86	0.77
samples	**ENMs per region**$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	4|1	16|13	27|2	27|2	28|1	4|1
		21|3	0|0	0|0	0|0	0|0	24|0

(int.rel), while Apolipoprotein B-10 (P04114) was chosen as the partition feature from the group of biological descriptors. Grouping of ENMs in four regions in the two-dimensional space, is presented graphically in Figure 5.6.

The full 2D MILP model for the *Gold ENMs* case study is presented (Eq. 5.37). The MILP model uses 7 physicochemical and 7 biological descriptors (see Tables 5.14, Γ′.1 and Γ′.2). Figure 5.7 presents the experimental vs. predicted values of $net.cell$. The 2D MILP model did not improve the prediction statistics compared to the 1D model. However, the 2D model is much simpler and uses a significantly lower number of descriptors.

$$
net.cell = \begin{cases}
-0.415 \cdot class + 0.113 \cdot lspri.serum \\
+0.094 \cdot zav.serum + 2.178 \cdot int.serum \\
+0.269 \cdot pdi.rel + 2.654 \cdot zp.synth.sign \\
-0.426 \cdot AS.total - 0.140 \cdot P05154 + 1.648 \cdot P19823 \\
-0.364 \cdot P03952 + 0.464 \cdot P00742 + 0.268 \cdot P09871 \\
+0.664 \cdot P20851 - 0.441 \cdot P23528 \\
-5.627 \quad \text{if } int.rel \leq 0.742 \ \& \ P04114 \leq 0.447 \text{ region A} \\
-0.682 \cdot class - 6.122 \quad \text{if } int.rel \leq 0.742 \ \& \ P04114 > 0.447 \text{ region B} \\
0.113 \quad \text{if } int.rel > 0.742 \ \& \ P04114 \leq 0.447 \text{ region C} \\
-7.590 \quad \text{if } int.rel > 0.742 \ \& \ P04114 > 0.447 \text{ region D}
\end{cases} \tag{5.37}
$$

The training *Gold ENMs* that belong to regions A, B, C and D are presented in Table 5.15. For each unknown ENM belonging to region A, B, C or D the training ENMs of that region are its neighbours. Again, the majority of samples in groups B and C have the same modification with DDT.

The results of the application of the LOO cross-validation and the Y-randomisation methods on the best predictive model, are presented in Table 5.16 and Table 5.17 respectively.

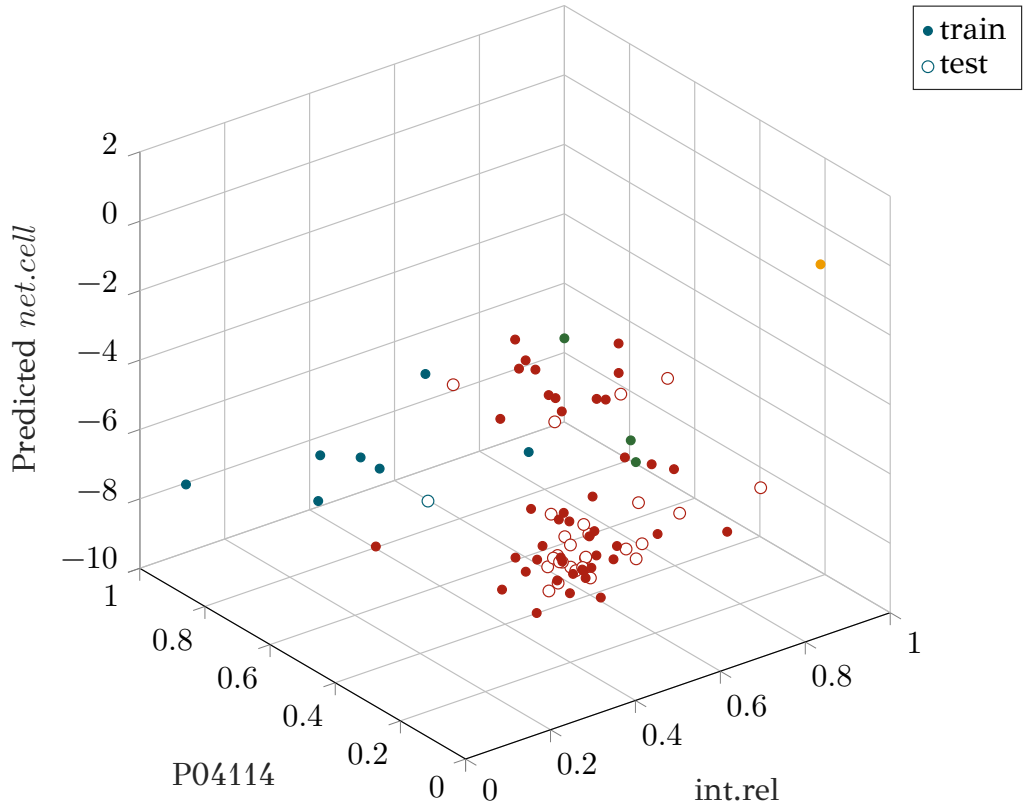


Figure 5.6: Breakpoints, and region distribution of the *Gold ENMs* samples, resulted by application of the grouping/read-across 2D MILP workflow using the sequential approach. Red markers indicate the samples belonging to region A, blue markers indicate the samples belonging to region B, the yellow marker the sample of region C and the green markers the samples of region D. Train samples are represented by circular markers whereas, test samples are represented by triangular markers.

Table 5.14: Variables involved in the 2D *Gold ENMs* model.

Role	Symbol	
Endpoint	*net.cell*	
Partition features	int.rel	
	P04114	
MLR variables	class	P05154
	lspri.serum	P19823
	zav.serum	P03952
	int.serum	P00742
	pdi.rel	P09871
	zp.synth.sign	P20851
	AS.total	P23528

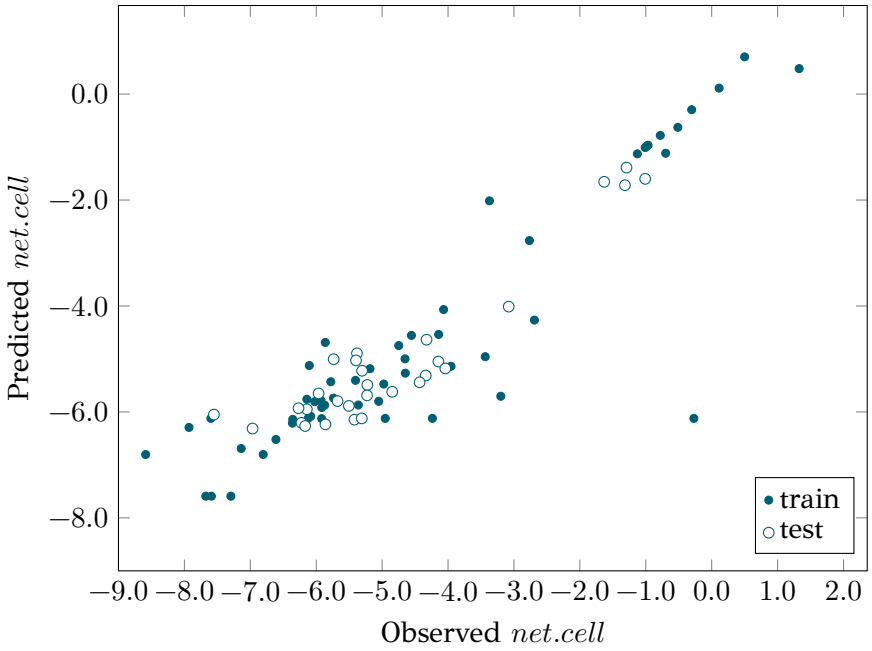


Figure 5.7: Plot of experimental versus predicted $net.cell$ values produced by the application of the grouping/read-across 2D MILP workflow using the sequential approach to the *Gold ENMs* dataset. Training samples are represented by circular markers, while test samples are represented by triangular markers.

Table 5.15: Groups of *Gold ENMs* training samples as produced by the application of the 2D MILP workflow using the sequential approach.

	Training *Gold ENMs*
Region A	G15.TP, G15.Met-SH, G15.DTNB, G15.DDT@DOTAP, G15.CTAB, G15.AC, G15.AHT, G15.CALNN, G15.F127, G15.Gly-SH, G15.MES, G15.MHDA, G15.MSA, G15.NT@PSMA-EDA, G15.NT@PSMA-Urea, G15.ODA, G15.PAH-SH, G15.PLL-SH, G15.PVA, G15.PVP, G15.SA, G15.SPP, G15.T20, G30.AC, G30.CFGAILS, G30.DDT@DOTAP, G30.LA, G30.MAA, G30.MUA, G30.MUTA, G30.PAH-SH, G30.Thr-SH, G30.TP, G60.AUT, G60.CIT, G60.CTAB, G60.CVVIT, G60.DDT@DOTAP, G60.DTNB, G60.MBA, G60.MUTA, G60.ODA, G60.Trp-SH, G60.PVA
Region B	G15.HDA, G15.DDT@ODA, G15.DDT@SA, G15.NT@DCA, G30.DDT@BDHDA, G30.DDT@CTAB, G60.DDT@BDHDA
Region C	G15.DDT@BDHDA, G15.DDT@CTAB, G15.DDT@SDS
Region D	G15.MUTA

Similarly to the previous results, the high Q^2_{LOO} value and the large predictive errors obtained in the Y-randomisation process illustrate the reliability of the model. Five random partitions of the dataset into training and test sets were used to train additional predictive models, which are presented in Table 5.18 and further illustrate the reliability of the proposed methodology.

Table 5.16: LOO cross-validation results of the 2D MILP workflow using the sequential approach applied on the *Gold ENMs* dataset for $\lambda = 0.03$.

λ	**MAE**	**RMSE**	Q^2_{LOO}
0.03	0.85	1.25	0.69

Table 5.17: Results of the 2D MILP workflow using the sequential approach applied on y-scrambled *Gold ENMs* dataset for $\lambda = 0.03$.

		# y-randomised endpoint				
		1	**2**	**3**	**4**	**5**
	Regions	4 (2×2)	4 (2×2)	4 (2×2)	4 (2×2)	4 (2×2)
	REG	11.07	11.07	10.28	7.70	5.59
	Selected variables	13	13	9	10	14
	Selected variables per region$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	13|0	13|0	9|0	10|0	14|0
		0|0	0|0	0|0	0|0	0|0
Training	**MAE**	1.25	1.25	1.01	1.37	1.38
set	**RMSE**	1.85	1.85	1.56	2.07	2.15
55	R^2	0.46	0.46	0.62	0.36	0.29
samples	**ENMs per region**$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	53|1	53|1	49|3	53|1	48|3
		1|0	1|0	3|0	1|0	3|1
Test	**MAE**	1.44	1.44	2.39	2.64	1.68
set	**RMSE**	2.03	2.03	2.73	3.39	2.25
29	$Q^2_{\mathbf{ext}}$	-0.41	-0.41	-1.55	-2.93	-0.73
samples	**ENMs per region**$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	27|1	27|1	27|2	22|7	26|3
		1|0	1|0	0|0	0|0	0|0

Table 5.18: Results of the 2D MILP workflow using the sequential approach applied on the *Gold ENMs* dataset for $\lambda = 0.03$ and for different random train-test set partitions.

		# random partition				
λ **= 0.03**		**1**	**2**	**3**	**4**	**5**
	Regions	4 (2×2)	4 (2×2)	6 (2×3)	4 (2×2)	4 (2×2)
	REG	11.23	5.51	10.64	11.23	5.51
	Selected variables	14	11	14	14	11
	Selected variables per region$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	13|1	9|0	14|0|0	13|1	11|0
		0|0	3|0	0|0|0	0|0	1|0
Training	**MAE**	0.44	0.63	0.46	0.42	0.61
set	**RMSE**	0.88	1.13	0.95	0.88	1.13
55	R^2	0.86	0.77	0.86	0.86	0.77
samples	**ENMs per region**$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	45|7	34|3	43|5|3	45|7	34|3
		1|2	11|7	1|2|1	1|2	11|7
Test	**MAE**	0.67	0.76	0.53	0.67	0.76
set	**RMSE**	0.95	0.90	0.68	0.95	0.90
29	$Q^2_{\mathbf{ext}}$	0.81	0.79	0.83	0.81	0.79
samples	**ENMs per region**$_{r_{\mathrm{M}}r_{\mathrm{N}}}$	27|1	22|0	29|0|0	27|1	22|0
		0|1	6|1	0|0|0	0|1	6|1

5.5.3 [r] Comparison with other models reported in the Literature and other techniques

For the *MWCNTs [a]* case study, the proposed approach was compared first to the model presented by Xia *et al.* [85] The Q^2_{LOO} metric reported by Xia *et al.* [85] was 0.923 which is slightly higher than the result produced by the MILP method (Table 5.5). The MILP model developed and tested with the Kennard-Stone data partitioning method was compared next with the Apellis GA for read-across (§4.5), with the *k*NN and with the LASSO algorithm (see also page 131). [123] The model built using the Apellis/GA workflow selected 3 features and the Q^2_{ext} metric was equal to 0.805 (see also page 70), which is lower compared to the results produced by the MILP model (Table 5.1).

In LASSO, the hyper-parameter α, which multiplies the ℓ_1 term, was optimised to the value of 0.043. The produced model selected 4 variables and the Q^2_{ext} metric was equal to 0.68. Finally, in the *k*NN model, using three neighbours and 4 features (see also page 69), the Q^2_{ext} metric was equal to 0.808, which is again lower compared to the results produced by the MILP model.

For the *Gold ENM* case study, the proposed MILP workflow was compared to previous results on the same dataset, with the *k*NN and with the LASSO method. The application of the our previously developed GA (§4.3.6.2) produced ten models where the Q^2_{ext} metric varied between 0.69-0.79 using one similarity criterion and between 0.72-0.86 using two similarity criteria (physicochemical and biological). The number of selected variables varied between 50-57 using one similarity criterion and 47-62 using two similarity criteria. For some samples in the test set, the GA workflow was not able to provide predictions due to the absence of close enough neighbours in the training set. In the LASSO algorithm, the hyper-parameter α was optimised to the value of 0.0055. The produced LASSO model selected 8 variables and the Q^2_{ext} metric was equal to 0.73 (see Table 6.2).

In the *k*NN algorithm (see page 69), using three neighbours and 11 features, the Q^2_{ext} metric was equal to 0.723.

By comparing these results with the results reported in Table 5.2 and Table 5.13, we can observe that the proposed MILP method is superior to previous approaches, because it improves the Q^2_{ext} metrics, selects fewer variables and is able to provide predictions for all testing samples.

Overall, the results in both case studies illustrated that the performance of the proposed MILP method is comparable or outperforms other alternative predictive modelling techniques. Taking also into account that the grouping, feature selection and model generation steps are fully automated, we can conclude that the proposed method can be considered as a promising new approach in the field of grouping/read-across modelling.

5.6 [m] Web implementation

All three models (1D model for the *MWCNTs [a]* and 1D and 2D MILP models for the *Gold ENMs*) are available through the vythos web service, which has been integrated into the Jaqpot e-infrastructure (`https://vythos.jaqpot.org/`) under GNU General Public License v3. vythos is a ready-to-use and user-friendly application implemented with R `shiny` (§A′.1.2.1) that produces read-across predictions following only a few simple steps. vythos web application was developed in R, using `shiny` v.1.4.0 package. The principal components of the R code are presented in the next paragraphs.

In order to facilitate the user experience, through the `ui.R` file we incorporated only the necessary parameters that create an elegant and intuitive environment for model use. The application consists of a single tab and the user with three steps, can acquire predictions for her/his query ENM samples:

1. Model selection
2. Data uploading
3. Predictions generation

To lead-help the user to insert the necessary data for the analysis, some features are enabled/disabled or hidden/shown, and pop-up warning messages appear according to the provided information. The main packages used are the same as in §4.5.1.

Each particular model is defined as an independent function included in the `server.R` file. This file also includes the necessary processes for prediction calculation and presentation of the

results. Due to its code simplicity, vythos app, can be easily extended with models developed using the workflow described in the previous paragraphs transformed in R functions, to facilitate their dissemination to the scientific community.

5.6.1 [m] Deployment

In order to "package" vythos application and deliver it in any server without any dependency impediments, we used again Docker (see §A′.1.5) to develop a portable version of it. All the necessary actions were performed on Linux OS.

The *Dockerfile* was similar to the one presented in §4.5.1.1 and following a similar procedure for the creation of the docker *image*, we *pushed* the final *image* to Docker Hub.

Interested users can *pull* the *image* from the following link: `hub.docker.com/r/demetradanae/vythos` (see Listing 5.1). Later, as presented in Listing 5.2 users can run vythos application locally.

```
docker pull demetradanae/vythos
```

Listing 5.1: Command for vythos download through Docker.

```
docker run --rm -p 3838:3838 demetradanae/vythos
```

Listing 5.2: Command for local execution of vythos through Docker.

5.6.2 [m] The vythos web application

To initiate the prediction process users must select one of the provided datasets from the *Select dataset* drop-down menu: *CNTs adsorption coefficients* that corresponds to the *MWCNTs [a]* set or *Gold ENMs toxicity* that corresponds to the *Gold ENMs* set. The users must also select the type of model they want to use; the 1D or 2D MILP model from the corresponding radio-buttons. Information about the required input variables, the endpoint predicted by the models and the external explained variance (Q^2_{ext}) is provided to the user (Figure 5.8).

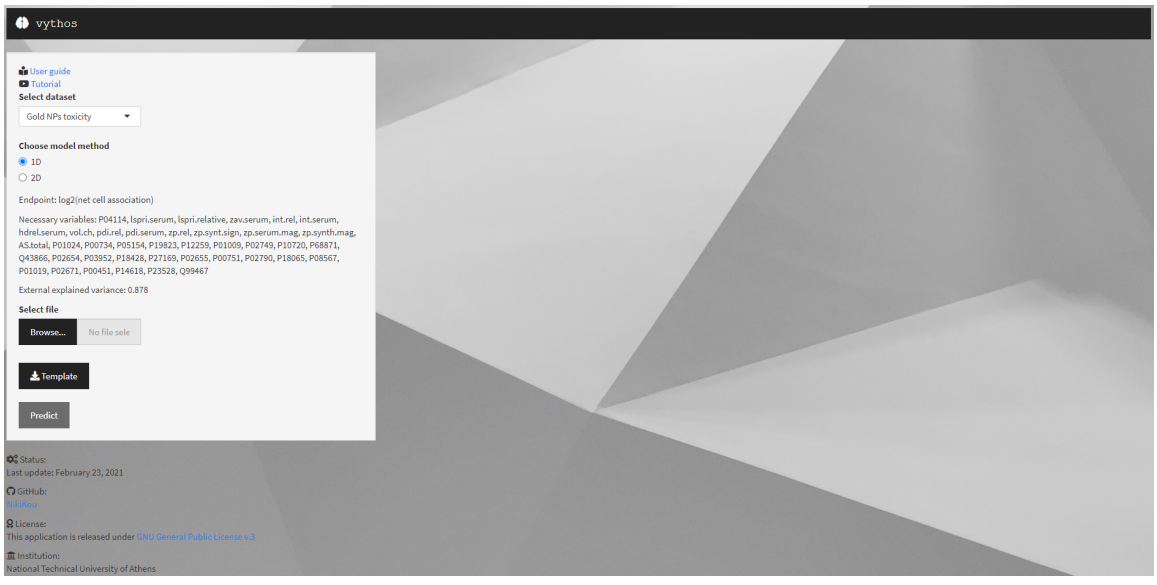


Figure 5.8: The user interface of vythos application. Stakeholders can choose one of the available models from the drop-down menu, and upload their data according to the template. Results are presented on the right-hand side. At the top of the page a link for a user-guide and a video tutorial are available.

Users should upload one CSV file containing the dataset of interest by clicking on the *Browse* button in the *Select file* field. The file must contain the values of the necessary descriptors (in columns), including the ENMs names in the 1st column. Missing values cannot be handled by this approach. Users are advised to download and fill the template input file different for each model that can be downloaded from the app by clicking on *Template* button.

Input data are automatically normalised between 0-1, according to Eq. 2.1. When a dataset is uploaded, by pressing the *Predict* button, the prediction process starts, according to the regions where each input sample belongs, otherwise the corresponding button remains disabled till the necessary file is provided.

The application produces a table containing the toxicity index predictions for each of the unknown ENMs, the group where they belong and if the ENMs are within the domain of applicability of the model. The application also produces a graph where the unknown samples are positioned in space, along with the training ENMs. The "space" is defined by the partition feature(s) and the toxicity index values thus, it can be a 2D graph in the case of one breakpoint (1D MILP model) model (Figure 5.9) or a 3D graph (2D MILP model) in the case of two breakpoints (Figure 5.10). The space is defined by and the toxicity index values, and the different groups of ENMs are depicted in color code. Both the table and the diagram can be downloaded.

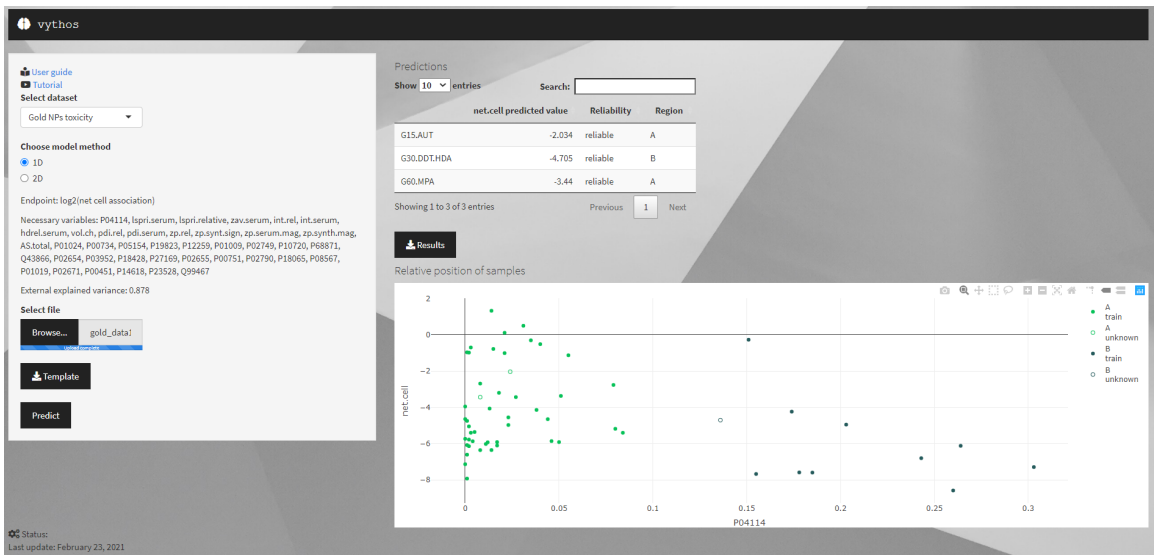


Figure 5.9: The produced results of vythos application running 1D-*Gold ENMs* set. The predictions for *net.cell* are presented along with the 2D regions plot.

5.7 Chapter summary

In this Chapter, the existing OPLRAreg methodology [119] was extended and adapted to the special requirements of nanotoxicity in order to first group ENMs and then predict toxicity related endpoints. The important advantages of the proposed workflow are: the automation of searching for the optimal grouping hypothesis, the selection of the most important input features and the possibility to take into account the multi-perspective characterisation of ENMs ("1D" and "2D MILD problem"). The produced grouping of ENMs is well-defined in terms of the partition feature(s), the grouping boundaries and the selected variables. In each group, read-across predictions are produced using the local multi-variable linear regression functions, which are also results of the MILP problem solution.

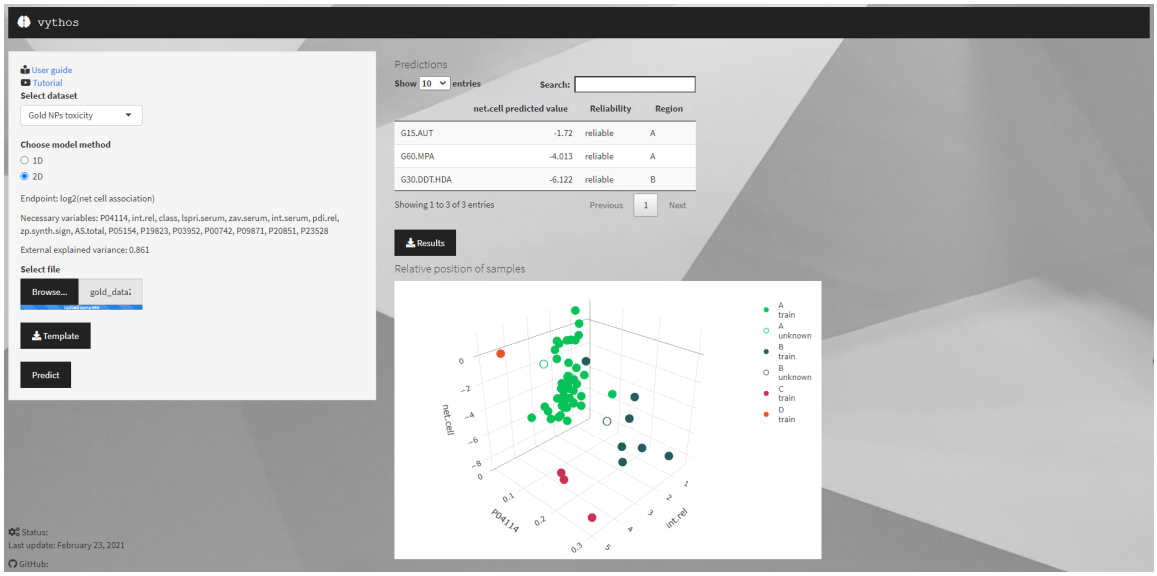


Figure 5.10: The produced results of vythos application running 2D-*Gold ENMs* set. The predictions for *net.cell* are presented along with the 3D regions plot.

The main features of this method are the following:

- Optimal read-across grouping based on similarities between ENMs.
- Automation of the grouping procedure without any need of the formulation of a prior grouping hypothesis.
- Inclusion of the multi-perspective characterisation of ENMs in grouping and prediction procedures.
- Automated variable selection.
- Control of the total number of selected variables (in order to prevent over-fitting) by the inclusion of a regularisation parameter.
- Development of a web application named vythos where the developed models are hosted.
- The application is freely available to the scientific community via: `https://vythos.jaqpot.org/`.

The proposed methodology was applied to two datasets containing different types of descriptors. One of these datasets was suitable for testing the full functionality of the proposed method. The performance of the method was illustrated, through the development of models that are comparable or improve the results of other modelling approaches. The present workflow, based on the concept of mathematical programming, can reduce dramatically the time required to arrive to reliable grouping and accurate read-across predictions and thus contribute to the effort of designing safer ENMs.

Chapter 6

Grouping/read-across modelling based on optimal partition of the response variable space

In this Chapter we are presenting a new grouping/read-across, which – in contrast to the method described in the previous Chapter - creates groups of similar ENMs based on the response variable and not on input variables.

Similarly to the previous approaches, the method formulates and solves a mathematical optimisation problem, which searches over the space of alternative hypotheses and converges to the one providing the most accurate predictions. Variable selection is performed simultaneously to the definition of the grouping boundaries. The type of the developed optimisation problem is MILP. Compared however to the model presented in Chapter 5, this new model is simpler faster because it does not involve the need of searching for the optimal partition feature.

The MILP grouping method is incorporated in an automated predictive workflow (§6.3) that indicates the most appropriate predictive model between MLR, LASSO and the proposed grouping/read-across strategy.

6.1 [m] Development of the grouping/read-across methodology

The grouping procedure is performed in of two consecutive steps: First a MILP problem is formulated, which includes, variable selection, the definition of the grouping boundaries with respect to the response variables and the development of linear regression models specific to each group of ENMs. The solution of the MILP problem is provided to the second step of the method, which computes the centroids in the input space, corresponding to each group of ENMs. The method is described in the next paragraphs (§6.1.1 and 6.1.2) and is integrated in a workflow presented in §6.3, which searches for the optimal number of groups.

6.1.1 [m] Step I: Formulation of MILP problems for grouping and generating local predictive models in each group

In this step we assume the number of groups is known. As mentioned before a MILP problem is formulated which searches for the grouping boundaries (based on the response variable) and generates multi-variable regression models using a subset of selected input variables in each region. The methodology is presented below:

6.1.1.1 Data

A dataset comprising the values of F descriptors and the endpoint for S ENMs is needed. The dataset is represented by A_{sf}, a matrix containing the values of the F descriptors of the S ENMs and Y_s, the vector of the endpoint values of the S ENMs. For this method the data should be first scaled in the range [0,1].

6.1.1.2 Variables

The indices used in the formulation of the MILP problem are:

- s: samples, $s = 1, 2, \dots, S$
- f: features, $f = 1, 2, \dots, F$
- r: regions, $r = 1, 2, \dots, R$

The parameters used for the formulation of the MILP problem are:

- U, U': arbitrarily large positive numbers
- λ: the regularisation parameter of the OF
- ε: the smallest difference between complementary breakpoints

The continuous variables associated with the model are:

- Y_*^r: breakpoint r on endpoint, $r = 1, 2, \dots, R$
- $Pred_s^r$: predicted output for sample s in region r, $s = 1, 2, \dots, S$, $r = 1, 2, \dots, R$
- W_f^r: regression coefficient for feature f in region r, $f = 1, 2, \dots, F$, $r = 1, 2, \dots, R$
- W_f^{r+}: absolute value of the regression coefficient for feature f in region r, $f = 1, 2, \dots, F$, $r = 1, 2, \dots, R$
- B^r: intercept of regression function in region r, $r = 1, 2, \dots, R$
- E_s: training error in prediction for sample s, $s = 1, 2, \dots, S$
- E_s^r: training error in prediction for sample s in region r, $s = 1, 2, \dots, S$, $r = 1, 2, \dots, R$

The binary variable associated with the model is:

- F_s^r: equal to 1 if sample s belongs to region r, equal to 0 otherwise, $s = 1, 2, \dots, S$, $r = 1, 2, \dots, R$

6.1.1.3 Constraints

At each iteration, the number of regions r is considered known. The breakpoints Y_*^r are consistent and arranged in an ordered way as shown in Eq. 6.1. Eq. 6.2 guarantees that the method will perform data partitioning and Eq. 6.3 sets the breakpoint of the last region equal to 1 given that the data are always normalised between $[0, 1]$.

$$Y_*^r \geq Y^{r-1} + \varepsilon \quad \forall r \in \{2, 3, \dots R\} \tag{6.1}$$

$$Y_*^r \geq \varepsilon \quad r = 1 \tag{6.2}$$

$$Y_*^r = 1 \quad r = R \tag{6.3}$$

The position of sample s in region r is encoded with binary variable F_s^r. If a sample s belongs to region r, then $F_s^r = 1$, otherwise $F_s^r = 0$. Eq. 6.4 guarantees that each sample s belongs to only one region.

$$\sum_{r=1}^{R} F_s^r = 1 \quad \forall s \in \{1, 2, ... S\} \tag{6.4}$$

Eq. 6.5 ensures that every region contain at least one ENM thus, no empty regions will be created.

$$\sum_{s=1}^{S} F_s^r \geq 1 \quad \forall r \in \{1, 2, ... R\} \tag{6.5}$$

Eqs. 6.6 and 6.7 assign samples to the correct region according to the breakpoints Y_*^r and the Y_s numeric value of sample s on the endpoint. U is an arbitrarily large positive number.

$$Y_s \geq Y_*^{r-1} + \varepsilon - U \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, ... S\}, r \in \{2, 3, ..., R\} \tag{6.6}$$

$$Y_s \leq Y_*^r - \varepsilon + U \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R - 1\} \tag{6.7}$$

Eq. 6.8 represents the multi-variable linear model which calculates the predicted value $Pred_s^r$ for sample s in region r, where W_f^r is the regression coefficient of feature f in region r and B^r is the intercept in region r.

$$Pred_s^r = \sum_{f=1}^{F} W_f^r \cdot A_{sf} + B^r \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{6.8}$$

Constraints 6.9 to 6.13 are used to compute the prediction error E_s for sample s, as the absolute difference between the actual endpoint value and the value predicted by the linear model corresponding to the region where samples s belongs, or equal to zero. Eq. 6.9 forces E_s to be a non-negative number.

The positive error in a specific region E_s^r takes values smaller than a large positive number if sample s belongs to region r, and is equal to zero if sample s does not belong to that region (Eq. 6.10). Eqs. 6.11 and 6.12 apply constraints on the absolute error in prediction E_s^r of sample s in region r.

$$E_s \geq 0 \quad \forall s \in \{1, 2, ... S\} \tag{6.9}$$

$$E_s^r \leq U' \cdot F_s^r \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{6.10}$$

$$E_s^r \geq Y_s - Pred_s^r - U' \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{6.11}$$

$$E_s^r \geq Pred_s^r - Y_s - U' \cdot (1 - F_s^r) \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{6.12}$$

$$E_s \geq E_s^r \quad \forall s \in \{1, 2, ... S\}, r \in \{1, 2, ..., R\} \tag{6.13}$$

6.1.1.4 Objective function

The OF (Eq. 6.14) includes two terms; the MAE which penalizes the summation of prediction errors over all samples s, divided by the number of samples and a regularisation term (REG) that reduces the risk of over-fitting by penalizing the summation of the absolute values of all regression coefficients. A parameter λ is used as a weight between the two terms. MAE and REG are defined by the equations shown below:

$$z = MAE + \lambda \cdot REG \tag{6.14}$$

$$MAE = \frac{\sum_{s=1}^{S} E_s}{|S|} \tag{6.15}$$

$$REG = \sum_{r=1}^{R} \sum_{f=1}^{F} W_f^{r+} \tag{6.16}$$

$$W_f^{r+} \geq W_f^r \quad \forall f \in \{1, 2, ... F\}, r \in \{1, 2, ..., R\} \tag{6.17}$$

$$W_f^{r+} \geq -W_f^r \quad \forall f \in \{1, 2, ... F\}, r \in \{1, 2, ..., R\} \tag{6.18}$$

where E_s is the absolute error for each sample s computed in Eqs. 6.9 to 6.13 and,
$|S|$ is the number of samples in the training set and W_f^{r+} indicates the absolute value of regression coefficients W_f^r.

6.1.2 [m] Step II: Computation of centroids for each group of ENMs

The solution of the MILP problem presented in Step I, defines groups of ENMs with respect to the response variables. However, when the model is used for obtaining read-across predictions for untested ENMs, they need to be assigned to a group using only the known input features. This is achieved by computing centroids in the input space for the different groups, as follows:

The selected variables in all regions are joined in a common list (F_{union}). Using these variables and the samples belonging to each region, the centroid ($centr_r = f(F_{\text{union}})$, Eq. 6.19) of each group is calculated as a vector containing the mean value of each selected variable considering the ENMs assigned to this specific group.

$$centr_r = \left(\frac{\sum_{s=1}^{S} A'_{s,1}}{S}, \frac{\sum_{s=1}^{S} A'_{s,2}}{S}, ..., \frac{\sum_{s=1}^{S} A'_{s,F_{union}}}{S} \right) \tag{6.19}$$

where F_{union} are all the selected variables considering all the created regions, $F_{\text{union}} < F$,
A'_{sf} is a matrix containing the values of the F_{union} descriptors of the S ENMs. A'_{sf} is in fact a subset of A_{sf} matrix,
S is the number of ENMs used in training and,
$centr_r$ is a vector containing the coordinates of the centroid of the r region,

After both steps have been completed, the produced read-across model is completely defined and includes the grouping boundaries of the domain, the prediction model in each group and a representative centroid for each group that will be used for the classification of untested ENMs -based on distances- and the prediction of their endpoint of interest.

6.1.3 [m] Step III: Using the grouping/read-across model for performing endpoint predictions

The produced grouping/read-across model is used for the prediction of the endpoint for untested ENMs, as follows: First a distance matrix is created containing all the Euclidean distances between the centroids ($centr_r$) and the ENMs (Eq. 6.20) based in the selected variables defined by F_{union}. Each ENM is assigned to the region which representative centroid has the minimum distance value from the ENM.

$$dist_{s,centr_r} = \sqrt{\sum_{f=1}^{F_{\text{union}}} \left(A'_{sf} - centr_{r,f}\right)^2} \quad s = 1, \dots, S^*, \forall r \in \{1, 2, \dots, R\} \tag{6.20}$$

where F_{union} are all the selected variables considering all the created regions, $F_{\text{union}} < F$,
A'_{sf} is a matrix containing the values of the F_{union} descriptors of the S^* ENMs,
S^* is the number of the untested ENMs,
$centr_r$ is a vector containing the coordinates of the centroid of the r region,
$dist_{s,centr_r}$ is the distance between the s ENM and the region's r centroid.

For each ENM, the endpoint predictions is computed by using the local multi-variable linear model corresponding to the region where the ENM has been allocated.

6.2 [m] Validation

For the validation of the proposed methodology, an external validation approach is again used that divides the full dataset into training and test sets. The Kennard and Stone method is used for data partitioning (§2.3.2.1).

The quality of fit between the predicted and experimental values of the training set and the test sets are expressed by the correlation coefficient R^2 (Eq. 2.6) and the external explained variance metric, Q^2_{ext} (Eq. 2.10), respectively.

The RMSE index was computed on both the training and test sets to further assess the validity and accuracy of the produced models. Eq. 2.9 presents the formula that was used to compute the RMSE index on the training set.

In order to validate the robustness of the method, two additional validation procedures were applied, the LOO cross validation method (§2.3.1) and the Y-randomisation test (§2.3.4). In the LOO method, the explained variance in prediction (Q^2_{LOO}, Eq. 2.11) is calculated.

6.3 [m] Proposed workflow

The proposed methodology can be considered as an extension to standard regression methods, like MLR or LASSO, which can be considered as methods where all ENMs are assigned to only one group. In some cases, the proposed method may result in an over-trained model of reduced accuracy and performance compared to the standard techniques, and therefore MLR or LASSO should be selected as the best-performing model. In this section we describe a computational workflow, called demos, that includes the application of the MLR and LASSO methods and the optimisation of the proposed grouping/read-across method with respect to the number of partition groups. All methods are applied on the training set and are validated on the test set and the best performing model is the final outcome of the workflow. The full workflow is presented schematically in Figure 6.1.

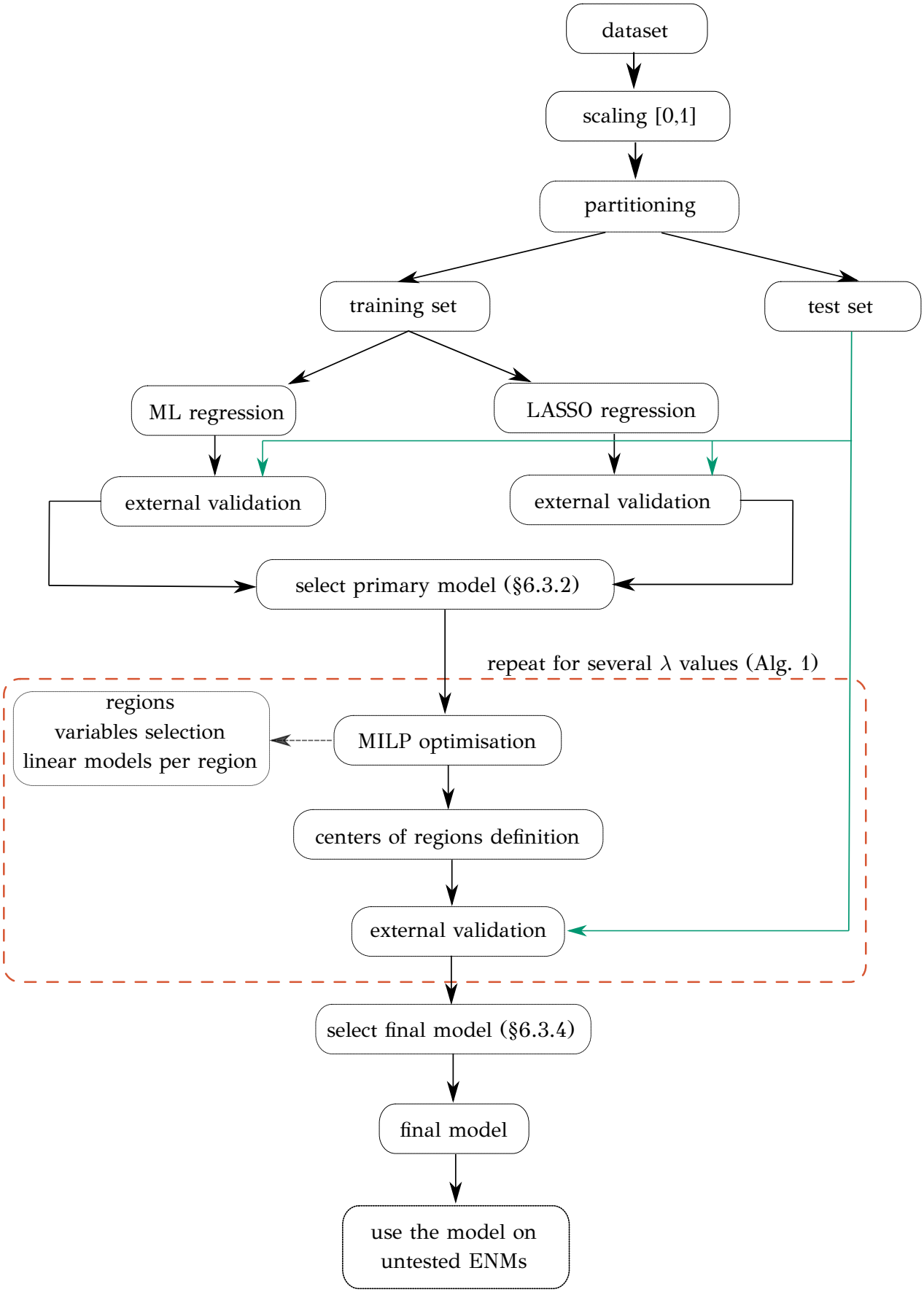


Figure 6.1: Schematic description of the automated methodology for model selection between ML regression, LASSO and grouping/read-across based on the response variable grouping (demos workflow).

6.3.1 Data preprocessing

All the data (descriptors and endpoint) are scaled in the range [0, 1] in order to be comparable, contribute equally to the analysis and to be ready for the application of the MILP grouping problem (§6.1.1). By the end of the training-validation process it is possible to convert the predicted endpoint values to the original range of the endpoint using the Eq. 2.1 solving for X_i. In addition, for external validation purposes, the A_{sf} matrix is partitioned into train and test subsets.

6.3.2 Development of the primary model

This step applies the MLR and LASSO algorithms on the training data without defining multiple regions. The produced models are applied on the test set and the RMSE metrics are recorded ($\text{RMSEtest}_{\text{MLR}}$, $\text{RMSEtest}_{\text{LASSO}}$). The model that produces the lowest RMSEtest value is selected as the primary model.

[t] The LASSO methodology Compared to MLR, the LASSO algorithm adds a regularisation term (see also Chapters 4 and 5) in the OF that is minimized:

$$\min \left\{ \frac{1}{2} \sum_{s=1}^{S} (\hat{y_s} - \sum_{f=1}^{F} W_f \cdot A_{sf} - B)^2 + \alpha \sum_{f=1}^{F} |W_f| \right\} \tag{6.21}$$

where $\hat{y_s}$ is the predicted endpoint value,
W_f is the regression coefficient of feature f,
B is the intercept of the regression model and,
α is the regularisation weight.

The regularisation term is used for penalizing the sum of the absolute values of the linear coefficients (ℓ_1 regularisation) [124] and is weighted by the parameter α, in the OF. The main advantage of the LASSO algorithm is that it shrinks the regression coefficients and reduces some of them to zero values. In principle, the LASSO method selects the most important variables (the ones with non-zero coefficients) and reduces the chance of over-fitting. An iterative fitting procedure is used for the selection of the optimal α parameter based on cross-validation method. The final model produced by both algorithms is described by the equation 6.22 and it is validated on the test set.

$$\hat{y_s} = \sum_{f=1}^{F} W_f \cdot A_{sf} + B \tag{6.22}$$

where, $\hat{y_s}$ is the predicted endpoint value,
W_f is the regression coefficient of feature f and,
B is the intercept of the model.

In our computational experiments, we used the `scikit-learn` [125] implementation of the MLR and the LASSO methods.

6.3.3 Grouping/read-across model optimisation

In this step, the proposed grouping/read-across problem is optimised in terms of the number of groups defined in the response domain, using an iterative procedure. The starting point in this iteration is the model produced in the previous step, (i.e. the best performing model between LASSO and MLR), which is considered as the single group model. The value of the OF (Eq. 6.14) is calculated and is denoted by $z_{\text{R=1,test}}$. The method described in §6.1.1

and 6.1.2 is applied considering two groups and the produced model is validated on the test set. The corresponding value of the OF (Eq. 6.14) is denoted by $z_{\text{R=2,test}}$. Eq. 6.23, where R is set equal to 1, is applied in order to decide if the iterative procedure will continue with defining more groups in the response domain or the algorithm will be terminated. More specifically, if Eq. 6.23 is not satisfied, the algorithm stops, and the single-group model is the final model of the iterative algorithm. Otherwise, the two-group model is accepted and the iterative algorithm proceeds with partitioning the ENMs into three groups. This process is iterated until the improvement in the value of z_{test} in consecutive iterations does not satisfy Eq. 6.23.

$$z_{\text{R,test}} \leq (1-\beta) \cdot z_{\text{R-1,test}} \tag{6.23}$$

where β a user defined parameter in the region (0,1) which represents the minimum % improvement in the OF that needs to be achieved in order to further iterate the algorithm and,
$z_{\text{R-1,test}}$ and $z_{\text{R,test}}$ are the OF values for the test set between two consecutive iterations.

An outer iteration loop optimises the algorithm in terms of the regularisation parameter λ: Lower and upper bounds λ_{min}, λ_{max} on the selection of the parameter λ are defined, as well as the increment value $definedStep$. Starting with λ_{min}, the parameter λ is increase by $definedStep$ in each iteration, until the upper bound is reached. The two-loop optimisation procedure is presented in detail in Alg. 1. The initial and final λ values, along with the iterative step are user-defined parameters.

The minimum RMSE metric is denoted by $\text{RMSEtest}_{\text{group}}$ and the corresponding model is the final model of the iterative procedure.

Algorithm 1: Grouping/read across model optimisation

Result: Grouping/read across models
define λ_{min}, λ_{max}, *definedStep*, ϵ, U, U', β;
for $\lambda \in [\lambda_{\text{min}}, \lambda_{\text{max}}]$ *with definedStep* **do**
  set $R = 1$;
  apply the best one-group model on the test set, compute $z_{\text{1,test}}$;
  solve MILP problem considering 2 regions (§6.1.1);
  define representative centroids of groups (§6.1.2);
  apply on test set, compute $z_{\text{2,test}}$;
  while *Eq. 6.23 is satisfied* **do**
    set $R = R + 1$;
    solve MILP problem (§6.1.1);
    define representative centroids of groups (§6.1.2);
    apply on test set, compute $z_{\text{R+1,test}}$;
  end
  save the model with R groups as the grouping/read-across model corresponding to the specific λ value;
  calculate validation statistics including RMSE metric on the test set using all saved model (corresponding to different values of λ);
end

6.3.4 Selection of the final model

The $\text{RMSEtest}_{\text{MLR}}$, $\text{RMSEtest}_{\text{LASSO}}$, $\text{RMSEtest}_{\text{group}}$ values are compared in the final step of the workflow. The lowest value defines the final model among the MLR, LASSO and the proposed grouping/read-across approach.

6.4 [m] Domain of applicability

In this methodology, where -in any case- multiple linear equations constitute the final model, a prediction is considered reliable only if all the independent descriptors are within the ranges defined by the training samples. Therefore, before using the final model for predicting the endpoint of an untested ENM, the input variables should be scaled first according to the original dataset's min-max values. Only if all scaled values are within the range [0,1], the read-across prediction can be accepted.

6.5 [m] Implementation

The implementation of the presented methodology has been entirely performed in the Python programming language, using the Anaconda Navigator and Jupyter Notebooks. The grouping/read-across problem (§6.1.1) is solved using the `mip` module and CBC (Coin-or branch and cut) solver. The MLR and the LASSO regressions are solved using the corresponding `sklearn` functions. The `multiprocessing` module is employed in order to achieve parallelisation and rapid solution of the MILP problems.

6.6 [r] Results and discussion

The proposed workflow was applied on the *Gold ENMs* dataset that was priory normalised according to Eq. 2.1. The dataset was first partitioned into training and test sets using the Kennard-Stone method and the training ENMs were submitted to the developed workflow (§6.3) using the parameters of Table 6.1.

Table 6.1: Values for the user-defined parameters of the demos read-across workflow applied on the *Gold ENMs* dataset.

Parameter	Value
λ_{min}	0.001
λ_{max}	0.07
step	0.001
U	10
U'	$\lvert\sum y_{\text{test}}\rvert$
β	0.05
ϵ	0.05
train ratio	0.66

The best model -in terms of the RMSE validation metric on the test set- using the training ratio of 0.66 was produced by the proposed grouping/read-across method which partitions the ENMs into two groups, using the optimal regularisation parameter $\lambda = 0.001$ (Table 6.2). The model selected 28 independent features, which are presented in Table 6.3 and are further explained in Tables Γ′.1 and Γ′.2.

The endpoint boundaries are presented in Table 6.4: in Region A the normalised *net.cell* obtains values between 0 and 0.6449, whereas in Region B obtains values between 0.6449 and 1. The full predictive model is presented in Eq. 6.24. As described in §6.1.2, the test samples are allocated to the created regions according to their distance from the centroids of Table 6.5.

Table 6.2: Summarized results of the demos workflow applied on the *Gold ENMs* dataset.

	demos workflow steps	
	MLR	
Train set	R^2	1.000
55 ENMs	RMSE	4.13E-16
Test set	Q^2_{ext}	0.4424
29 ENMs	RMSE	0.1304
	LASSO	
	Selected variables	8
	optimal α	0.0055
Train set	R^2	0.8485
55 ENMs	RMSE	0.1027
Test set	Q^2_{ext}	0.7320
29 ENMs	RMSE	0.0904
	primary model	LASSO
	grouping/read-across	
	regions	2
	Selected variables	28
	Selected variables per region$_r$	20 – 11
	optimal λ	0.001
Train set	R^2	0.9387
55 ENMs	RMSE	0.0623
	train samples per region	43 – 12
Test set	Q^2_{ext}	0.8332
29 ENMs	RMSE	0.0713
	test samples per region	25 – 4
	Final model	grouping/read-across

$$net.cell = \begin{cases} 0.037 \cdot lspri.serum + 0.070 \cdot zav.synth + 0.213 \cdot zav.serum \\ +0.028 \cdot vol.synth + 0.054 \cdot zp.synth.sign + 0.242 \cdot P02649 \\ +0.038 \cdot P0C0L5 + 0.090 \cdot P10720 + 0.181 \cdot P02749 \\ +0.279 \cdot P02654 + 0.181 \cdot Q13103 + 0.068 \cdot P00748 \\ +0.017 \cdot P00740 + 0.018 \cdot P03950 + 0.072 \cdot P09871 \\ +0.129 \cdot P27169 + 0.134 \cdot P20851 + 0.024 \cdot P08567 \\ +0.077 \cdot P00451 + 0.110 \cdot Q99467 \quad \text{region A} \\ \\ 0.037 \cdot lspri.serum + 0.198 \cdot vol.synth + 0.056 \cdot int.serum \\ +0.016 \cdot zav.rel + 0.020 \cdot P01009 + 0.049 \cdot P00738 \\ +0.026 \cdot P01011 + 0.037 \cdot P00736 + 0.094 \cdot P01019 \\ +0.024 \cdot P02671 + 0.034 \cdot Q99467 + 0.671 \quad \text{region B} \end{cases} \tag{6.24}$$

Table 6.3: Variables involved in the grouping/read-across model based on the response variable grouping applied on the *Gold ENMs* dataset

Selected variables			
int.serum	P00736	P00738	P00740
lspri.serum	P00748	P01009	P01011
vol.synth	P01019	P02649	P02654
zav.rel	P02671	P02749	P03950
zav.serum	P08567	P09871	P0C0L5
zav.synth	P10720	P20851	P27169
zp.synth.sign	Q13103	P00451	Q99467

Table 6.4: Coordinates of the *Gold ENMs* model's endpoint boundaries derived by the response variable grouping.

net.cell	0	Region A	0.6449	Region B	1

Table 6.5: Coordinates of the *Gold ENMs* model's centroid derived by the response variable grouping.

Descriptor	**Centroid A**	**Centroid B**
lspri.serum	0.248	0.475
zav.synth	0.155	0.155
zav.serum	0.307	0.557
vol.synth	0.146	0.148
int.serum	0.348	0.583
zav.rel	0.318	0.390
zp.synth.sign	0.186	1.000
P02649	0.215	0.042
P01009	0.139	0.515
P0C0L5	0.190	0.565
P10720	0.280	0.032
P02749	0.178	0.000
P02654	0.113	0.000
P00738	0.085	0.145
P01011	0.108	0.312
Q13103	0.060	0.416
P00736	0.108	0.137
P00748	0.118	0.000
P00740	0.093	0.218
P03950	0.204	0.000
P09871	0.077	0.067
P27169	0.094	0.083
P20851	0.100	0.146
P08567	0.063	0.179
P01019	0.032	0.315
P02671	0.105	0.139
P00451	0.068	0.128
Q99467	0.016	0.187

In Table 6.6 the two groups of training *Gold ENMs* samples are presented. This time there is no obvious pattern regarding the coating of the ENMs as in the groups presented in Tables 5.9 and 5.15 created by the feature grouping. However, all the ENMs of Region B (higher cell association levels) have the same surface charge modifications (all "cationic") which is consistent with the Literature. [76]

Table 6.6: Training groups of *Gold ENMs* samples as created by the grouping/read-across model based on the response variable grouping.

	Training *Gold ENMs*
Region A	G15.NT@DCA, G60.DTNB, G15.CTAB, G30.TP, G15.DDT@ODA, G15.LA, G60.PVA, G60.CIT, G15.AC, G15.NT@PSMA-EDA, G15.MES, G60.DDT@BDHDA, G15.DDT@SDS, G60.MBA, G15.CALNN, G15.T20, G15.SA, G15.DDT@BDHDA, G60.CVVIT, G15.DDT@SA, G60.CTAB, G30.Thr-SH, G15.MSA, G15.SPP, G15.PVP, G60.ODA, G30.MAA, G15.DTNB, G15.DDT@CTAB, G15.Gly-SH, G30.CFGAILS, G30.LA, G15.ODA, G15.TP, G30.MUA, G15.MHDA, G60.NT@PSMA-AP, G30.DDT@BDHDA, G15.F127, G30.DDT@CTAB, G60.Trp-SH, G15.PVA, G15.Met-SH
Region B	G15.MUTA, G60.AUT, G60.MUTA, G15.PAH-SH, G15.PLL-SH, G15.HDA, G15.AHT, G30.DDT@DOTAP, G30.PAH-SH, G60.DDT@DOTAP, G15.DDT@DOTAP, G30.MUTA

Figure 6.2 presents graphically the predicted and the actual values for the training and test samples in the two regions. Table 6.7 presents the summarised results of Y-randomisation. The large predictive errors obtained in the random shuffles of the Y-vector, indicate that the model development phase is reliable and that the predictive power of the model is not due to a coincidental outcome.

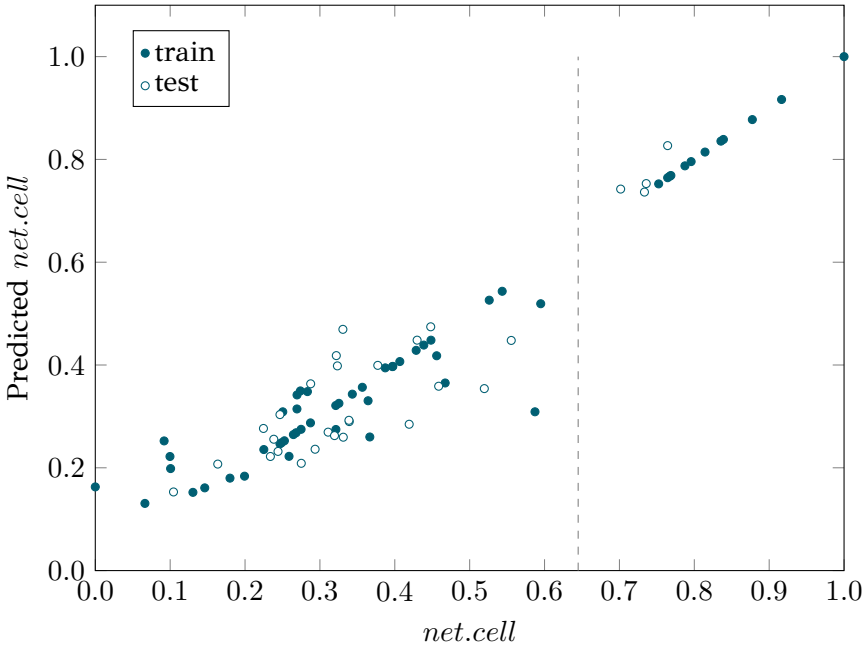


Figure 6.2: Plot of experimental *net.cell* values versus the corresponding predicted values in grouping/read-across based on the response variable grouping methodology using a training ratio of 0.66. Training samples are represented by green markers whereas, test are represented by red markers. The two regions are distinguished by the grey dashed line.

Internal validation was also performed by applying the LOO scheme using the α (LASSO)

and λ (MILP) parameters that occurred from the optimisation workflow (Table 6.2). Results are presented in Table 6.8.

Table 6.7: Y-randomisation results of the demos workflow on the *Gold ENMs* dataset.

		# y-randomised endpoint				
		1	**2**	**3**	**4**	**5**
	α (LASSO)	0.006	0.006	0.006	0.006	0.006
	λ (grouping/read-across)	0.001	0.001	0.001	0.001	0.001
	selected final model	LASSO	LASSO	LASSO	LASSO	grouping/read-across
	regions	1	1	1	1	2
	number of used features	15	17	14	17	31
	features per region	-	-	-	-	[22, 11]
Training	RMSE	0.1940	0.2006	0.1865	0.1918	0.0996
set	R^2	0.5202	0.4680	0.5182	0.5345	0.8431
55 ENMs	train samples per region	-	-	-	-	[43, 12]
Test	RMSE	0.1947	0.2254	0.1993	0.1722	0.1561
set	Q^2_{ext}	-0.2430	-0.6658	-0.3020	0.0273	0.2008
29 ENMs	test samples per region	-	-	-	-	[26, 3]

Table 6.8: LOO results of the demos workflow on the *Gold ENMs* dataset

α (LASSO)	λ (grouping/read-across)	Q^2_{LOO}	RMSE
0.0055	0.001	0.740	0.115

6.6.1 [r] Comparison with other models reported in the Literature and other techniques

The results of the presented case study using the *Gold ENMs* dataset, were compared to the results on the same dataset produced in the previous Chapters and in other publications. The application of the previously developed GA using the same train ratio, produced a model using 27 variables, where the Q^2_{ext} metric was equal to 0.779 using one similarity criterion and a regularisation factor equal to 0.05 (Table 4.12). By comparing these results with the results reported in Table 6.2, we can observe that the proposed workflow is superior to previous approaches, because it improves the Q^2_{ext} metrics using approximately the same number of variables.

Compared to modelling the model built using *k*NN algorithm and 11 variables (see also page 69), the Q^2_{ext} value (equal to 0.723) is lower compared to the Q^2_{ext} value produced by the present model (equal to 0.833).

Compared to the models of Chapter 5, the performance of the current model is slightly deficient. However, in case of the "1D MILP" model, the model produced by grouping based on the endpoint, uses significantly less variables. In terms of internal validation, the developed model using the endpoint grouping methodology (Table 6.8) has better performance in terms of the Q^2_{LOO} statistic than the corresponding models of Chapter 5, where regions are created based on one or two properties respectively (Tables 5.10 and 5.16).

6.7 Chapter summary

A new grouping/read-across strategy is proposed in this Chapter, where each query ENM is allocated according to the shortest distance from the "centroid" of each group. Then a multi-variable linear regression function is applied to predict the endpoint of interest. Variable

selection is integrated in the workflow and it is performed automatically. Also, the multi-variable function formulation and the number of groups is produced automatically needing minimum interaction from the user. The proposed methodology was applied on one benchmark dataset and its performance was illustrated, through the development of models that are comparable to the results of other modelling approaches. The developed workflow is fully-automated and reduces drastically the necessary time for a formulation of a grouping hypothesis.

The main features of the proposed algorithms are summarized next:

- Optimal read-across grouping based on similarities between ENMs.
- Automation of the grouping procedure without any need of the formulation of a prior grouping hypothesis.
- Automated variable selection.
- Control of the total number of selected variables (in order to prevent over-fitting) by the inclusion of a regularisation parameter.

Chapter 7

Development of read-across models based on the *k*-nearest neighbours strategy

In this Chapter we are presenting the application of the *k*-Nearest Neighbours (*k*NN) machine learning algorithm in read-across problems. The *k*NN can be actually considered as a read-across strategy, [53], [126] as it requires experimental observations of only a few neighbours (similar ENMs) to the query ENM, in order to predict the endpoint of interest (§1.3).

7.1 Development of a safe-by-design tool for decorated MWCNTs

Decorated nanostructures have different levels of structural complexity and heterogeneity (presence of inorganic/organic elements and coatings, varying stoichiometry between the particles etc.) and thus extracting quantitative parameters for the characterisation of the structural and chemical properties of the nanostructures is a very challenging task that is not yet fully addressed computationally. The development of *in silico* methods is therefore hindered by the absence of sufficiently large physicochemical, geometrical, structural and biological datasets of different nanostructures in available databases. [58] The hypothesis that decorated nanostructures can be represented by its surface modifiers when the core remains identical, can be considered pragmatic, especially taking into account the near- and long-term hazard and risk assessment goals, and the time and cost required for a full characterisation – experimental and/or computational – of all available nanostructures. This hypothesis has already been accepted and used in different studies found in the Literature. [56], [88], [127], [128]

Fourches *et al*. [87] built and validated classification models for the prediction of the protein binding and cytotoxicity of MWCNTs (see also §3.5, *MWCNTs [b]* dataset), and made the underlying experimental dataset at least partially available for further analysis. These models were based on Molecular Operating Environment (MOE) and Dragon molecular descriptors computed only from the surface-modifying compounds, assuming that the MWCNT core was the same in all samples. Support vector machines, random forest and *k*NN, have been employed as machine-learning techniques, and the reported accepted CCR (Correct Classification Rate, mean of sensitivity and specificity) of the validation sets ranged from 73 to 75% for the protein binding, and from 70 to 77% for the toxicity endpoint.

Singh *et al*. [57] reported an ensemble learning approach based nanoQSAR model for predicting biological effects of ENMs based on molecular descriptors, calculated with Chemistry Development Kit (CDK). Here, the 29 most toxic surface-modified (decorated) MWCNTs from the Zhou *et al*. [86] dataset have been used for the prediction of their impact on cellular viability. For model development, decision tree boost and decision tree forest methods were

implemented based on six molecular descriptors of the decorators. The models resulted in R^2 values of 0.903 and 0.922 respectively. Shao *et al*. [129] used the 29 most toxic samples in order to build QSAR models based on different sets of descriptors. The MWCNT–decorator complex was geometrically optimised using the molecular dynamics simulation package GROMACS with the ffgmx force field. All possible combinations of calculated MOE, VolSurf, and 4D-fingerprints descriptors have been used. MLR and trial QSAR models were built, in a genetic function approximation scheme. For the CA protein binding endpoint, using only the decorators for the descriptor calculations, R^2 and Q^2_{LOO} accuracy was reported as 0.892 and 0.832 respectively, while using the combination of a 10Å nanotube and the decorators, the R^2 and Q^2_{LOO} measures were reported as high as 0.903 and 0.851 respectively. For the cell viability endpoint, using only the decorators, R^2 and Q^2_{LOO} were equal to 0.922 and 0.863 respectively, while using the combination of a 10Å nanotube and the decorators the R^2 and Q^2_{LOO} measures were 0.857 and 0.759 respectively.

These results suggest that depending on the endpoint being modelled, and the role of the core versus surface in the specific interaction, inclusion of both components should be assessed to determine whether the core plays a role or not. Unsurprisingly, in the case of protein binding, a minor contribution from the MWCNTs was found, whereas in the case of toxicity, the surface functionalisation played the dominant role, probably by controlling the amount of cellular adhesion and internalization of the MWCNTs. This reinforces the hypothesis that the decorated MWCNT with the same core can be represented by their surface modifiers for prediction of protein binding and cellular receptor attachment. Given that following the attachment step, ENMs including MWCNTs are actively taken up into THP-1 cells via an active endocytotic process (e.g. phagocytosis), we can safely assume that the particle scaffold (core), which is common to the whole dataset, is the driver once attachment, which is ligand-specific, has occurred, and thus the discrimination in terms of the amount of uptake (and thus toxicity) is driven by the ligands, allowing us to ignore the role of the core.

In this section, the *k*NN method is employed in the development of a fully-validated predictive read-across workflow which can assess the biological and toxicological profile of MWCNTs, based solely on calculated molecular descriptors of the surface decorators. Each MWCNT sample has been evaluated against two different endpoints -protein binding of carbonic anhydrase and toxicity- and was classified as a "binder" or "non-binder" and "toxic" or "non-toxic", respectively. The driving force for adsorption of Human CA II (HCAII) to ENMs has been shown previously to be electrostatic in nature, driven by attraction to negatively charged particle surfaces, and the hydrophobic effect alone was shown not to be strong enough to drive the initial binding at least to positively charged hydrophobic polystyrene ENMs. [130] We used as many of the available MWCNT samples as possible, and not only the most toxic ones as in previous computational studies. The main target of the proposed workflow was to offer a computational tool that will simplify the design and screening of novel MWCNTs by allowing prediction of the CA binding and cellular toxicity based only on the chemical structure of the surface decoration molecule, as part of a safe-by-design strategy that would allow elimination of potentially toxic modifications at the design stage.

7.1.1 [m] Development of the predictive workflow

For each endpoint the full dataset was randomly divided into training and test sets in the proportion 75:25. Due to the abundance of samples, the training set was further divided into calibration and validation sets in a proportion of 50:25 of the initial set. The calibration set was used for variable selection and model development, whereas the validation set was used for the determination of the accuracy of the produced models. The decorators of the test set were excluded from model training, and were used as a blind set to assess the unbiased performance of the model (Figure 7.1).

As many of the molecular descriptors had considerably different numerical ranges, they

were normalised prior to modelling. [6] More specifically, Gaussian normalisation method (Eq. 2.2) was applied on the calculated descriptors of the training set with mean values equal to 0 and standard deviation equal to 1. The normalisation function used for the training set was later applied to the test set.

A variable selection method included in `WEKA` (§B′.2.4.1) was used in order to remove noisy variables and to retain only the ones relevant to each endpoint. The most significant descriptors were selected using the InfoGain variable selection (`InfoGainAttributeEval`) with Ranker evaluator. `InfoGainAttributeEval` measured the attribute's information gain with respect to the current endpoint, whereas Ranker prioritized the variables and removed the lower-ranking ones. [52] In this way the modelling computational time and space were reduced, and the predictive performance was greatly improved.

The *k*NN method was employed for developing the read-across models, using the `EnaloskNN` KNIME node. With this node, apart from the endpoint predictions, we were able to identify the *k* neighbours of each test decorated MWCNT and map the analogous area, as required by the read-across framework. [38]

Among the modelling parameters, an optimal *k* value has been selected, with Euclidean distance between the chosen descriptors and the inverted distance as the weighting factor for the majority vote.

It should be emphasized that even though the modelling was performed for the surface ligands, the biological activities and the toxicity are related to the whole decorated MWCNT structure and not only the surface-modifying compounds, due to the realistic hypothesis that the differences in the biological behaviour of MWCNTs of the same core, were mostly due to the structural characteristics of their surface ligands. [87]

Model validation The read-across models were validated both externally and internally in terms of goodness-of-fit, robustness and predictivity, as recommended by the OECD. [54] For the calibration subset a model was developed, and its performance was tested on the corresponding validation set by computing the following validation metrics: sensitivity, specificity and accuracy (Eqs. 2.21, 2.22, 2.23) and the confusion matrix (Table 2.2).

Moreover, the Y-randomisation (§2.3.4) test was performed in the internal loop, in order to validate the robustness and the statistical significance of the produced models. In addition to the previous validation practices, internal validation was performed in order to reduce the bias produced from a possible unbalanced representation of the two classes between the two subsets. Both for the calibration sets (internal loop) and the training set (external loop), LOO and leave-five-out (L5O) cross-validation methods (§2.3.1) were employed for both models (protein binding and cell viability).

The selected model was finally validated using the external test set by calculating the same accuracy measurements (Eqs. 2.21, 2.22, 2.23).

Applicability domain For the two proposed models the APD threshold was calculated according to Eq. 2.25 (Z in this case was set equal to 0.5). The assessment of the applicability domain of the proposed models was performed in our KNIME workflow, using the `Enalos+ Domain-APD` node that executes the procedure described in §2.4. [69], [90]

7.1.2 [r] Results and discussion

All preprocessing and modelling activities, including the calculation of molecular descriptors, were performed within the freely-available KNIME platform, using the available nodes and the Enalos proprietary nodes developed by NovaMechanics Ltd (see also §A′.1.4).

Two read-across models were trained to classify samples as "binders" and "non-binders" as well as "toxic" and "non-toxic" to assess their CA binding and toxicity.

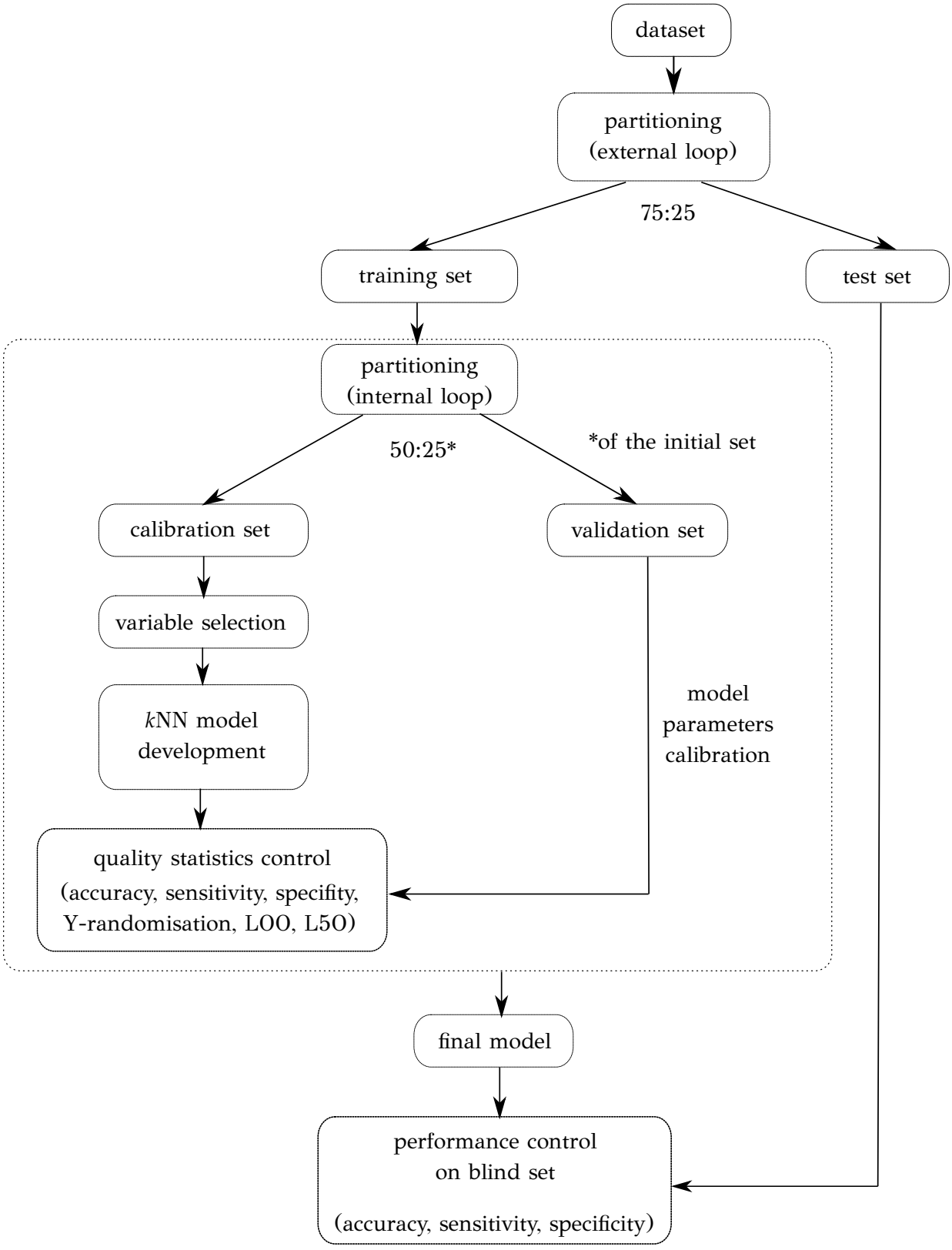


Figure 7.1: Schematic description of the validation workflow used in *MWCNTs [b] k*NN model development. A validation scheme using three data subsets is employed in order develop unbiased read-across models.

Table 7.1: Selected descriptors for the CA binding and the toxicity endpoints of *MWCNTs [b]* dataset, ranked in order of significance.

CA binding		Toxicity	
D522	Mean molecular topological order-2 charge index	D468	Geary topological structure auto-correlation length-6 weighted by atomic Sanderson electronegativities
D473	Geary topological structure auto-correlation length-3 weighted by atomic polarizabilities	D173	Mohar order-2 index
D472	Geary topological structure auto-correlation length-2 weighted by atomic polarizabilities	D454	Geary topological structure auto-correlation length-8 weighted by atomic masses
D269	Information content order-0 index	D254	Radial centric index
D133	Mean value of atomic composition index	D250	EXP5 of Path-distance/Walk-distance over all atoms
D541	Lowest eigenvalue from Burdex matrix weighted by van der Walls order-2	D255	Vertex distance count equality index

Table 7.2: Accuracy statistics of the *MWCNTs [b]* *k*NN/read-across predictive models for the validation and the test sets.

Model	Set	Accuracy	Sensitivity	Specificity
CA binding	Validation	0.750	0.778	0.727
	Test	0.857	0.727	1.000
Toxicity	Validation	0.778	0.778	0.778
	Test	0.842	0.875	0.818

The InfoGain variable selection with Ranker evaluator method (which are included in the `WEKA` node), was applied to the calibration data, to select the most critical, among the 403 available descriptors. The method selected six descriptors for each read-across model, which are depicted in Table 7.1.

The workflow was applied to the calibration data with an optimised value for the number of neighbours equal to 3 for the CA binding model and, equal to 7 for the toxicity model. The values of the accuracy, sensitivity and specificity metrics for the calibration and the test set are presented in Table 7.2.

For comparison purposes, the popular tree classification method (J48 algorithm) was applied to the same calibration-validation and test subsets. As shown in Table 7.3, the validation metrics corresponding to the *k*NN based read-across method are higher compared to the J48 machine learning algorithm, which clearly indicated that *k*NN outperforms J48 in this case study.

Table 7.3: Accuracy statistics of the *MWCNTs [b]* J48 predictive models for the validation and the test sets.

Model	Set	Accuracy
CA binding	Validation	0.550
	Test	0.619
Toxicity	Validation	0.722
	Test	0.684

The Y-randomisation robustness test (§2.3.4) illustrated that the statistical significance of the proposed models. Random shuffles of the endpoints were performed while the descriptor matrix of the calibration set remained intact. Predictions using the validation set demonstrated that the resulting models (same parameters as the proposed ones) presented statistically lower predictive power (0.40–0.55 for the CA binding and 0.33–0.53 for the toxicity model) in comparison to the models using the original training values, thus the possibility of chance correlation was eliminated. As far as internal validation is concerned, the models' stability to the inclusion–exclusion of data was tested by performing LOO and L5O cross-validation, in the training sets. The accuracy values of cross-validation for both models are presented in Table 7.4 and are higher than 0.7 thus, both models can be considered stable.

Table 7.4: Accuracy values of the *MWCNTs [b]* *k*NN/read-across predictive models for the calibration and training sets in LOO and L5O cross-validation.

	CA binding	**Toxicity**
LOO	0.810	0.750
L5O	0.833	0.722

Finally, the APD has been determined in order to define the area of reliable predictions. The APD threshold was calculated, according to the training set, to be 2.166 for the CA binding model. All samples in the test set had values in the range of 0.219–2.297. Similarly, for the toxicity model, the APD threshold was calculated equal to 1.805 and the decorators in the test set had values in the range of 0.25–2.305. Therefore, in both cases, the prediction for the samples that exceeded the APD threshold was considered unreliable. For comparison reasons we altered the Z values (Eq. 2.25) by ±0.1 and we observed the new APD values and whether the percentage of reliable predictions has changed. In both models, the percentage of the reliable predictions was the same both for the initial Z value (0.5) and for the modified Z values. We considered that a moderate Z value and thus a moderate APD value would lead to predictions that could be considered reliable with enough certainty. A rather strict or a rather flexible APD could lead to the exclusion, from the final report, of reliable predictions or to the inclusion of unreliable predictions, respectively.

A representative case of the read-across process is presented below using the sample "AMOO4AC008" which belongs to both test sets for CA binding and toxicity. In Figure 7.2, the 3 CA binding and the 7 toxicity neighbours are presented and their structural similarity in terms of common substituents is depicted using a color code. In Table 7.5 the neighbours, along with their distance from the "AMOO4AC008" sample, are presented.

Table 7.5: CA binding and toxicity training neighbours of the test *MWCNT [b]* sample "AMOO4AC008" in the training set.

AMOO4AC008	**Actual values**: non-binder/toxic		**Predicted values**: non-binder/toxic		
	CA binding			**Toxicity**	
Neighbours	Distance		Neighbours	Distance	
AMOO1AC008	0.1793	non-binder	AMOO5AC008	0.042	toxic
AMOO3AC008	0.2212	non-binder	AMOO5AC006	0.0704	toxic
AMOO7AC006	0.3317	non-binder	AMOO3AC008	0.0733	toxic
			AMOO3AC007	0.0909	non-toxic
			AMOO4AC006	0.0928	toxic
			AMOO2AC006	0.1158	toxic
			AMOO8AC006	0.1185	toxic

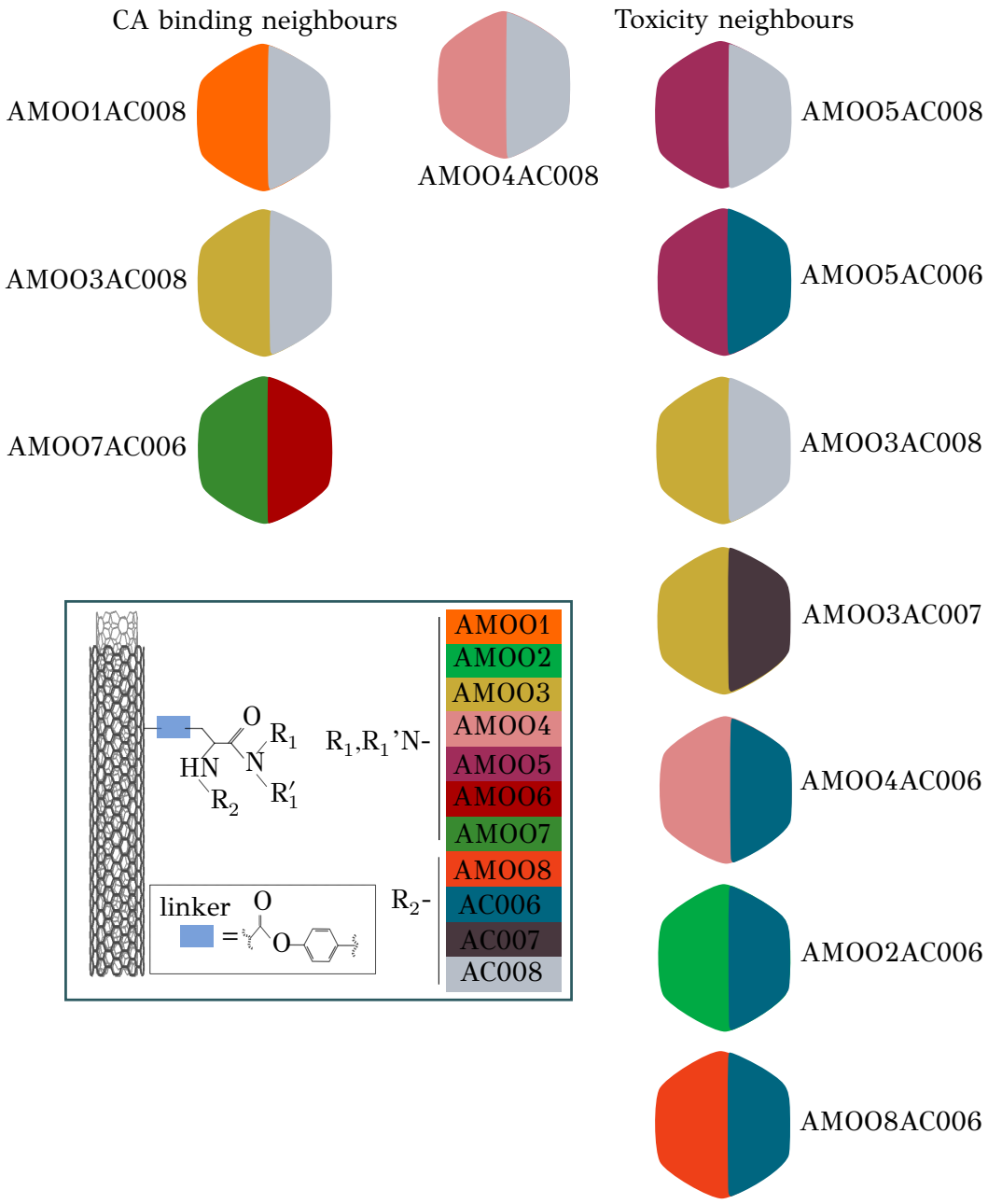


Figure 7.2: A qualitative representation of the neighbours of the decorated test *MWCNT [b]* sample AMOO4AC008. Both the CA binding and toxicity neighbours are ordered according to their distance from the query sample. The color code for the substituents R1/R1' and R2 of the MWCNTs surface decorators are presented.

7.1.2.1 [t] Discussion on selected descriptors

Most of the selected descriptors, as presented in Table 7.1, are derived from the structural graph representation of the molecules and quantify their molecular topology. [131] Geary coefficients are topochemical indices that encode spatial autocorrelation, a function of spatial separation that measures the strength of the relationship between atoms. Burdex eigenvalues, that belong to the class of Burden eigenvalue descriptors, [132] have emerged as significant variables for model development. Burden eigenvalues are topochemical indices, which reflect both the topology of the whole molecule and the chemical properties of atoms such as their chemical identity or their hybridization state. Mohar indices are topostructural indices, which encode useful information about the adjacency and distances between atoms within the molecular structure. In addition, Vertex distance counts, which express the distance de-

gree between the atoms of a molecule (e.g. the order of their neighbours), were identified. The majority of the aforementioned descriptors belong to the family of molecular topological indices, including among others, the structure of the molecules and the distances between atoms. [131] More details about the selected descriptor can be found in Appendix Γ′.1.

Here, we focus on the descriptors with the highest ranking during the variable selection process. Descriptors related to the topological charge index express the charge transfer between pairs of atoms and consequently the overall transfer of charge in the molecule. The Geary topological structure autocorrelation descriptors, embedded with a physicochemical property as a weighting factor (such as the Sanderson electronegativities or the atomic polarizabilities) also emerged as important ones for modelling during variable selection. Considering that the molecules in question are the MWCNTs decorators and the surface area of the decorator is also their "contact area" with the biological environment, the surface electrostatic status influences the MWCNT behaviour in the exposed environment. For example, it is reported in the Literature [133], [134] that electrostatic interactions directly induce the adsorption of proteins onto ENMs, thus surface charge of the MWCNTs, which is conferred by the decorating ligands, is an important factor, greatly related to the CA binding endpoint. Surface charge is also an important parameter for the cytotoxicity endpoint, given that it contributes to the cellular uptake of ENMs. [135], [136] Beyond the molecular scale of these descriptors, the electrostatic status of the ENMs is expressed by their surface charge or their zeta-potential.

7.1.2.2 [r] Web implementation

The models are available for public use and verification through the Enalos Cloud platform (`enaloscloud.novamechanics.com/EnalosWebApps/CNT/`), and can be used in order to observe the effects of the different inputs (decorating molecule structures) on the prediction of CA binding to the MWCNTs and the toxicity of the resultant decorated-MWCNTs. The user-friendly web service will facilitate the computer-aided design of novel MWCNTs by the interested users (computational experts or not); the Enalos Cloud platform can be easily accessed and can be directly explored by anyone interested in MWCNTs design to optimise functionality and safety (i.e. safe-by-design), without any need for prior programming skills. The user-friendly interface can be seen in Figure 7.3.

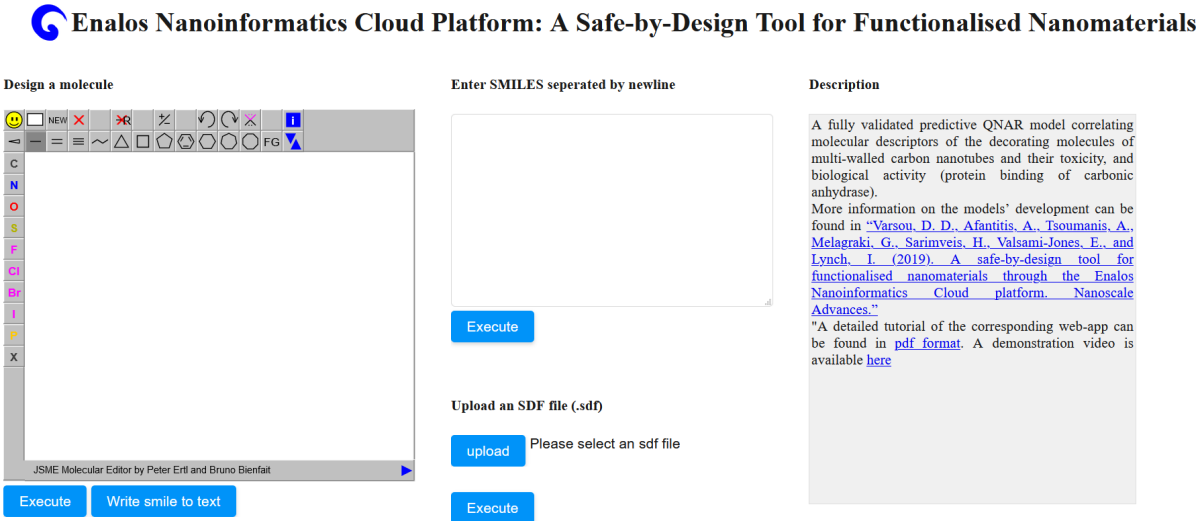


Figure 7.3: Enalos Nanoinformatics Cloud platform interface for the MWCNTs *k*NN models. At the left-handed side the molecular drawing tool is found. At the top right-handed the SMILES input form can be seen followed by the option of importing an SDF file.

The user can insert one or several structures of compounds being considered as potential decorating molecules for MWCNTs and get, within seconds, the prediction of the CA binding

and their toxicity profile, along with a warning on the reliability of the predictions. During a safe-by-design process, different datasets with decorators of interest can be imported, and their effects on the biological and toxicity behaviour of the resulting decorated MWCNTs can be studied.

Required input The user has three different options for providing the structures of the compounds to be screened:

- By drawing the chemical structure of interest. When using the drawing tool, only one structure can be submitted at a time. The user can easily select from the different panels' specific atoms, bonds or substructures and then construct the decorating molecule. The functionality also enables the user to open, save and convert files with a variety of chemical formats such as, SMILES, IUPAC Chemical Identifier etc. using the drop-down menu of the on-line sketcher. It is also possible to search via InChIKey identification, information about the molecular structure of interest. More information about the drawing tool can be found in JSME Molecule Editor website (`peter-ertl.com/jsme/2013_03/help.html`).
- By SMILES notation of the compounds in the appropriate field (Figure 7.4). Even if the SMILES notation is not initially known, the included chemical sketcher gives the users the opportunity to draw the molecular structure and then copy the structure as SMILES by right mouse click or by using the drop-down menu. This facilitates the generation of several structures, by allowing a multitude of modifications to be performed using the sketcher and then coping all structures as SMILES and pasting in the appropriate field, so that a prediction for the whole set of produced structures is obtained. Thus, the modifications can be visualized, and multiple predictions can be realized at once.

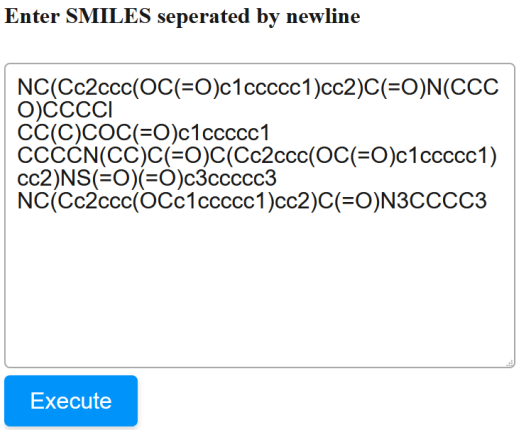
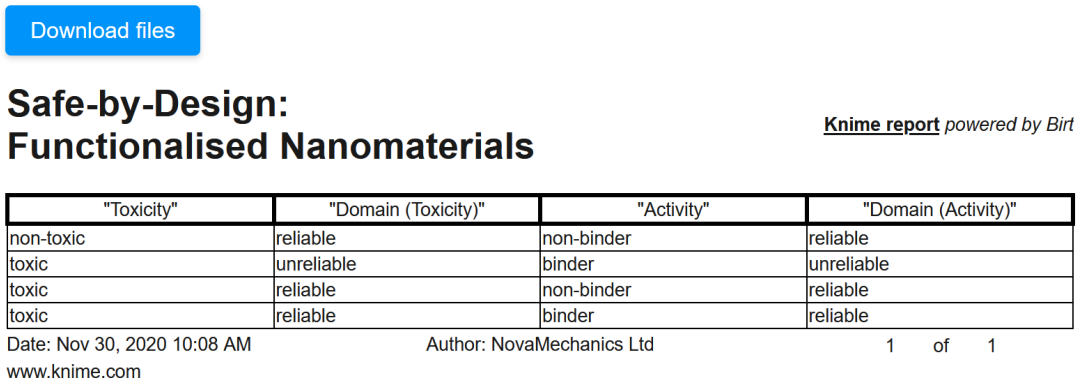


Figure 7.4: SMILES identification input in the MWCNTs web service. Each SMILE must be separated by newline.

- By uploading a Standard Database Format (SDF) file with a batch of compounds. The user can select and import an SDF file with several structures, by clicking the *Upload* button. This type of files contains molecular structure records, used as a standard exchange format for chemicals information. The decorating molecule's structure in SDF format can be extracted from PubChem Data base or other repositories.

Output When properties are uploaded for a set of decorating molecules, a prediction is generated by submitting the input structures by clicking on the *Execute* button of the corresponding field. The results include the predicted CA binding class ("binder"/"non-binder") to the MWCNTs, the toxicity class ("toxic"/"non-toxic") of the resultant decorated-MWCNTs for

each structure entered, and an indication of whether these predictions could be considered reliable based on the domain of applicability of the models (Figure 7.5). Two options are available: The "reliable" option which indicates a prediction within the domain of applicability limits of the model and the "unreliable" option which is a warning for the opposite case.



Download files

Safe-by-Design: Functionalised Nanomaterials

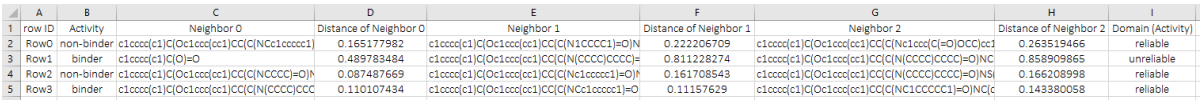
Knime report *powered by Birt*

"Toxicity"	"Domain (Toxicity)"	"Activity"	"Domain (Activity)"
non-toxic	reliable	non-binder	reliable
toxic	unreliable	binder	unreliable
toxic	reliable	non-binder	reliable
toxic	reliable	binder	reliable

Date: Nov 30, 2020 10:08 AM Author: NovaMechanics Ltd 1 of 1
www.knime.com

Figure 7.5: MWCNTs web service results table. The first and the third columns contain the prediction for each input sample and the second and fourth columns contain the reliability of the corresponding prediction according to the domain of applicability.

By clicking the *Download files* button, the above table is downloaded on both CSV and HTML format. In the produced CSV files, the interested users can observe the neighbours of the training set used for the prediction of each one of the input samples (Figure 7.6).



	A	B	C	D	E	F	G	H	I
1	row ID	Activity	Neighbor 0	Distance of Neighbor 0	Neighbor 1	Distance of Neighbor 1	Neighbor 2	Distance of Neighbor 2	Domain (Activity)
2	Row0	non-binder	c1cccc(c1)C(Oc1ccc(cc1)CC(C(NCc1ccccc1)	0.165177982	c1cccc(c1)C(Oc1ccc(cc1)CC(C(N1CCCC1)=O)N	0.222206709	c1cccc(c1)C(Oc1ccc(cc1)CC(C(Nc1ccc(C(=O)OCC)cc1	0.263519466	reliable
3	Row1	binder	c1cccc(c1)C(O)=O	0.489783484	c1cccc(c1)C(Oc1ccc(cc1)CC(C(N(CCCC)CCCC)=	0.811228274	c1cccc(c1)C(Oc1ccc(cc1)CC(C(N(CCCC)CCCC)=O)NC	0.858909865	unreliable
4	Row2	non-binder	c1cccc(c1)C(Oc1ccc(cc1)CC(C(NCCCC)=O)N	0.087487669	c1cccc(c1)C(Oc1ccc(cc1)CC(C(Nc1ccccc1)=O)N	0.161708543	c1cccc(c1)C(Oc1ccc(cc1)CC(C(N(CCCC)CCCC)=O)NS(	0.166208998	reliable
5	Row3	binder	c1cccc(c1)C(Oc1ccc(cc1)CC(C(N(CCCC)CCC	0.110107434	c1cccc(c1)C(Oc1ccc(cc1)CC(C(NCc1ccccc1)=O	0.11157629	c1cccc(c1)C(Oc1ccc(cc1)CC(C(NC1CCCCC1)=O)NC(c	0.143380058	reliable

Figure 7.6: Example of the output file of the MWCNTs web service containing the neighbours and their distance in the training set used for the activity prediction of each input sample.

7.1.2.3 [r] Virtual screening

The developed models can be used under a virtual screening framework for the development of novel, plus safe, decorated MWCNTs. As an initial case study, we tried to improve the profiles of MWCNT samples identified in the initial dataset as having unsatisfactory toxicity and high protein binding properties (a toxic and CA binder sample). It should be mentioned at this point that, depending on the nature of the specific proteins that bind, protein binding can increase an ENM's engagement with specific cellular receptors thus enhancing uptake, or can increase or reduce the susceptibility to phagocytosis (depending on whether the corona presents opsonising or disopsonising proteins) or can create cryptic epitopes in cellular signaling proteins causing toxic responses. [86] As a second case study we performed a sensitivity analysis in order to explore the toxicity and the protein binding limits of the samples, by inserting, deleting or modifying substituents at different positions of the decorators. These safe-by-design case studies are presented below.

Case study 1: Design of MWCNTs with desired properties To begin with, we selected three MWCNT samples with unsatisfactory toxicity and CA protein binding responses and through a similarity search in the PubChem database, [137] we proposed a group of potential surface modifying compounds that could lead to samples with the desired (low) toxicity and (low) protein binding levels. Therefore, we selected the AMOO4AC002, the AMOO7AC002

and the AMOO8AC002 [87] samples from the initial dataset, which are toxic and bind CA. For their substituents – as presented in Figure 3.2 – using the `Enalos+ PubChem Similarity` and the Main PubChem KNIME nodes, we searched the whole PubChem repository for similar substituents to the reference substituents of the initial samples. Tanimoto similarity measure was selected equal to 98% for both substituents R_1 and R_2.

After filtering the duplicate generated substituent SMILES, we created a list of 942 candidate surface modifiers by combining the different substituents in positions R_1 and R_2 with the core molecule. We uploaded an SDF file including these structures to the web service, and within seconds we acquired the predictions for their CA binding and toxicity profiles, as well as the reliability of these predictions according to the APD limits. According to our initial plan we were only interested in MWCNTs with reduced toxicity and low protein binding, thus from the generated predictions we focused only on non-toxic and CA non-binder results. From these, we excluded the samples with unreliable outcomes and 32 MWCNT samples with desired properties remained. In a final step we checked if the valence on the atoms of the structure is correct in KNIME, using the `Valence Checker` node. The valence was correct for the structures, therefore they can be considered feasible. Three candidate surface decorators are presented in Figure 7.7.

Case study 2: Sensitivity analysis In order to test the sensitivity of the proposed method to vary the decorator compounds, we slightly altered (Tanimoto similarity over 91%) the decorator's structure of a sample with desired properties from the initial dataset. Sample AMOO3AC005(1) [87] is a non-toxic CA non-binder that was used as the input structure for extracting similar compounds in the way described for the previous case study. After filtering the duplicate generated SMILES, 26 compounds remained, to be tested in the dedicated MWCNTs web service we have developed as described above. From the produced predictions we focused only on the 13 reliable ones, according to the calculated applicability domain. Finally, in order to be consistent with the initial structure of the MWCNTs as depicted in Figure 3.2, we excluded the compounds that did not meet its' main components; i.e., the structure of the linker and the substituent base. The selected altered decorators are presented in Figure 7.8.

7.1.3 Discussion

A fully validated workflow for prediction of the binding of a representative protein, CA, to organic molecule functionalised MWCNTs and for prediction of the toxicity of the functionalised MWCNTs has been developed and has been offered to the community as a user-friendly web service through the Enalos Cloud platform.

The predictive power of the proposed models compared to the models developed by Fourches *et al*. [87] is improved in terms of sensitivity and specificity, especially in the case of the toxicity endpoint. Singh *et al*. [57] and Shao *et al*. [129] reported high accuracy statistics nevertheless, their findings are not directly compared with the results reported here, as they considered a smaller dataset consisting of only 29 MWCNTs.

The main advantages of the models presented here compared to other relevant models proposed in the Literature, [57], [87], [129] are: the immediate release and dissemination of the models to all interested parties, the important new insights into the significant molecular descriptors and the determination of the domain of applicability of the model allowing for the discrimination between reliable and unreliable predictions. The web service is publicly available and ready-to-use by any interested user (e.g. experimentalists or regulators) in the computer-aided design of novel MWCNTs or in the prioritization of novel potent MWCNTs based on their predicted toxic effects, taking into account that predictions can be produced rapidly (about 30 seconds) along with an indication of their reliability. Thus, it represents a

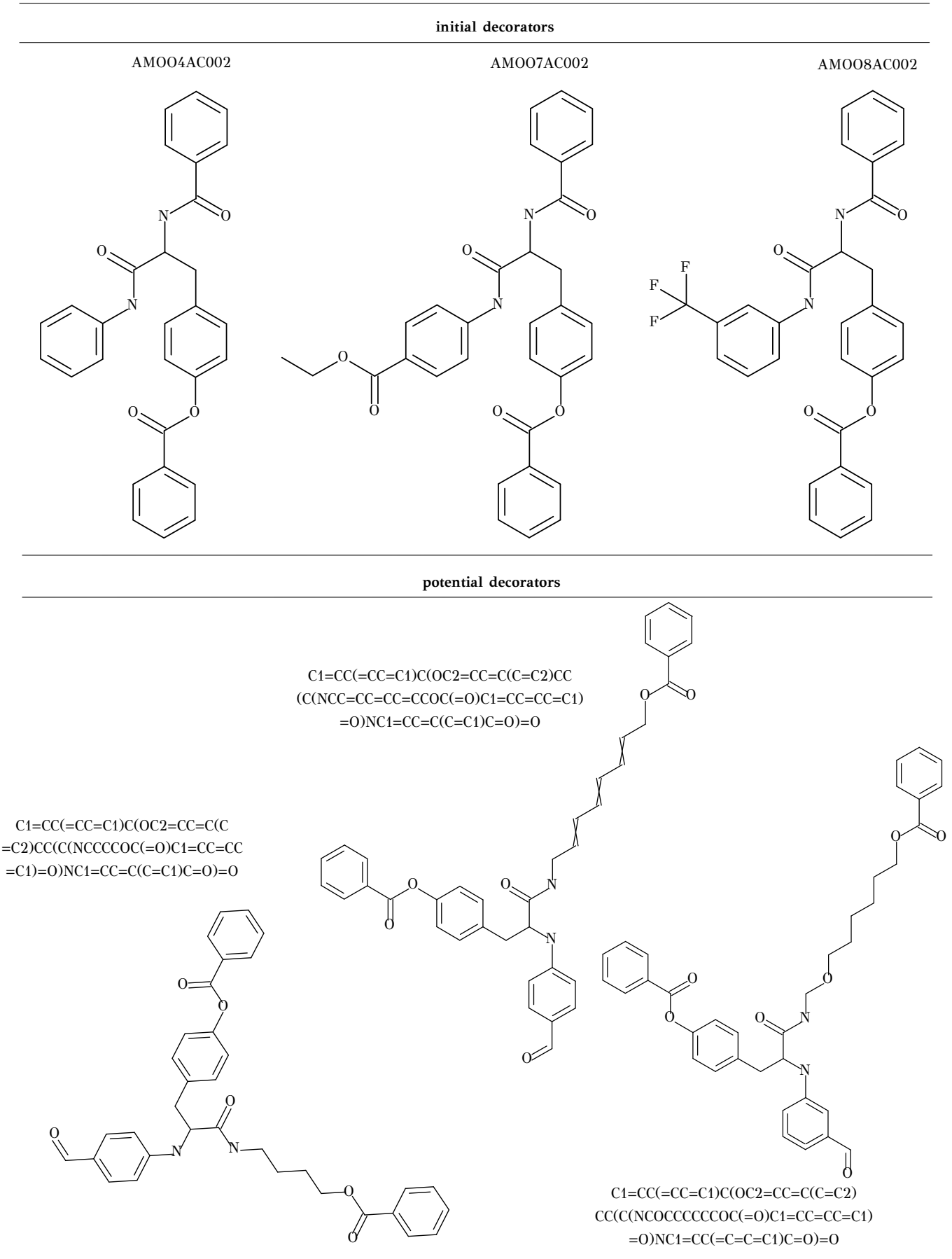
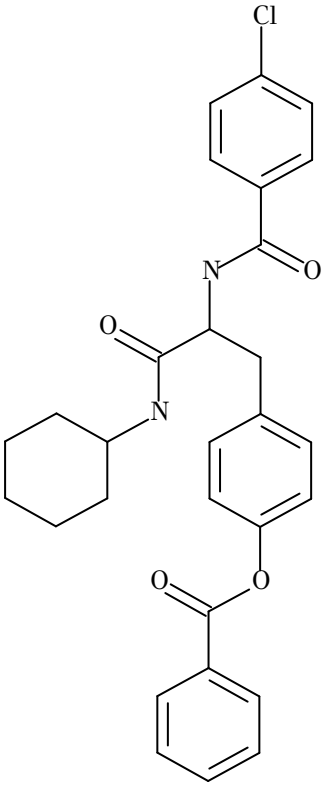


Figure 7.7: Potential decorators for designing MWCNTs with desired properties (non-toxic, non-protein binders) based on the decorators of three inadequate (i.e. toxic and CA binders) samples of the initial dataset.

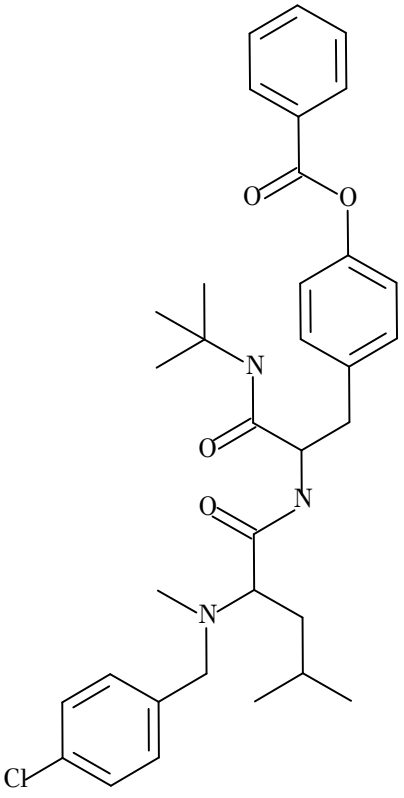
initial decorator

AMOO3AC005(1) non-toxic/non-binder

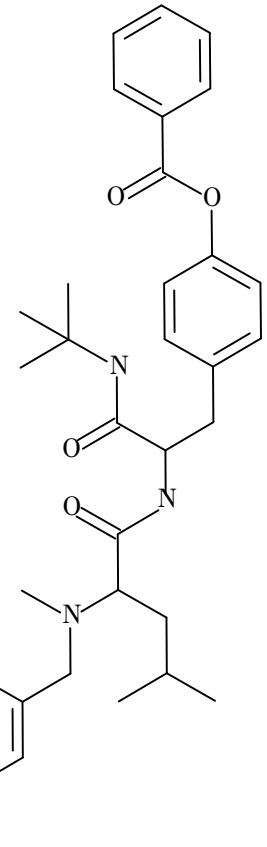


altered decorators

toxic/non-binder



non-toxic/non-binder



non-toxic/binder

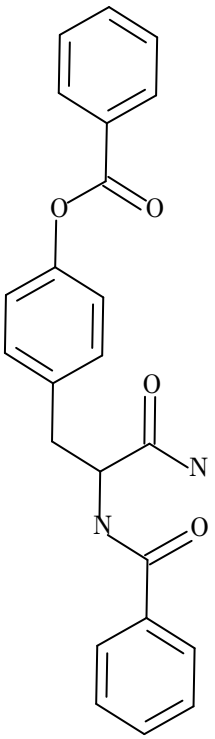


Figure 7.8: Altered *MWCNTs [b]* decorators according to an initial decorator with desired properties in sensitivity analysis.

useful tool within a safety-by-design framework and can contribute to the reduction of *in vivo* experiments and their replacement by *in vitro* and in due course only *in silico* experiments. Finally, the dissemination of the models facilitates their utility as they are easily expandable and adjustable to address the requirements of other ENMs, other decorating molecules or other toxicity endpoints, provided sufficient experimental data is available to train the extended models.

While it was not possible based on the current dataset to link the binding and toxicity read-across models, since the uptake studies were performed in serum-containing medium rather than on the single protein-bound MWCNTs (i.e. CA-MWCNT complexes) it is clear that as suitable datasets become available where protein binding and toxicity are performed under the same conditions, a linked model, that can determine whether high protein binding correlates with high or low toxicity, would be possible. Indeed, reduction of protein binding via surface decoration of ENMs with polyethylene glycol (PEG) or other hydrophilic polymers has been suggested as a route to reducing their recognition and phagocytosis as a "stealth" strategy for nanomedicines. [138], [139] Conversely, corona thickness as driven by use of different media supplemented with 10% fetal bovine serum was shown to affect cellular uptake and toxicity for gold ENMs: while DMEM elicited the formation of a large time-dependent protein corona, Roswell Park Memorial Institute medium (RPMI) showed different dynamics with reduced protein coating which correlated with more abundant internalized by two cell lines (HeLa and U937) cells and higher cytotoxic effects as compared to DMEM. [140]

7.2 Development of a read-across model for zeta-potential prediction

Beyond the characterisation of the pristine ENMs, it is vital to understand their behaviour under the relevant exposure conditions, e.g. dispersed in the relevant OECD or other test medium. For example, zeta-potential provides an estimation of the surface charge, and therefore the electrostatic stability, in medium and, in practical terms, controls the ENM's tendency to form agglomerates.[1] [93], [99] Thus, models extracting descriptors from TEM images, usually prepared in simple medium, need to have input information on the surface charge and the type of capping (e.g. small molecules such as citrate) or coating (typically larger polymers such as PEG, PVP etc.) molecules present on the ENMs.

As previously mentioned, experimental approaches are often costly and time consuming therefore, computational approaches can provide significant aid to prioritization of ENMs for experimental testing, and indeed to prioritize which characterisation endpoints, from the quite extensive lists of minimal characterisation needs, [143] are most useful for correlating with toxicity and other biological or environmental effects. One promising approach is to develop computational tools that extract additional information from existing experimental datasets – i.e., to enrich the experimental datasets with computationally determined descriptors, thus maximizing the utility of the experimental datasets. For example, SEM (scanning electron microscopy) or TEM images are currently utilized by experimentalists to determine size and size distribution, and occasionally for characterisation of shape, aspect ratio or other morphological parameters, although there are no agreed methodologies or reporting conventions for these, and it is very challenging to extract ENMs morphological data directly from such images. Tools and workflows for extracting image descriptors from high-throughput fluorescence images of cells, such as global intensity level, cell count, cell shape, cellular and subcellular constellations, colocalization information etc., already exist in the field of Biology

[1] Zeta- potential values of $\pm$ 30 mV are often considered to denote good electrostatic stabilization. [141] However many ENMs may also have a strong contribution from steric stabilization so, close to neutral zeta-potentials cannot always be used to predict instability and agglomeration potential. [142]

[144] and can also be used for the analysis of ENMs microscopy images with only a few modifications. Currently, many efforts for the extraction of image analysis descriptors have been presented, including the NanoXtract, [3] the open-source ImageJ [145] tool and similar tools like Fiji, [146] or implemented in programming environments like MATLAB, [146], [147] to mathematically describe and "quantify" the different shapes of the ENMs and utilize these image-descriptors as additional input information for predictive models of ENMs toxicity. [146], [148]

In this section of the Thesis, we present a fully validated read-across model trained on the *NanoMILE ENMs* dataset (§3.6) that predicts zeta-potential of ENMs, based on calculated descriptors from the available TEM images. A key element in our study was the discussion on the selected descriptors, to highlight their influence on the zeta-potential. Zeta-potential was selected as the endpoint of interest because it is easy to measure, but also because work has shown that this property correlates highly with ENM behaviour and cytotoxicity.

7.2.1 [m] Development of the predictive workflow

The KNIME platform was used (see §A′.1.4) to perform all the various components of the nanoinformatics analysis under a common interface, including variable selection, model development and validation. In this way we ensured a systematic approach for the analysis and a complete supervision of the workflow. In the developed workflow the available nodes were combined with the `Enalos+` nodes, developed by NovaMechanics Ltd.

The calculated image nanodescriptors had substantially different numerical ranges, and therefore, in order to force them to contribute equally to the rest of the analysis, we performed Gaussian (z-score) normalisation. [6] The zeta-potential endpoint values were also normalised.

7.2.1.1 [m] Variable selection

The Best First (`BestFirst`) variable selection was performed using the CFS Subset Evaluator (`CfsSubsetEval`) included in `WEKA` node, to select the attributes that are the most relevant to the specific endpoint (zeta-potential in our case). `BestFirst` search method searches the space of attribute subsets by greedy hill-climbing with backtracking facility. `CfsSubsetEval` is a correlation-based attribute subset evaluator, that takes into account subsets of uncorrelated features but that are highly correlated with the predicted endpoint. [52] By applying variable selection, noisy attributes were excluded, the algorithm's performance was greatly improved and over-fitting of the model to the dataset was avoided.

7.2.1.2 [m] *k*NN/Read-Across Model Development

The *k*NN method incorporated into KNIME through the `EnaloskNN` node, was used for developing the read-across model. [69] The endpoint was numerical, hence the prediction was the distance weighted average of the endpoint of the selected neighbours. An optimal *k* value was selected based on the calculated Euclidean distance between all instances and used as weighting factors the inversed distance. [52], [56] Another important aspect of the analysis -apart from the simple endpoint prediction- was to clearly define and present the groups of *k* neighbours of each test ENM, and therefore to specify and map the analogous space, which is a prerequisite of the read-across framework. [38]

Model validation In order to fully validate the proposed model, external validation was performed by separating the full dataset into training and test sets, with the test set left out of modelling and used subsequently for validation purposes. Nevertheless, as the full dataset was small in size, we performed multiple divisions in order to eliminate the possible bias of the splitting on the predictive accuracy. Various random –but stratified- splits were performed,

keeping the previous proportion between training and test sets, and the results of modelling based on the different training sets were compared to each other, until it was ensured that the model is sufficiently robust.

To evaluate the models' performance, the goodness-of-fit on the test data was measured, using the coefficient of multiple determination (R^2_{pred} - Eq. 2.12) [54] and the statistical indices, as proposed by Tropsha *et al*., to assess the predictive power of regression predictive models (§2.3.3.1). [63], [73] A further set of conditions were tested also, in accordance with the approach of Tropsha *et al*. [73].

Finally, as an additional test of the robustness of the proposed model, the Y-randomisation test (§2.3.4) was performed.

Applicability domain Considering that in this study a local (read-across) methodology was applied, the APD could not be defined using all the samples of the training set; therefore, for each query ENM, the APD was defined using similarity measurements based on the Euclidean distance among the *k* selected neighbours of the training set and the test ENM. The distance of a test ENM to its nearest neighbour in the training set is compared to the predefined APD threshold (Eq. 2.25) and its prediction is considered unreliable if the distance exceeds this APD limit. The assessment of the APD of the proposed model was introduced into a KNIME workflow, using the `Enalos+ Domain - APD` KNIME node. [69], [90]

7.2.2 [r] Results and discussion

In order to develop our predictive workflow, the complete dataset (18 parameters extracted from each of the 68 TEM images and 2 additional data points (core and pH) for each of the 37 ENMs, in total 740 data points) was partitioned randomly for external validation purposes into training and test sets in the proportion 75:25. The training set was used in the model development and the test set was not involved in this process but was kept as a blank set for subsequent model validation. The `BestFirst` variable selection along with the `CfsSubsetEval` evaluator were applied to the training set, in order to select the most significant among the 20 nanodescriptors.

The *k*NN methodology using `Enalos+kNN` KNIME node, with the optimal value of $k = 7$ neighbours, produced both satisfactory and reliable predictions, as it was successful in all Tropsha's [73] recommended tests (Eqs. 2.17 to 2.20) to assess the predictive ability of developed models:

$$R^2_{\text{pred}} = 0.898 > 0.6 \tag{7.1}$$

$$Q^2_{\text{ext}} = 0.907 > 0.5 \tag{7.2}$$

$$\frac{R^2_{\text{pred}} - R^2_0}{R^2_{\text{pred}}} = 0.003 < 0.1 \tag{7.3}$$

$$0.85 < k = 1.056 < 1.15 \tag{7.4}$$

In Figure 7.9 the predicted zeta-potential values for the test ENMs are presented along with the corresponding experimental values for the test set.

As described above, for validation purposes, various random, stratified (regarding the type of core) partitions with the same proportion (75:25) were performed, to assess the predictive power of the approach independently of the data partitioning. In every case variable selection (as described above) was performed in order to clearly define the image nanodescriptors space among the initial set of 18 descriptors. All models were successful in Tropsha's [73]

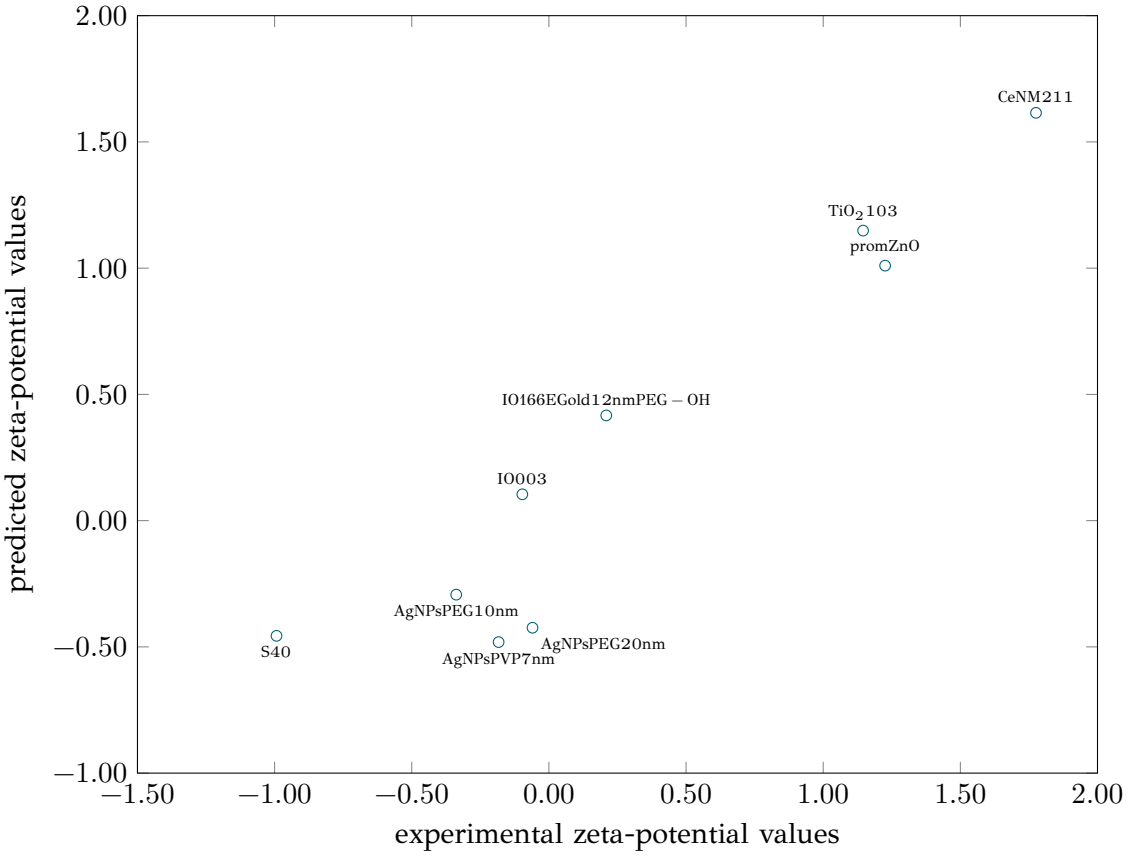


Figure 7.9: Predicted zeta-potential values (normalised) using the proposed *k*NN/read-across model on the test set.

recommended tests and the results for the squared correlation for all different splits, are presented in Table 7.6.

Table 7.6: R^2_{pred} values of the test set for different *NanoMILE ENMs* dataset splits, using the zeta-potential *k*NN/read-across model.

Splitting	$R^2_{\textbf{pred}}$
Initial	0.898
Random 1	0.613
Random 2	0.689
Random 3	0.829
Random 4	0.785
Random 5	0.702

Finally, the Y-randomisation test was performed, which confirmed that the proposed predictive model is robust. Five random shuffles of the endpoint variable vector (zeta-potential) were performed, whereas the descriptor matrix remained intact. The correlation coefficient (R^2_{pred}) was tracked, as can be seen in Table 7.7, and the models presented have statistically lower predictive power than the initial one, therefore it can be considered that the accuracy of the proposed model is true and is not due to chance correlation.

For the calculation of the APD cut-off limit, the type of core and the pH were excluded as categorical variables. However, given that in both training and test sets these variables had the same possible values, the calculation of the APD was not affected by this exclusion. The APD threshold was calculated for each test ENM according to its seven selected neighbours

Table 7.7: Y-randomisation results (correlation coefficient and number of satisfied tests) of the zeta-potential *k*NN/read-across model.

	$R^2_{\mathbf{pred}}$	**Satisfied criteria**
Initial	0.898	4
Y-random 1	0.015	0
Y-random 2	0.098	0
Y-random 3	0.317	0
Y-random 4	0.416	1
Y-random 5	0.386	1

in the training set. As can be seen in Table 7.8, the test ENMs with domain values higher than their corresponding APD threshold, are considered unreliable.

Table 7.8: Domain of applicability and reliability of predictions for each test ENM of the *NanoMILE ENMs* set using the zeta-potential *k*NN/read-across model.

NM in test set	**Domain**	**APD threshold**	**Reliability**
IO166E Gold 12 nm PEG-OH	0.026	0.060	reliable
AgNPs PEG 10nm	0.008	0.031	reliable
AgNPs PEG 20nm	1.230	0.036	unreliable
AgNPs PVP 7nm	0.021	0.036	reliable
TiO2 103	0.075	1.163	reliable
prom ZnO	0.331	1.163	reliable
Ce NM 211	0.011	1.222	reliable
S40	0.033	0.036	reliable
IO003	0.123	0.047	unreliable

7.2.2.1 [t] Discussion on selected descriptors

As a next step, interpretation of the variable selection results and clear definition of the variables that emerged as important for modelling the zeta-potential endpoint are provided. In all partitions the type of core of the ENMs and their main elongation emerged as important variables. Here, the information encoded by these descriptors is analysed in order to understand how they affect the zeta-potential value for a specific ENM. The predominant variables of the initial model are presented below in order of significance.

Main elongation The elongation (el, Eq. 7.5) is calculated using the parameters of the minimum bounding box (rectangle); the larger side (L) and the shortest side (S) of the minimum bounding box. The smallest enclosing box (or minimum bounding rectangle – Figure 7.10) is the smallest rectangle that contains every point of the particle. The main elongation variable expresses the lengthening of the particle and is similar to aspect ratio descriptor (L/S). [149]

$$el = 1 - \frac{S}{L} \tag{7.5}$$

The fact that this parameter has a high appearance rate for zeta-potential is interesting for 2 reasons: Firstly, there are significant questions over the reliability of experimental measurements for zeta-potential as ENMs become less spherical and/or agglomerated, since the underpinning mathematical models for zeta-potential are based on Stoke's law and assume

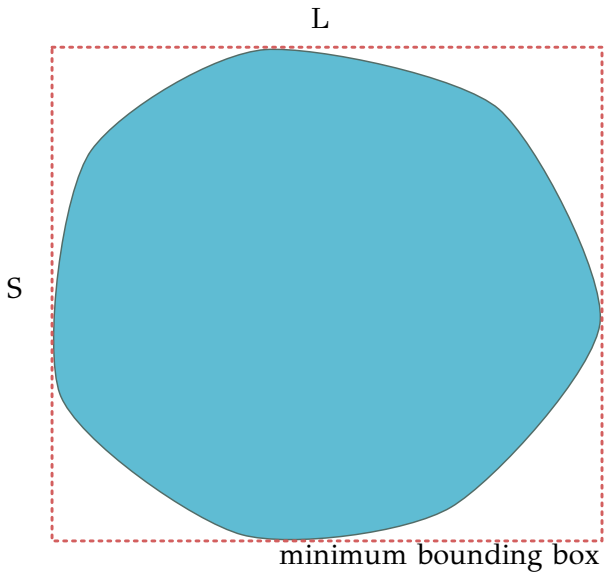


Figure 7.10: Minimum bounding box (or rectangle) of a particle. This image can be used under the following terms: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) https://creativecommons.org/licenses/by-nc-sa/4.0/.

spherical particles. [98] Secondly, there is a well-established paradigm for increasing toxicity of ENMs with increasing aspect ratio, resulting in the definition of a sub-group of ENMs described as HARNs (High Aspect Ratio ENMs), which are perceived as more hazardous, especially when combined with rigidity and persistence. [150]

Type of core The inner material of an ENM, its core, is one of the principal factors that define their behaviour during production and processing and their interaction with the environment and humans. The core of the ENMs is also responsible for their main intrinsic physicochemical properties, such as electrical, magnetic and catalytic properties, selectivity, solubility etc. [151] For example, some ENMs (e.g. Ag, Cu/CuO, ZnO) may dissolve quickly in a medium, while others –like cerium dioxide or titanium dioxide- dissolve at a slower rate. [18] Finally the core of the ENMs consists of the matrix for the application of the coating that –as mentioned before- can alter their properties. [151] One of the major paradigms suggested as being predictive of metal/MeOx ENMs toxicity is band-gap, [83], [152] i.e. the energy gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) which is linked to core composition and crystal structure, and thus the appearance of core composition as one of the main nanodescriptors is not surprising.

7.2.2.2 [t] The neighbouring space

Another important aspect that has been considered was to "unbox" the *k*NN algorithm and study the selected training neighbours for each test ENM. In that way it was possible to search for patterns and similarities in the "neighbourhood" space and to do a preliminary grouping of the ENMs as can be seen in the qualitative illustration of the neighbouring relationships in Figure 7.11. In Table 7.9 the selected training neighbours for each of the ENMs in the initial test set are presented.

During the study of the neighbourhood (Figure 7.11 and Table 7.9), it was expected that some patterns in the selection of neighbours would be recognized, due to obvious characteristics such as having the same core, coating, production method etc. of the ENMs. At a first glance this is clearly the case for the samples "AgNPs PEG 10nm", "AgNPs PEG 20nm", "AgNPs PVP 7nm" and "S40", that have neighbours with the same core (silver) and coating,

and the case of samples "IO003" and "IO166E Gold 12nm PEG-OH" which have neighbours with the same core (gold). For the rest of the samples ("Ce NM 211", "Prom ZnO" and "TiO2 103") there are not enough (at least seven) ENMs with the same core in the training set to be selected as neighbours, therefore the furthest neighbours are training samples with different core compositions.

It could be assumed in this case study that the type of core could be (along with the medium pH variable) the only variable that controls the whole grouping and prediction of zeta-potential process. However, the role of the main elongation variable must not be underestimated; its values have a balancing function in the selection of similar ENMs (for example in the case where not enough training ENMs with the same core are available). In the methodology of the neighbours' selection, the categorical values have either a great (equal to 1, when the compared attributes have different values) or a null participation (equal to 0, when the compared attributes have the same values) in the distance calculation. The rest of the nanodescriptors which are numeric values, have a tuning role in the calculation of the distance, as they encode the subtle similarities or dissimilarities between the ENMs, and they finally contribute to a better selection of neighbours. What is more their participation in the calculation of distance (and thereby in the calculation of weighting factors) leads to more accurate predictions.

It can therefore be concluded that the selected variables are indeed the appropriate ones, among the initial set of nanodescriptors, for modelling the zeta-potential index. To demonstrate this, the modelling process was repeated using as input descriptors only the core of the ENMs and the medium pH. This model presented lower predictive power than our initial one ($R^2_{\text{pred}} = 0.85$ and passed only three out of the four validation criteria [64], [73]; thus its predictive ability cannot be considered as high as the predictive ability of the initial model utilizing the core, the pH of the medium and the ENM main elongation extracted from the ENMs TEM images. It is worth noting that the used dataset was rather small and heterogeneous (some training ENMs are not selected as neighbours); as the dataset becomes more complete in size it is expected to achieve a better neighbour tuning and better prediction results.

7.2.2.3 [r] Web implementation

Based on the selected set of significant descriptors (type of core and main elongation) and the pH of interest the developed predictive model was released through the Enalos Cloud platform (Figure 7.12). The web service is freely available via `enaloscloud.novamechanics.com/EnalosWebApps/ZetaPotential/`. Users can import different datasets with ENMs of interest and study the effects of different inputs on the zeta-potential value, a decisive step during a safety-by-design process.

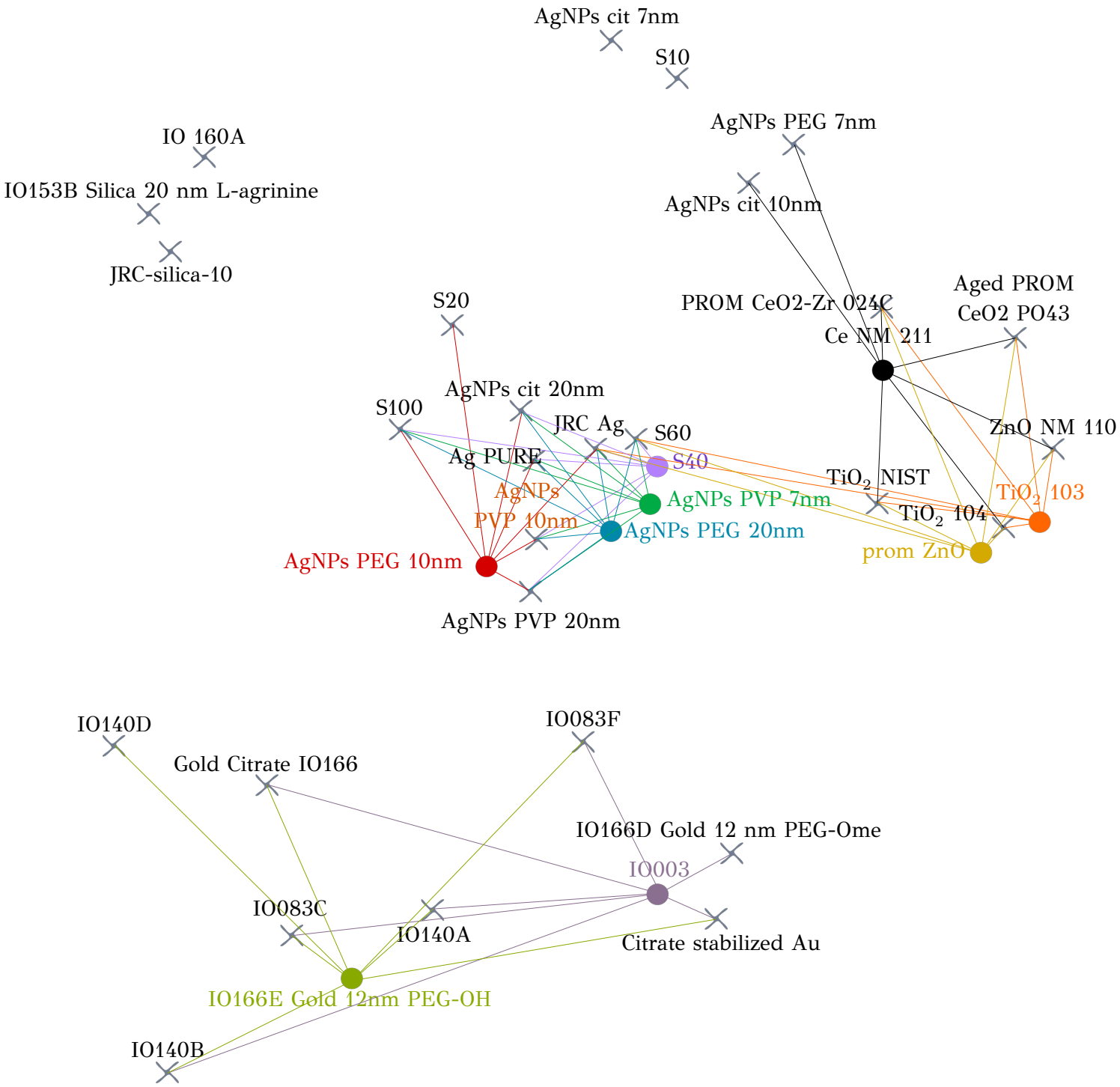


Figure 7.11: A qualitative representation of the neighbouring space of the training and the test ENM sets in the *k*NN/read-across zeta-potential model. Test ENMs are depicted with coloured circles, whereas training ENMs are illustrated with gray crosses. The seven selected-closest neighbours for each test NM are clearly defined via lines.

Table 7.9: Training neighbours for the test ENMs of zeta-potential *k*NN/read-across model, presented in order of accretive distance.

	Neighbours in training set			
ENM in test set	**# 1**	**# 2**	**# 3**	**# 4**
IO166E Gold 12 nm PEG-OH	IO083C	IO140 A	Gold Citrate_IO166	IO140 B
AgNPs PEG 10nm	AgNPs PVP 20nm	AgNPs PVP 10nm	AgPURE	AgNPs cit 20nm
AgNPs PEG 20nm	S60	JRC Ag (JRC Silver)	AgNPs PVP 10nm	AgNPs PVP 20nm
AgNPs PVP 7nm	S60	JRC Ag (JRC Silver)	AgNPs PVP 10nm	AgNPs PVP 20nm
TiO2 103	TiO2 104	ZnO 110	TiO2 NIST	Aged PROM_CeO2_PO43
prom ZnO	TiO2 104	ZnO 110	TiO2 NIST	Aged PROM_CeO2_PO43
Ce NM 211	PROM CeO2-Zr 024C	Aged PROM_CeO2_PO43	TiO2 NIST	ZnO 110
S40	S60	JRC Ag (JRC Silver)	AgNPs PVP 10nm	AgNPs PVP 20nm
IO003	IO166D Gold 12 nm PEG-Ome	Citrate stabilized Au	IO083F	IO140 A

continued from Table 7.9

	Neighbours in training set		
ENM in test set	**# 5**	**# 6**	**# 7**
IO166E Gold 12 nm PEG-OH	IO083F	IO166E Gold 12 nm PEG-COOH	Citrate stabilized Au
AgNPs PEG 10nm	S100	JRC Ag (JRC Silver)	S20
AgNPs PEG 20nm	AgPURE	AgNPs cit 20nm	S100
AgNPs PVP 7nm	AgPURE	AgNPs cit 20nm	S100
TiO2 103	PROM CeO2-Zr 024C	S60	JRC Ag (JRC Silver)
prom ZnO	PROM CeO2-Zr 024C	S60	JRC Ag (JRC Silver)
Ce NM 211	TiO2 104	AgNPs PEG 7nm	AgNPs cit 10nm
S40	AgPURE	AgNPs cit 20nm	S100
IO003	IO083C	Gold Citrate_IO166	IO140 B

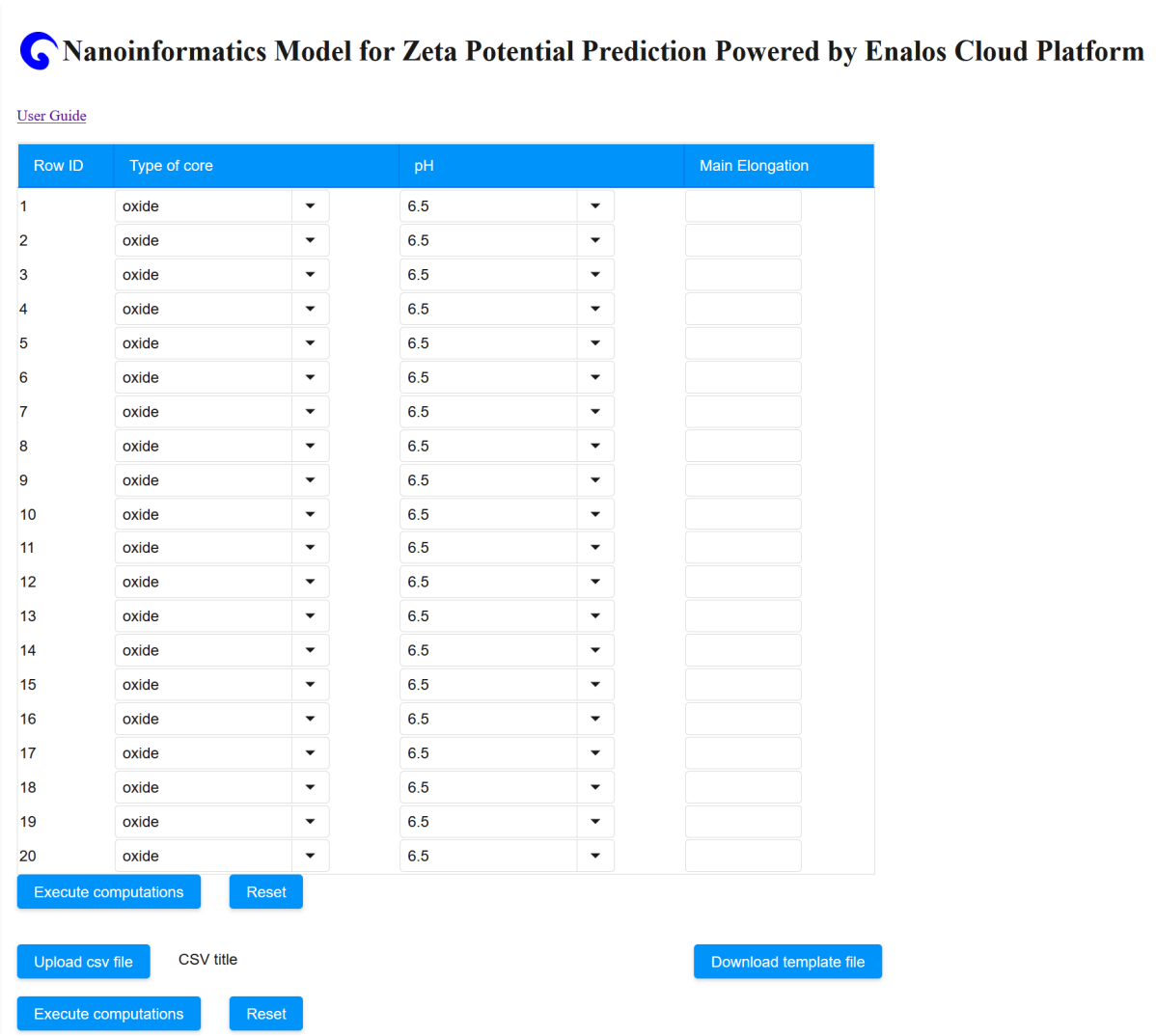


Figure 7.12: Enalos Zeta-Potential Prediction platform. At the top page the input form can be seen and at the bottom by clicking on *Upload csv file button*, the user can import a CSV file with all the required properties

Required input To start the prediction process, the user must indicate the type of core of the NPs and the pH of the solution where the experimental measurement of the zeta-potential would have been made. In addition, the user must provide the value of the main elongation of the NPs. Main elongation is an image descriptor that can be calculated from TEM images using the Enalos NanoXtract tool or an image analysis tool of their choice (e.g. ImageJ).

The user has two alternative ways to provide the above information:

- Users can enter manually the three required parameters using the form given in the website (advisable for small ENM sets -up to 20 ENMs). Each row corresponds to one ENM. The users select from the first drop-down menu the type of core of the ENM (pure metal/MeOx) and from the second the pH value (6.5/7) of the medium of interest. Finally, the users enter a numerical value of main elongation between 0 and 1.

- Users can import a file in CSV format, containing the ENM samples and their properties, by clicking the *Upload csv file* button. The file must have a specific format (Figure 7.13) as described below. A template file is also available to download.

	A	B	C	D
1	NP sample	Type of core	pH	Main Elongation
2	IO166E Gold 12 nm PEG-OH	pure metal	6.5	0.044918882
3	AgNPs PEG 10nm	pure metal	7	0.031727412
4	AgNPs PEG 20nm	pure metal	7	0.180722892
5	AgNPs PVP 7nm	pure metal	7	0.042328042
6	TiO2 103	oxide	7	0.580700546
7	prom ZnO	oxide	7	0.609965636
8	Ce NM 211	oxide	7	0.019771807
9	S40	pure metal	7	0.043717949
10	IO003	pure metal	6.5	0.017090053

Figure 7.13: Required format of the CSV file with a sample of input data the zeta-potential web service.

Output When properties are uploaded for a set of ENMs, a prediction is generated by submitting the input values (clicking on *Execute computation* button of the corresponding field), the predictive model is then applied to the data provided and the output is generated within seconds. As can be seen in Figure 7.14, the results include the predicted zeta-potential values for each included ENM as well as a warning on the prediction reliability according to the domain of applicability limits.

Download files

Row ID	Predicted zeta potential [mV]	Reliability
"IO166E Gold 12 nm PEG-OH"	-0.5815616505801628	"reliable"
"AgNPs PEG 10nm"	-23.554501736363807	"reliable"
"AgNPs PEG 20nm"	-27.79291401386539	"unreliable"
"AgNPs PVP 7nm"	-29.631770892835036	"reliable"
"TiO2 103"	23.107595387224467	"reliable"
"prom ZnO"	18.615889644766884	"reliable"
"Ce NM 211"	38.20268062475627	"reliable"
"S40"	-28.832095935920286	"reliable"
"IO003"	-10.691819408446678	"unreliable"

Figure 7.14: Generated output page of the zeta-potential web service. The first column of the results table contains the prediction for each input ENM and the second column contains the reliability of each prediction based on the model's domain of applicability. This table can also be downloaded in CSV format by clicking in the corresponding button.

By clicking the *Download files* button, the above table is downloaded in CSV format. In the downloaded files the neighbours of each particle in the training set can be also found (Figure 7.15).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	row ID	Prediction [mV]	Neighbor 0	Distance	Neighbor 1	Distance	Neighbor 2	Distance	Neighbor 3	Distance	Neighbor 4	Distance	Neighbor 5	Distance	Neighbor 6	Distance
2	IO166E Gold 12 r	-0.581561651	IO083C	0.0052	IO140 A	0.0055	Gold Citrate_IO1(	0.0058	IO140 B	0.0066	IO083F	0.0099	IO166E Gold 12 r	0.0195	Citrate stabilized	0.0205
3	AgNPs PEG 10nm	-23.55450174	AgNPs PVP 20nm	0.0017	AgNPs PVP 10nm	0.0025	AgPURE	0.0062	AgNPs cit 20nm	0.0067	S100	0.0085	JRC Ag (JRC Silver	0.0126	S20	0.0126
4	AgNPs PEG 20nm	-27.79291401	S60	0.2482	JRC Ag (JRC Silver	0.2502	AgNPs PVP 10nm	0.2603	AgNPs PVP 20nm	0.2645	AgPURE	0.269	AgNPs cit 20nm	0.2695	S100	0.2713
5	AgNPs PVP 7nm	-29.63177089	S60	0.0042	JRC Ag (JRC Silver	0.0061	AgNPs PVP 10nm	0.0162	AgNPs PVP 20nm	0.0204	AgPURE	0.0249	AgNPs cit 20nm	0.0254	S100	0.0272
6	TiO2 103	23.10759539	TiO2 104	0.0152	ZnO 110	0.5515	TiO2 NIST	0.6209	Aged PROM_CeO	0.8554	17. PROM CeO2-2	0.9776	S60	1.3819	JRC Ag (JRC Silver	1.3832
7	prom ZnO	18.61588964	TiO2 104	0.0668	ZnO 110	0.6031	TiO2 NIST	0.6726	Aged PROM_CeO	0.907	17. PROM CeO2-2	1.0292	S60	1.418	JRC Ag (JRC Silver	1.4194
8	Ce NM 211	38.20268062	17. PROM CeO2-2	0.0117	Aged PROM_CeO	0.134	TiO2 NIST	0.3684	ZnO 110	0.4378	TiO2 104	0.9742	AgNPs PEG 7nm	1	AgNPs cit 10nm	1
9	S40	-28.83209594	S60	0.0066	JRC Ag (JRC Silver	0.0086	AgNPs PVP 10nm	0.0186	AgNPs PVP 20nm	0.0228	AgPURE	0.0273	AgNPs cit 20nm	0.0278	S100	0.0296
10	IO003	-10.69181941	IO166D Gold 12 r	0.0249	Citrate stabilized	0.0286	IO083F	0.0392	IO140 A	0.0436	IO083C	0.0439	Gold Citrate_IO1(	0.0549	IO140 B	0.0557

Figure 7.15: Example of the output file of the zeta-potential web service containing the neighbours and their distance in the training set used for the zeta-potential prediction of each input ENM.

7.2.3 Discussion

In silico assessment of various ENM properties and biological effects prior to their use or even prior to actual synthesis is significantly contributing to a reduction of the cost and time

required for experimental procedures required to generate, for example, regulatory dossiers. ENM specific descriptors are thus highly desired to develop significant correlations between ENM descriptors and ENM properties and biological effects. Image descriptors derived from ENM TEM images are emerging as an important source of additional information providing an enriched parameter space to explore in order to find correlations between ENMs properties and their effects, although to date these have not been extensively studied in part because of the lack of tools available to extract additional nanodescriptors from TEM images.

Based on NanoXtract tool for image descriptors calculation a workflow to demonstrate the utility of the extracted image nanodescriptors for prediction of ENMs physico-chemical properties was implemented, utilizing zeta-potential as a first example as experimental values were available for the 37 ENMs included in the dataset. The predictive model for ENM zeta-potential was based on grouping of the ENMs according to their nearest neighbours and provided some interesting insights into the most important similarity features between the ENMs and their nearest neighbours. Thus, in addition to grouping based on the ENMs core composition (e.g. CeO_2, Ag, TiO_2 etc.), the main elongation emerged as an important grouping parameter, which links to key drivers of ENM toxicity, such as aspect ratio. Importantly, the ability to predict zeta-potentials values for non-spherical ENMs fills a gap where experimental measurement reliability and meaningfulness is poor and thus the image-based nanodescriptors and predicted zeta-potentials can be used to improve subsequent predictions of ENMs toxicity and adverse outcomes.

To ensure its accessibility to the wider community, the zeta-potential predictive model has been made publicly available as web through the Enalos Cloud Platform, enabling future *in silico* exploitation of ENM properties and behaviour based on image descriptors.

7.3 Chapter summary

In this Chapter the *k*NN machine learning methodology is included in read-across workflows for the prediction of ENMs properties. It is applied to two different case studies and proved that *k*NN is one of the most reliable, competitive and easy-to-employ read-across strategies. In brief, in each workflow the following elements are included:

- Variable selection.
- Complete validation of the produced models.
- Y-randomisation.
- Applicability domain.
- Discussion on the selected descriptors in each case.
- Dissemination of the models at: `http://enaloscloud.novamechanics.com/EnalosWebApps/CNT/` and `http://enaloscloud.novamechanics.com/EnalosWebApps/ZetaPotential/`.

The *k*NN machine learning algorithm, proved to produce reliable predictions in small or larger datasets based on similarities between ENMs. The ease to apply the *k*NN algorithm to any dataset through user-friendly platforms (e.g. KNIME, `WEKA`) allows its use from any interested user, even the not computationally high-skilled ones. Therefore, it is possible to perform fast screening on ENMs data and acquire preliminary estimations of their possible toxic effects.

Chapter 8

Conclusions

The aim of the Dissertation was the development of novel computational read-across methodologies for the prediction of properties and biological endpoints of chemical substances, with particular emphasis on addressing the modelling opportunities and challenges that have emerged in the nanoinformatics area, due to the fact that (engineered) nanomaterials (ENMs) datasets are small, with high dimensionality and high variance in the feature space. The concept behind the development of the proposed methodologies, was to improve, automate and optimise existing read-across approaches, which are time consuming and provide results which are not guaranteed to be optimal. The read-across methods are based on the development of advanced mathematical programming formulations and solution algorithms, which are able to integrate multiple objectives (maximisation of prediction accuracy, optimal grouping of ENMs, selection of the most informative descriptors). More specifically, the following goals were taken into account in the development of the read-across approaches:

- Optimal variable selection.
- Consideration of the multi-perspective characterisation of ENMs.
- Automated and optimised workflows.
- Evaluation of the proposed methodologies using internal and external validation schemes and y-scrambling.
- Definition of the applicability domain of the models.
- Dissemination of the developed methodologies/models in the form of user-friendly web applications.

The proposed read-across methodologies are summarized next:

- A read-across methodology, which defines one or more similarity thresholds that indicate neighbouring ENMs to a query ENM (in short "Thresholding grouping").
- A read-across approach, which partitions ENMs into groups, in terms of one or more of their known input features (in short "Grouping using the features").
- A read-across methodology, which partitions ENMs into groups, in terms of their corresponding endpoint values (in short "Grouping using the endpoint").
- Predictive read-across workflows using the *k*-Nearest Neighbour machine learning methodology (in short "*k*NN/read-across models").

The first read-across approach was a mixed integer non-linear mathematical optimisation problem (MINLP), designed to include one or multiple similarity criteria corresponding to different types of ENMs properties. A tailor-made genetic algorithm was developed in order to compute an approximate solution, because the problem cannot be solved globally and efficiently by standard optimisation methods. In this scheme, controlled variable selection was integrated and the multi-perspective characterisation of ENMs was taken into account by selecting neighbours systematically. Both numerical and categorical endpoints can be predicted with this approach.

The second approach extended and adapted the mixed integer linear programming (MILP) model developed by Cardoso *et al*, [119] to the complex structures and the multi-perspective characterisation of ENMs. This grouping/read-across method selects from the available features-properties of the ENMs, the ones that best divide the multi-dimensional input space into regions and creates groups of similar ENMs. In each group a special linear regression function is automatically created for predicting the endpoint of interest. The grouping boundaries, the regression models and the optimal variable selection are the results of the solution of the MILP problem.

Building further on the idea of grouping, we developed an alternative read-across methodology for the formation of regions of similar ENMs, but in this case the endpoint was partitioned into regions for defining the groups. Again a MILP problem was structured, which defines the grouping boundaries, selects the most informative variable and creates the regression models in each group. The "characteristic centroid" of each group is later calculated and is used for the allocation of untested ENMs into groups.

Finally, the *k*NN machine learning method was tested and explored as a read-across methodology, because it uses the concept of similarity for defining neighbours and producing the endpoint predictions. The advantages and disadvantages of the four proposed read-across approaches are summarized in Table 8.1.

The application of the methods on benchmark datasets illustrated the efficiency of the algorithms in terms of providing accurate read-across predictions, since in most cases they outperformed the models that have been presented in the Literature.

The most comprehensive dataset on which all four read across models were applied, was the *Gold ENMs* dataset. This particular dataset defines a regression problem that contains both physicochemical and biological features. It gave us the opportunity to apply all four read-across models and exploit the full functionalities of the methods, which are able to define multiple similarity thresholds and group ENMs in the multidimensional space. The results are summarised in Table 8.2, which compares the algorithms in terms of the predictive Q^2_{ext} metric and the computational time needed to run the algorithms in a personal computer with the characteristics presented in Table B′.1. The results illustrate that computational requirements were comparable between all four methods. With respect to prediction accuracy, the models developed using the mathematical programming approaches clearly outperformed the *k*NN machine learning algorithm.

The read-across methods and workflows developed in this Dissertation have been implemented as user-friendly web applications which increase the visibility and the sustainability of the methods, and offer to the stakeholders the possibility to use the existing models, but also apply the read-across algorithms to new datasets and develop and share new predictive models.

The produced methodologies can be applied in various stages during the life cycle of ENMs. That includes the design of new safer and more efficient ENMs and the risk assessment of using ENMs in human and environmental applications. All methods select the most informative features and estimate their effects on the predicted endpoints. This is important information for material designers, because they could define the ENM characteristics that need to be designed out to eliminate or reduce the hazardous effects and the features that can

be optimised to improve their functionalities for the particular applications for which they are intended. They can also guide experimental designs for gaining additional experimental information and evidence, in specific regions of the ENM, where data are sparse.

To conclude, we believe that the methodologies developed in this Dissertation contribute greatly to the emerging field of nanoinformatics, because they introduce new approaches and mathematical foundations to the read-across concept, they have been validated extensively in benchmark datasets and can find practical applications in the areas of computer aided material design and regulatory risk assessment.

8.1 Future challenges

Read-across approaches are used extensively in computational modelling and predictive applications in the field of nanoinformatics to address the limited availability of experimental data and the high variance in the input space, which is a consequence of the plethora of protocols and methods that are used for characterising the ENMs. Most existing read-across methods are based on manual and iterative techniques. Experts in the field are employed to examine the validity of several grouping hypotheses before arriving at read-across predictions, which in many situations are not optimal, because they have explored only a limited part of the search space. This Thesis has contributed to the field of nanoinformatics by developing fully automated read-across algorithms that address the aforementioned limitations. In particular the proposed methods are able to automate and optimise the processes of selecting the most informative features, defining the grouping boundaries, and determining the mathematical functions that predict the functionalities and the adverse effects of ENMs.

However, the application of the methods on big datasets may reveal some limitations on the applicability of the proposed approaches, because they are computationally demanding. The performance of the models proposed algorithms is closely coupled with the advancements and improvements on mathematical programming theory, algorithms and available software. Further increase of computational power will support the application of the methods to larger case studies. In particular, the consideration of omics data in the proposed approaches, can be perfectly combined with the concept and recent developments in the field of Adverse Outcome Pathways (AOPs), and can result in methodologies that will be able to group ENMs or other substances in terms of triggering initiation of the same or similar AOPs. We have already illustrated that the methods can perfectly handle biological information, like protein corona datasets, so the adaptations and the adjustments of the proposed read-across approaches to omics data will be straightforward.

Finally, the increased visibility the proposed read-across methodologies through the web implementations and the outreach to the wider scientific and stakeholder communities, may extend the applicability of the methods to other scientific domains that can benefit from the development of predictive modelling approaches.

Table 8.1: Advantages and disadvantages of the developed methodologies

Developed methodology	Advantages	Disadvantages
Thresholding grouping methodology	Automated variable selection and regularisation Automated definition of grouping boundaries Susceptible to local similarities Inclusion of multi-perspective ENMs characterisation Numerical and categorical endpoint predictions Tested for different case studies Reliable predictions Exploration of "neighbourhoods" Disseminated methodology	Time consuming when dealing with large datasets Prediction may be impossible when neighbours are not found Many user-defined parameters (lack of total automation) Stochastic methodology (the exact same model is not produced for the same input data) Near-optimal solutions
Grouping using the features methodology	Automated variable selection and regularisation Automated definition of grouping boundaries Inclusion of multi-perspective ENMs characterisation Prediction for any input sample Reliable predictions Exploration of "neighbourhoods" Disseminated models	Not sufficiently tested in different case studies Prediction not available for categorical endpoints
Grouping using the endpoint methodology	Automated variable selection and regularisation Automated definition of grouping boundaries Total optimisation Prediction for any input sample Reliable predictions Exploration of "neighbourhoods"	Not sufficiently tested in different case studies Prediction not available for categorical endpoints
*k*NN/read-across models	Reliable and fast predictions Prediction for any input sample Numerical and categorical endpoint predictions Exploration of "neighbourhoods" Disseminated models	Lack of automation Variable selection is not automatically coupled with the prediction

Table 8.2: Training time of the different methodologies applied on the *Gold ENMs* dataset.

Developed methodology	Model	Training time [min]	Software	$Q^2_{\mathbf{ext}}$	Selected variables	Notes
Thresholding grouping	§4.3.6.2	14.91	MATLAB	0.78	27	*predFactor* = 0.3 & $wf_{\mathrm{OF}} = 0.05$ (prediction is not achieved for all ENMs)
Grouping using the features	§5.5.1	10.38	MATLAB	0.88	39	λ = 0.01 & 1D MILP
Grouping using the endpoint	§6.6	7.41	Python	0.83	28	
*k*NN/read-across models	page 69	10	KNIME	0.72	11	Approximate time (no automated process)

Appendix A′

Software tools

In this part we are briefly presenting the software tools used in this Thesis, in order to implement whole parts of the presented analysis or individual steps, such as data preprocessing or deployment of web applications. We have mainly used in Microsoft Windows 10 OS however, Docker platform was used under Linux Ubuntu OS. More information about the software used for the elaboration of this Thesis can be found in the Appendix B′.

A′.1 Programming languages and platforms

A′.1.1 MATLAB

MATLAB® is a commercial high-performance programming language developed by MathWorks, popular among scientists and engineers. It offers a great variety of data structures and powerful tools for programming, computation, visualisation and deployment of algorithms, as well as, it offers debugging tools and supports object-oriented programming. [153]

The integrated functions allow a wide variety of computations however, it can be extended via specialized "toolboxes" for use in different fields of applied mathematics and engineering including communications, control theory and signalling, dynamics, econometrics, image processing, optimisation etc.

A′.1.1.1 Optimisation in MATLAB

In the present Thesis, in order to formulate and solve mathematical optimisation problems we incorporated the free `YALMIP` toolbox. [121] It offers special commands and access to a variety of solvers either internal or external (e.g. `mosek`, `gurobi`, cplex etc.) thus a variety of mathematical optimisation problems can be efficiently solved (e.g. LP, MILP, quadratic programming, semi-definite programming etc.). [154] In this Thesis, mosek solver (`www.mosek.com`) was used in order to solve MILP problems (Chapter 5).

A′.1.2 R

R is a complete software for data manipulation, statistical analysis and graphical representation. R is an open-source, free software distributed under GNU General Public License (GPL) version 2 or 3 (`www.gnu.org/licenses/gpl-3.0.html`). It was initially developed by Robert Gentleman and Ross Ihaka (University of Auckland, New Zealand) inspired by S programming language. Currently, the "R core team" is responsible for R's development, however due to its open-source character any users from the broader community can contribute to its evolution. [155]

R facilitates standard programming structures (conditionals, loops, user-function creation, etc.), data storage and manipulation, due to its wide variety of flexible data structures and

indexing functions. It is highly extensible through "packages" that contain appropriate functions for specific data types (e.g. biological data) and specialized analysis tools. In that way it can perform robust statistical analysis and modelling, and produce high-quality graphical representations. [155] The vanilla R distribution offers 14 packages, however more packages are available from different repositories such as the "Bioconductor" (`www.bioconductor.org/`), the "Comprehensive R Archive Network (CRAN)" (`cran.r-project.org/`) and, the "Omega Project for Statistical Computing" (`www.omegahat.net/`). Packages can be also found in GitHub repositories. Some of the packages used in this Thesis, are briefly presented in Table A′.1 and in more detail in Appendix B′.2.2.1.

R was selected for its simplicity in data handling, its free and open character, its popularity and for the possibility of developing web applications through `shiny` package. Scripting in R was performed in Rstudio (`www.rstudio.com`), which is a free, open-source integrated development environment (IDE) for R.

A′.1.2.1 The `shiny` package

An extensive reference should be made for `shiny` and `shiny`-related packages built by RStudio (`shiny.rstudio.com`), witch are appropriate for the development of interactive web applications. At the time of writing there have been already released `shiny`-based apps dedicated in the bioinformatics and nanoinformatics field (e.g. toxFlow [45] and eUTOPIA [156]). In this Thesis, three web applications developed in `shiny` are presented; toxFlow, Apellis and vythos (see §A′.2.1, Chapters 4.5 and 5 respectively).

A `shiny` application is composed of two main parts, usually written in two different scripts that communicate with each other: an interactive web-page which is intended for the "communication" of the user with the application and a server file (main-code) that "propels" the application. The application runs under the developer's personal computer but it can also be packaged up in a container (e.g. using Docker) or hosted in a remote server.

In the user-interface file (`ui.R`) the graphical layout of the application is defined, using HTML/CSS/JavaScript elements wrapped in `shiny` functions; input and output fields are defined, as well as formatted text and elements (e.g. images or videos) that frame the interface. Input elements, including buttons, sliders, checkboxes, drop-down menus etc., support the communication between the users and the application. Output elements include plain text, tables, graphs etc. depending on application's scope. The main-code file (`server.R`) is responsible for the proper operation of the application: it processes the input data and produces results that are presented back in the web-page, when a user is interacting with the application. Input data can be altered any-time due to the aforementioned interaction nevertheless, developers can tune the levels of interaction in order to avoid unnecessary re-execution of some parts of the server script. [157], [158]

There are many online and free tutorials about `shiny` applications development, and an active community of `shiny` apps developers that constantly enrich `shiny` functionalities. [158]–[160]

A′.1.3 Python

Python is a powerful general-purpose programming language distributed as an open-source software. It was initially created in the late 80s by Guido van Rossum (Stichting Mathematisch Centrum, Netherlands) and currently it is supported by the "Python Software Foundation". Python was selected as an equivalent alternative to R. Scripting in Python was performed in PyCharm and in Jupyter Notebook (see §B′.2.3).

Python has a simple and straightforward syntax that renders it easy-to-learn and ideal for rapid scripting. It offers a wide variety of data types including -apart from the ordinary data types- *lists*, *tuples* and *dictionaries*. The common programming operations such as conditionals,

loops etc. can be extended using user-defined functions or using "modules" or "packages" (collection of modules) available in Python's library or in external repositories such as PyPI (`pypi.org`). Furthermore, considering that it is an object-oriented language, users are able to create easily their own classes and methods that are specified for their needs. Finally, Python is used in numerous applications including machine learning, deep learning and data science, web and game development, image processing, web scraping etc. [161]

A′.1.3.1 Optimisation in Python

In order to formulate and solve MILP problems in Python, we used the `mip` module (`www.python-mip.com`). It offers special commands and access to solvers including CBC solver [162] and Gurobi. In this Thesis, CBC solver was used in order to solve MILP problems (Chapter 6).

A′.1.4 KNIME

KNIME (Konstanz Information Miner) platform is a user-friendly and open-source software for data integration, reporting and analysis. It was initially developed as a proprietary software at the University of Konstanz from a software developer team headed by Michael Berthold. From version 2.1 and beyond, KNIME Analytics platform is released under GPL v.3 however, licensed commercial software extensions are also available. [163] In addition due to its open-source character is highly extensible, by integrating tools from other open-source projects (e.g. R, Keras and WEKA) or by creating custom-made nodes. [164]

One of the main advantages of KNIME is the intuitive nature and the modular character of its graphical environment. It enables users to create visual data pipelines (workflows), consisting of nodes and connections between them. For this reason, KNIME does not require any programming skills, in order to be used. Data are processed and/or transformed in the nodes, that represent concrete tasks or functions, and are transported through the connections. Workflows can be selectively or entirely executed, giving users the flexibility to experiment easily between different methodologies and compare the results among different tuning parameters. [3], [164]

KNIME also bridges tools from different disciplines, projects, databases and software suites under the same platform (e.g. CDK, RDKit, `WEKA`, Enalos+, ImageJ etc.). It is also possible to write and execute R or Python scripts through the appropriate nodes in order to address specialized needs that cannot be covered by existing KNIME nodes. Through the aforementioned features KNIME enables data mining from different sources (e.g. ChEMBL molecules database, Twitter, Google Analytics etc.) and of various types (e.g. CSV, JSON and XML files, images, molecules etc.), (big) data analysis, machine learning modelling, image analysis, data visualisation and reporting, rendering it a powerful tool in different fields including cheminformatics, pharmaceutical research, customized enterprise data management etc. [3], [90], [163], [164]

A′.1.5 Docker

Docker (`www.docker.com`) is an open platform that offers development, delivery and deployment of applications and software using *containers*. The *containers* are virtual isolated packages comprised of the application, the necessary libraries and its dependencies and, they can run on any Linux kernel. In that way a dockerized application is independent of the hardware and can be distributed quickly on any host (e.g. personal computers, private or public cloud, virtual machines etc.). Docker *containers* need a few resources and are "lightweight" compared to other virtualization techniques thus, more *containers* can run on a given host. [165]

Table A′.1: Summary of programming software employed in the Thesis

Software	Version	Extensions - modules - packages - toolboxes
R	3.6.2	`base`, `caret`, `DT`, `e1071`, `extrafont`, `ggplot2`, `mltools`, `plotly`, `prospectr`, `rapportools`, `RColorBrewer`, `shiny`, `shinyalert`, `shinycustomloader`, `shinyjs`, `shinyLP`, `shinythemes`, `shinyWidgets`, `stats`
MATLAB®	9.2 (2017a)	`YALMIP`, `mosek`
Python	3.9.0	`math`, `matplotlib`, `mip`, `multiprocessing`, `numpy`, `pandas`, `scikit-learn`, `statistics`
KNIME	3.7.1	`Enalos+`, `R`, `WEKA`

In order to build a *container* a file-system must be provided through a Docker *image*. The *image* incorporates all the necessary elements of the application (e.g. script, libraries and other dependencies, run-time, system configurations etc.). The instructions for "composing" a portable *image* are recorded in a *Dockerfile* which provides the environment for the *container*, including system files, script files and communication endpoints (ports). The developed Docker *images* can be stored (*pushed*) in public (e.g. Docker Hub) or proprietary *registries* and they can be later downloaded (*pulled*) and executed on any host.

A′.2 Web applications

A′.2.1 toxFlow

toxFlow (`https://toxflow.jaqpot.org/`) developed by Varsou *et al*. [45] is a user-friendly R `shiny` web application developed for enrichment analysis of omics data paired with a read-across workflow for the prediction of ENMs toxicity-related endpoints. The development of the application is founded on the multi-perspective characterisation of ENMs thus, physicochemical, theoretical, omics, biology etc. information data can be integrated in a read-across analogue prediction. More specifically, the toxicity endpoint prediction of the target ENM is performed using the weighted average of the corresponding values of the neighbour ENMs. Neighbouring ENMs of every target ENM are selected by calculating pairwise similarity measures with all available ENMs separately for physicochemical (or other type of available descriptors) and biological descriptors and by excluding those ENMs for which one or both similarity measures do not fulfil predefined thresholds.

The novelty of the toxFlow application is the possibility of filtering omics data (in case of availability) using an enrichment analysis scheme through GSVA. [81] Using GSVA analysis tools over- or under-expressed genes in omics experiments are estimated and organized into specific functional groups (statistically significant gene sets). The user in the read-across prediction can select whether only the significant genes or proteins (according to the data) from the GSVA analysis will be used. Therefore, by exploiting prior biological information, the results may be interpreted in a more accurate background.

A′.2.2 NanoXtract

NanoXtract developed by Varsou *et al*. [3] is a user-friendly online tool for the calculation of image descriptors from ENMs TEM images. Interested users can upload an ENM TEM image and with just a few clicks, they can obtain 18 image descriptors encoding the size, the shape, the geometry and the morphology of the depicted nanoparticles. Therefore, already existing sets of nano-descriptors can be enriched, without any need of further experiments,

with the generated information from TEM images and they can be later used in a predictive modelling framework. NanoXtract is made available through the Enalos Nanoinformatics platform (enaloscloud.novamechanics.com/EnalosWebApps/NanoXtract/).

Appendix B′

Software packages list

For the completion of the present Thesis, including coding, software development and writing, the following software packages were used. A detailed presentation of the necessary software for the main results generation, is presented in Appendix A′.

B′.1 Operating systems

`Windows 10 Home` (v1903) It is part of the OS family produced by Microsoft Co. The system details are presented in Table B′.1. More information in: `www.microsoft.com/`

Table B′.1: System details

CPU	Intel® Core™ i7-7500U CPU @2.70GHz 2.90GHz
RAM memory	8 GB
System type	OS 64 bit

`Ubuntu` (v16.04) It is a Linux OS distribution released under a GNU GPL license. More information in: `ubuntu.com`

B′.2 Programming software

B′.2.1 MATLAB

`MATLAB`® (v9.2/2017a) It is a powerful numerical computing environment and a programming language developed by MathWorks. More information in: `www.mathworks.com/products/matlab.html`

B′.2.1.1 Toolboxes

`YALMIP` (R20190425) YALMIP is a software framework for mathematical optimisation, free to use and openly distributed. More information in: `yalmip.github.io`

`Mosek` (v9.0) Mosek is an optimisation package developed to solve large-scale mathematical optimisation problems (linear, mixed-integer linear, conic and quadratic). It is supported by different platforms including MATLAB®, R, Java and Python. In the present Thesis, mosek was used under a Personal Academic License. More information in: `www.mosek.com`

B′.2.2 R

`R` (v3.6.2) A programming language and free software environment for statistical calculations and graphical representations developed by the R Foundation for Statistical Computing. More information in: `www.r-project.org`

`RStudio` (v1.1.423) An IDE for R programming language, distributed under AGPL v3. `shiny` package is available only via RStudio. More information in: `rstudio.com`

B′.2.2.1 Packages

`base` (v3.6.2) This package includes the basic-essential R functions (e.g. arithmetic functions, input/output functions, basic programming support, etc.). More information in: `stat.ethz.ch/R-manual/R-devel/library/base/html/base-package.html`

`caret` (v6.0.85) `caret` package offers functions for training of regression and classification models. More information in: `www.rdocumentation.org/packages/caret/versions/6.0-86`

`DT` (v0.13) `DT` package gives access to the DataTables JavaScript library through R. Data frames or simple matrices in R can be rendered as interactive tables on HTML pages, providing features for data filtering, pagination, sorting etc. More information in: `cran.r-project.org/web/packages/DT/index.html`

`e1071` (v1.7.3) This package includes a collection of machine learning functions. More information in: `https://cran.r-project.org/web/packages/e1071/index.html`

`extrafont` (v0.17) This package contains tools to handle fonts in R. More information in: `cran.r-project.org/web/packages/extrafont/index.html`

`ggplot2` (v3.3.0) This package is based on "The Grammar of Graphics" and offers a powerful graphics language for creating elegant and complex plots for either numerical or categorical data, including also grouping of data with color, symbol, size etc. More information in: `ggplot2.tidyverse.org`

`mltools` (v0.3.5) This package offers a collection of machine learning exploratory and diagnostic functions. More information in: `cran.r-project.org/web/packages/mltools/index.html`

`plotly` (v4.9.2.1) `plotly` is an R package for creating interactive web and publication-quality graphics (e.g. line or scatter plots, heatmaps, area and bar charts, 3D graphs etc.) via the open source JavaScript graphing library plotly.js. More information in: `plotly.com/r/`

`prospectr` (v0.1.3) `prospectr` package contains functions for spectroscopic data preprocessing and for representative sample selection. More information in: `cran.r-project.org/web/packages/prospectr/index.html`

`RColorBrewer` (v1.1.2) `RColorBrewer` package provides color palettes for graphics in R. More information in: `cran.r-project.org/web/packages/RColorBrewer/index.html`

`rapportools` (v1.0) This package offers a collection of helper functions that wrap advanced statistical methods. More information in: `cran.r-project.org/web/packages/rapportools/`

index.html

shiny (v1.4.0) This package offers a wide variety of necessary tools for the easy development of interactive web applications through R. Applications can be extended using CSS themes, JavaScript elements etc. More information in: shiny.rstudio.com/

shinyalert (v1.0) shinyalert package is used in order to show to the user a message in a modal (e.g. pop-up, dialog, or alert box). Modals may contain text, images, OK/Cancel buttons, an input to get a response from the user, and other tailor-made features. More information in: www.deanattali.com/blog/shinyalert-package/

shinycustomloader (v0.9.0) This package is used in order to add a custom CSS/HTML, a custom text or a gif/image file for the loading screen in shiny application. More information in: www.emitanaka.org/shinycustomloader/

shinyjs (v1.1) This package offers functions that perform useful JavaScript operations in shiny apps: functions to improve user-experience (e.g. disable/enable inputs, hide/show elements etc.) and functions that will help developers during apps building (e.g. call custom-made JavaScript functions through R). More information in: www.deanattali.com/shinyjs/

shinyLP (v1.1.2) This package contains functions that wrap HTML Bootstrap elements to create a landing page for shiny applications, in order to improve the user-experience. More information in: cran.r-project.org/web/packages/shinyLP/index.html

shinythemes (v1.1.2) This package offers a collection of CSS themes ready-to-use in a shiny application, in order to alter its appearance and style. More information in: rstudio.github.io/shinythemes/

shinyWidgets (v0.5.0) shinyWidgets package offers a collection of custom input controls and user-interface components for shiny applications. More information in: cran.r-project.org/web/packages/shinyWidgets/index.html

stats (v3.6.2) This package offers a broad collection of statistical functions. More information in: stat.ethz.ch/R-manual/R-devel/library/stats/html/stats-package.html

B′.2.3 Python

Python (v3.9.0) A powerful general-purpose programming language, distributed as a free software. More information in: www.python.org

PyCharm (v2020.1.1) An IDE for Python programming language. More information in: www.jetbrains.com/pycharm

Anaconda Navigator (v2020.11) A GUI included in Anaconda® distribution adequate for the management -without any need of command line- of packages and programming environments. More information in: docs.anaconda.com/anaconda/navigator/

Jupyter Notebook (v6.1.4) An open-source web application adequate for the creation and sharing of reports containing code, visualisations and explanatory text. More information in: jupyter.org

B′.2.3.1 Modules

math This module offers access to various mathematical functions. More information in: docs.python.org/3/library/math.html

matplotlib (v3.3.2) This module offers a collection of functions for the creation of static, animated and interactive diagrams, plots and other visualisations in Python. More information in: matplotlib.org

mip (v1.12.0) It offers a number of tools for modelling and solution of mixed-integer linear programs. More information in: www.python-mip.com

multiprocessing (v3.9.1) This module is used in order to run in parallel different processes in Python. More information in: docs.python.org/3/library/multiprocessing.html

numpy (v1.19.2) It is an open-source library for machine learning in Python. More information in: numpy.org

pandas (v1.1.2) It is a free Python library that permits efficient manipulation of simple or complex data structures. More information in: pandas.pydata.org

scikit-learn (v0.23.2) It is an open-source collection of powerful tools for predictive data analysis and machine learning; including regression, classification and clustering algorithms. More information in: scikit-learn.org [125]

statistics (v1.0.3.5) This module offers access to various mathematical statistics functions of numeric data. More information in: docs.python.org/3/library/statistics.html

B′.2.4 KNIME

KNIME (v3.7.1) A free and open-source data analytics, reporting and integration platform. More information in: www.knime.com

B′.2.4.1 Extensions

Enalos+ (v1.0.0) Enalos+ nodes, designed and developed by NovaMechanics Ltd., are a useful aid in cheminformatics and nanoinformatics problems. They simplify tasks performed in molecular modelling they give access to multiple chemical databases through the KNIME interface, and facilitate data mining and manipulation of data directly in KNIME. More information in: www.enalosplus.novamechanics.com/, www.knime.com/community/enalos-nodes

R (v3.3.0) KNIME includes a branch of nodes for the reduction of R scripts and the execution of ready-to-use R templates (tables manipulation, graphs creation etc.), in a local R installation (package Rserve from CRAN is required). Python and MATLAB® extensions are also available. More information in: www.knime.com/community/scripting

WEKA (v3.7) KNIME offers a WEKA extension that incorporates the tools of WEKA data mining suite. WEKA is an open-source software that offers machine learning functions and gives access to scikit-learn, R, and Deeplearning4j. WEKA can be used through a user-friendly interface, but also through terminal or a Java API. More information in: hub.knime.com/knime/extensions/org.knime.features.ext.weka_3.7/latest,

www.cs.waikato.ac.nz/ml/weka/

B′.3 Informatics platforms

Jaqpot Jaqpot 5, developed by UPCI in NTUA, is a user-friendly web-based e-infrastructure that allows model developers to deploy their predictive models and share them through the web. More information in: app.jaqpot.org

toxFlow A free to use web application developed for enrichment analysis of omics data and read-across toxicity prediction. More information in: https://toxflow.jaqpot.org/

Enalos Cloud Platform An online toxicity and drug discovery platform developed by NovaMechanics Ltd that hosts predictive models released as web services, for the support of cheminformatics and nanoinformatics-related challenges. More information in: enaloscloud.novamechanics.com/

nanoXtract An online tool, free to use, for the calculation of image descriptors from ENMs TEM images. More information in: enaloscloud.novamechanics.com/EnalosWebApps/NanoXtract/

B′.4 Applications deployment

Docker (v19.03.5) An open platform that offers development, delivery and deployment of applications and software using *containers*. More information in: www.docker.com

B′.5 Graphics editors

GNU GIMP (v2.8.22) The GNU Image Manipulation Program (GIMP) is a free and open-source raster graphics editor released under GPLv3+ licenses. It is used for image retouching and editing, free-form drawing and other more sophisticated tasks. More information in: www.gimp.org

Inkscape (v0.92.3) It is a free and open-source vector graphics editor used to create vector images. More information in: inkscape.org

BIOVIA™ Draw (v21.1.0.2363) It is a software adequate for drawing and editing complex molecules, chemical reactions and biological sequences with ease. BIOVIA Draw was used under an academic license. More information in: discover.3ds.com/biovia-draw-academic

Appendix Γ′

Datasets' descriptor details

In this Appendix information about the descriptors of the datasets included in Chapter 3 is presented.

Table Γ′.1: List of the biological descriptors included in the *Gold ENMs* dataset and a brief explanation of their physical meaning. [76]

Descriptor	Brief meaning
O43866	CD5 antigen-like (www.uniprot.org/uniprot/O43866)
P02654	Apolipoprotein C-I (www.uniprot.org/uniprot/P02654)
P04114	Apolipoprotein B-100 (www.uniprot.org/uniprot/P04114)
P01011	Alpha-1-antichymotrypsin (www.uniprot.org/uniprot/P01011)
P00739	Haptoglobin-related protein (www.uniprot.org/uniprot/P00739)
P02760	Protein AMBP (www.uniprot.org/uniprot/P02760)
P01009	Alpha-1-antitrypsin (www.uniprot.org/uniprot/P01009)
P03951	Coagulation factor XI (www.uniprot.org/uniprot/P03951)
P01008	Antithrombin-III (www.uniprot.org/uniprot/P01008)
P02656	Apolipoprotein C-III (www.uniprot.org/uniprot/P02656)
P02749	Beta-2-glycoprotein 1 (www.uniprot.org/uniprot/P02749)
P03952	Plasma kallikrein (www.uniprot.org/uniprot/P03952)
P02655	Apolipoprotein C-II (www.uniprot.org/uniprot/P02655)
P02774	Vitamin D-binding protein (www.uniprot.org/uniprot/P02774)
P00738	Haptoglobin (www.uniprot.org/uniprot/P00738)
P02671	Fibrinogen alpha chain (www.uniprot.org/uniprot/P02671)
P00751	Complement factor B (www.uniprot.org/uniprot/P00751)
P00734	Prothrombin (www.uniprot.org/uniprot/P00734)
P01019	Angiotensinogen (www.uniprot.org/uniprot/P01019)
P03950	Angiogenin (www.uniprot.org/uniprot/P03950)
P00740	Coagulation factor IX (www.uniprot.org/uniprot/P00740)
P01042	Kininogen-1 (www.uniprot.org/uniprot/P01042)
P01024	Complement C3 (www.uniprot.org/uniprot/P01024)
P04196	Histidine-rich glycoprotein (www.uniprot.org/uniprot/P04196)
P02790	Hemopexin (www.uniprot.org/uniprot/P02790)

continued from Table Γ.1

Descriptor	Brief meaning
P02743	Serum amyloid P-component (www.uniprot.org/uniprot/P02743)
P00736	Complement C1r subcomponent (www.uniprot.org/uniprot/P00736)
P01023	Alpha-2-macroglobulin (www.uniprot.org/uniprot/P01023)
P00742	Coagulation factor X (www.uniprot.org/uniprot/P00742)
P00748	Coagulation factor XII (www.uniprot.org/uniprot/P00748)
P05154	Plasma serine protease inhibitor (www.uniprot.org/uniprot/P05154)
P02788	Lactotransferrin (www.uniprot.org/uniprot/P02788)
P02649	Apolipoprotein E (www.uniprot.org/uniprot/P02649)
P00450	Ceruloplasmin (www.uniprot.org/uniprot/P00450)
P00451	Coagulation factor VIII (www.uniprot.org/uniprot/P00451)
P05546	Heparin cofactor 2 (www.uniprot.org/uniprot/P05546)
P06396	Gelsolin (www.uniprot.org/uniprot/P06396)
P08567	Pleckstrin (www.uniprot.org/uniprot/P08567)
P08709	Coagulation factor VII (www.uniprot.org/uniprot/P08709)
P09871	Complement C1s subcomponent (www.uniprot.org/uniprot/P09871)
P0C0L4	Complement C4-A (www.uniprot.org/uniprot/P0C0L4)
P0C0L5	Complement C4-B (www.uniprot.org/uniprot/P0C0L5)
P10720	Platelet factor 4 variant (www.uniprot.org/uniprot/P10720)
P10909	Clusterin (www.uniprot.org/uniprot/P10909)
P12259	Coagulation factor V (www.uniprot.org/uniprot/P12259)
P14618	Pyruvate kinase isozymes M1/M2 (www.uniprot.org/uniprot/P14618)
P15169	Inter-alpha-trypsin inhibitor heavy chain H2 (www.uniprot.org/uniprot/P19823)
P18065	Insulin-like growth factor-binding protein 2 (www.uniprot.org/uniprot/P18065)
P18428	Lipopolysaccharide-binding protein (www.uniprot.org/uniprot/P18428)
P19823	Inter-alpha-trypsin inhibitor heavy chain H2 (www.uniprot.org/uniprot/P19823)
P20851	C4b-binding protein beta chain (www.uniprot.org/uniprot/P20851)
P23528	Cofilin-1 (www.uniprot.org/uniprot/P23528)
P27169	Serum paraoxonase/arylesterase 1 (www.uniprot.org/uniprot/P27169)
P35542	Serum amyloid A-4 protein (www.uniprot.org/uniprot/P35542)
P49908	Selenoprotein P (www.uniprot.org/uniprot/P49908)
P68871	Hemoglobin subunit beta (www.uniprot.org/uniprot/P68871)
Q03591	Complement factor H-related protein 1 (www.uniprot.org/uniprot/Q03591)
Q06033	Inter-alpha-trypsin inhibitor heavy chain H3 (www.uniprot.org/uniprot/Q06033)
Q13103	Secreted phosphoprotein 24 (www.uniprot.org/uniprot/Q13103)
Q13790	Apolipoprotein F (www.uniprot.org/uniprot/Q13790)
Q14520	Hyaluronan-binding protein 2 (www.uniprot.org/uniprot/Q14520)
Q14624	Inter-alpha-trypsin inhibitor heavy chain H4 (www.uniprot.org/uniprot/Q14624)
Q99467	CD180 antigen (www.uniprot.org/uniprot/Q99467)

Table Γ.2: List of the physicochemical descriptors included in the *Gold ENMs* dataset and a brief explanation of their physical meaning. [76]

Descriptor	Brief meaning
AS.total	Total surface area
bca.density	Total adsorbed serum protein density
class	Surface classification: "anionic" (1) or "cationic" (0)
hdlayer.serum	zav.serum - tem.size[1]
hdlayer.synth	zav.synth - tem.size
hdrel.serum	zav.serum/tem.size
hdrel.synth	zav.synth/tem.size
int.serum	Intensity mean HD after serum exposure
int.ch	int.serum - int.synth
int.rel	int.serum/int.synth
int.synth	Intensity mean HD after synthesis
lspri.rel.ch	lspri.diff/lspri.synth
lspri.diff	lspri.serum - lspri.synth
lspri.relative	lspri.serum/lspri.synth
lspri.serum	LSPR index after serum exposure
lspri.synth	LSPR index after synthesis
num.ch	num.serum - num.synth
num.rel	num.serum/num.synth
num.serum	Number mean HD after serum exposure
num.synth	Number mean HD after synthesis
pdi.ch	pdi.serum - pdi.synth
pdi.rel	pdi.serum/pdi.synth
pdi.serum	Polydispersity index after serum exposure
pdi.synth	Polydispersity index after synthesis
vol.ch	vol.serum - vol.synth
vol.rel	vol.serum/vol.synth
vol.serum	Volume mean HD after serum exposure
vol.synth	Volume mean HD after synthesis
zav.ch	zav.serum - zav.synth
zav.rel	z-average HD after serum exposure
zav.serum	z-average HD after serum exposure
zav.synth	z-average HD after synthesis
zp.ch	zp.serum - zp.synth
zp.rel	zp.serum/zp.synth
zp.serum	Zeta-potential after serum exposure
zp.serum.mag	Magnitude of zeta-potential after serum exposure
zp.synth	Zeta-potential after synthesis
zp.synth.mag	Magnitude of zeta-potential after synthesis
zp.synth.sign	Sign (signum) of zeta-potential after synthesis

[1]Core ENM size

Table Γ΄.3: List of the descriptors included in the *MeOx ENMs [a]* dataset and a brief explanation of their physical meaning. [82]

Descriptor	Brief meaning
x(H2O)	Hydration rate
n_oxy_M	Oxidation degree of the metal
r_cat	Radius of the metalic cation
EN_M	Pauling electronegativity
s(log Mtot)	Solubility
zeta	Zeta-potential at pH = 7
SSA	Specific surface area
dmin	Minimum diameter
dmax	Maximum diameter
SF	Shape factor = (dmax/dmin)
CSF	Corrected shape factor = $\log(d_{axis}/d_{perp})$
Agg	Agglomeration/Aggregation state

Table Γ΄.4: List of the descriptors included in the *MeOx ENMs [b]* dataset and a brief explanation of their physical meaning. [83], [84]

Descriptor	Brief meaning
d	Primary ENM size
HD_{water}	MeOx hydrodynamic size in water
HD_{BEGM}	MeOx hydrodynamic size in BEGM
HD_{DMEM}	MeOx hydrodynamic size in DMEM
E_C	ENM energy of conduction band
E_V	ENM energy of valence band
E_{Amz}	MeOx atomization energy
χ_{MeOx}	MeOx electronegativity
ΔH_{sub}	MeOx sublimation enthalpy
ΔH_{IE}	MeOx ionization energy
$\Delta H_{IE,1+}$	First molar ionization energy of metal
ΔH_{sf}	MeOx standard molar enthalpy of formation
ΔH_{Lat}	MeOx lattice enthalpy
Z^2/r	Ionic index of metal cation
IEP	ENM isoelectric point
ZP	Zeta-potential at pH = 7.4
χ_{Me}	Electronegativity of metal
n_{Me}	Metal atoms
n_O	Oxygen atoms
AM	Atomic mass of metal
MW	MeOx molecular weight
μ	Chemical potential
η	Chemical hardness
ω	Electrophilicity

Table Γ΄.5: List of the descriptors included in the *MWCNTs [a]* dataset and a brief explanation of their physical meaning. [85]

Descriptor	Brief meaning
R	Lone-pair electrons
π	Polarizability
α	Hydrogen-bond acidity
β	Hydrogen-bond basicity
V	Lipophilicity interaction

Table Γ΄.6: List of the descriptors included in the *NanoMILE ENMs* dataset and a brief explanation of their physical meaning. [166]–[168]

Descriptor	Brief meaning
Core	The type of ENM core (pure metal or MeOx)
pH	The pH where the zeta-potential is measured
Area	The area of the ENM
Boundary size	Total length of the ENM's boundary (the perimeter calculated by different method)
Boxivity	The extent to which an ENM approaches a rectangle
Circularity[2]	The degree to which an ENM approaches a perfect circle
Convexity[3]	The ENM's edge roughness
Diameter	The ENM's diameter
Eccentricity	The measure of how much the ENM deviates from being circular
Extent	The boxivity calculated using different method
Main elongation	The lengthening of the ENM
Major axis	The longest diameter of the best fitting ellipse to the ENM
Maximum Feret's diameter	The longest distance between any two points along the selection boundary (calliper diameter)
Minimum Feret's diameter	The shortest distance between any two points along the selection boundary
Minor axis	The shortest diameter of the best fitting ellipse to the ENM
Perimeter	Total length of the ENM's boundary
Roundness	Compares the surface of the ENM to the surface of the disc of diameter equal to major axis
Solidity	The degree of the overall concavity or convexity of an ENM

[2]NanoXtract produces two circularity values, calculated by different KNIME nodes
[3]NanoXtract produces two convexity values, calculated by different KNIME nodes

Table Γ′.7: *NanoMILE ENMs* dataset details.

ENM sample	Alternative names	Type of core	pH	Zeta-potential [mV]
Aged PROM_CeO2_PO43	Aged PROM_CeO2_PO43-aged_21days	metal oxide	7	-7.12
AgNPs cit 10nm	-	pure metal	7	-35.00
AgNPs cit 20nm	-	pure metal	7	-46.00
AgNPs cit 7nm	-	pure metal	7	-45.00
AgNPs PEG 10nm	-	pure metal	7	-25.00
AgNPs PEG 20nm	-	pure metal	7	-16.00
AgNPs PEG 7nm	-	pure metal	7	-18.00
AgNPs PVP 10nm	-	pure metal	7	-24.00
AgNPs PVP 20nm	-	pure metal	7	-19.00
AgNPs PVP 7nm	-	pure metal	7	-20.00
AgPURE	-	pure metal	7	-1.21
Ce NM 211	-	MeOx	7	43.40
Citrate stabilized Au	-	pure metal	6.5	-41.80
Gold Citrate_IO166	Au NPs pristine, IO074	pure metal	6.5	-40.70
IO 160A	IO160A Copper 14 nm Dopamine, Cu NPs dopamine	pure metal	6.5	28.20
IO003	Au NPs pristine	pure metal	6.5	-17.20
IO083C	Au NPs amino-functionalized	pure metal	6.5	61.80
IO083F	Au NPs amino-functionalized	pure metal	6.5	41.20
IO140 A	Au NPs hydroxyl-functionalized	pure metal	6.5	-6.60
IO140 B	Au NPs carboxylic acid-functionalized	pure metal	6.5	-25.30
IO153B Silica 20 nm L-agrinine	JRC-silica-01	MeOx	6.5	-31.65
IO166D Gold 12 nm PEG-Ome	Au NPs methoxy functionalized (hydrophobic), IO166D	pure metal	6.5	-40.60
IO166E Gold 12 nm PEG-COOH	IO140D	pure metal	6.5	-38.90
IO166E Gold 12 nm PEG-OH	IO140C	pure metal	6.5	-7.30
JRC Ag (JRC Silver)	Ag NM-300K (JRC)	pure metal	7	-5.52
JRC-silica-10	Fluorescent SiO2 NPs pristine 20nm, SiO2-un RuBPy <20 nm JRC	MeOx	6.5	-31.65
PROM CeO2-Zr 024C	-	MeOx	7	45.90
prom ZnO	-	MeOx	7	25.60
S10	Sigma silver 10 nm	pure metal	7	-30.30
S100	Sigma silver 100 nm	pure metal	7	-48.80
S20	Sigma silver 20 nm	pure metal	7	-29.50
S40	Sigma silver 40 nm	pure metal	7	-46.20
S60	Sigma silver 60 nm	pure metal	7	-50.30
TiO2 103	-	MeOx	7	23.00
TiO2 104	-	MeOx	7	25.30
TiO2 NIST	-	MeOx	7	41.00
ZnO 110	-	MeOx	7	-20.30

Table Γ′.8: List of the descriptors included in the *SPIONs* dataset and a brief explanation of their physical meaning. [101]

Descriptor	Brief meaning
Magnetic core	Core of maghemite or magnetite
Overall size	Particle size
Relaxivity	Relaxation rate as a function of concentration
B_0	Magnetic field strength
Fe/cell	Iron concentration per cell

Γ′.1 MWCNTs *k*NN models molecular descriptor details

In this part, information about the selected descriptors used in the MWCNTs *k*NN models of §7.1 is presented. Descriptor information is derived from the following Bibliography items. [131], [169], [170] In the same sources, information about all the 777 calculated descriptors can be found.

- **D133 - Mean value of atomic composition index**: The atomic composition indices are molecular 0D descriptors with high degeneracy, derived from the chemical formula of compounds and defined as information indices of the elemental composition of the molecule. They can be considered molecular complexity indices that take into account the molecular diversity in terms of different atom types. The mean information content on atomic composition is the mean value of the total information content, and is calculated as:

$$\bar{I}_{AC} = -\sum_{g} \frac{A_g}{A^h} \cdot \log_2 \frac{A_g}{A^h} = -\sum_{g} p_g \cdot \log_2 p_g \tag{Γ′.1}$$

  where A^h is the total number of atoms (hydrogen included),
  A_g is the number of atoms of type g and,
  p_g is the probability to randomly select a gth atom type.

- **D173 - Mohar order-2 index**: The first Mohar index (TI_1) and the second Mohar index (TI_2) are calculated from the eigenvalues of the Laplace matrix as follows:

$$TI_1 = 2 \cdot \log \frac{B}{A} \cdot QW_L \tag{Γ′.2}$$

$$TI_2 = \frac{4}{nSK \cdot \lambda_{A-1}} \tag{Γ′.3}$$

  where QW_L is the quasi-Wiener index,
  A and B are the number of vertices in the molecular graph and the number of graph edges respectively and,
  λ_{A-1} is the first non-zero eigenvalue of the Laplace matrix.

◦ **D250 - EXP5 of Path-distance/Walk-distance over all atoms**: The path-distance map matrix, denoted as **PD**, resembling the bond length-weighted distance matrix of a molecular graph, is defined as:

$$[PD]_{i,j} = \min_{p_{i,j}}([EC]_{k,q})_{i,j} \tag{Γ′.4}$$

where $[\mathbf{ED}]_{kq}$ denotes entries of the Euclidean-distance map matrix,
p_{ij} is a path connecting vertices i and j and the summation goes over all the pairs of adjacent vertices along the considered path. Then, each entry of the path-distance map matrix is the shortest distance between two vertices measured along the path by summing the geometrical length of the edges connecting adjacent vertices along the path.

◦ **D254 - Radial centric index**: Centric indices are molecular descriptors proposed to quantify the degree of compactness of molecules by distinguishing between molecular structures organized differently with respect to their centers. Based on the recognition of the graph center, these indices are mainly defined by the information theory concepts applied to a partition of the graph vertices made according to their positions relative to the center.

Radial centric information index ($V_{\bar{I}_{C,R}}$) is defined as:

$$V_{\bar{I}_{C,R}} = -\sum_{g=1}^{G} \frac{n_g}{A} \log_2 \frac{n_g}{A} \tag{Γ′.5}$$

where n_g is the number of graph vertices having the same atom eccentricity (the maximum distance from a vertex to any other vertex in the graph),
G is the number of different vertex equivalence classes and,
A is the number of graph vertices.

◦ **D255 - Vertex distance count equality index**: The vertex distance counts, indicate the frequencies of distances equal to 1, 2, 3 etc. from vertex v_i to any other vertex. The vertex distance count of first-order ${}^1f_i$ coincides with the vertex degree δ_i, that is, the number of first neighbours, while ${}^2f_i$ and ${}^3f_i$ correspond to the connection number (i.e., number of second neighbours), and polarity number (i.e., number of third neighbours) for the i^{th} vertex, respectively.

◦ **D269 - Information content order-0 index**: The information content based descriptors are calculated from the information content of a molecule (I_c). I_c is used to measure the degree of diversity of the atoms or bonds in a molecule (Eq. Γ′.6).

$$I_c = \sum_{c=1}^{C} n_c \cdot \log_2 n_c \tag{Γ′.6}$$

where, C is the number of different types of atoms or bonds and,
n_c is the number of atoms or bonds of the cth type.

Mean information content, mean information content on edge equality, and redundancy index are some examples of information content-based descriptors.

◦ **D454 - Geary topological structure autocorrelation length-8 weighted by atomic masses | D468 - Geary topological structure autocorrelation length-6 weighted by atomic Sanderson electronegativities | D472 - Geary topological structure autocorrelation length-2 weighted by atomic polarizabilities | D473 - Geary topological structure autocorrelation length-3 weighted by atomic polarizabilities**: The Geary coefficient (c_k) is a general index of spatial autocorrelation that, if applied to a molecular graph and can be defined as:

$$c_k = \frac{\frac{1}{2\Delta_k}\sum_{i=1}^{A}\sum_{j=1}^{A}(w_i - w_j)^2 \cdot \delta(d_{ij};k)}{\frac{1}{A-1}\cdot\sum_{i=1}^{A}(w_i - \overline{w})^2} \tag{Γ'.7}$$

where w_i is any atomic property as a weighting factor,
$\overline{w}$ is its average value on the molecule,
A is the number of atoms,
k is the lag considered,
d_{ij} is the topological distance between ith and jth atoms and
$\delta(d_{ij};k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise.
Δ_k is the number of vertex pairs at distance equal to k.

Geary coefficient is a distance-type function varying from zero to infinite. Strong autocorrelation produces low values of this index; moreover, positive autocorrelation translates in values between 0 and 1 whereas negative autocorrelation produces values larger than 1; therefore, the reference "no correlation" is $c_k = 1$.

◦ **D522 - Mean molecular topological order-2 charge index**: Descriptors related to the topological charge index are derived from the adjacency matrix and distance matrix of a molecule, which estimate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule.

◦ **D541 - Lowest eigenvalue from Burden matrix weighted by van der Waals order 2**: Burden eigenvalues are demonstrated to reflect the topology of the whole molecule; the highest and the lowest eigenvalues reflect relevant aspects of molecular structure, useful for similarity searching, identification and ordering of molecular structures. In detail, Burden matrices (***B***) fall into the category of weighted adjacency matrices that encode information about proximity between vertices of molecular graphs that represent molecules containing heteroatoms and/or multiple bonds. The elements B_{ij} that represent two bonded atoms i and j are equal to $\pi^* 10^{-1}$, where π^* is the conventional bond order (0.1, 0.2, 0.3 and 0.15 for a single, double, triple and aromatic bond respectively; elements corresponding to terminal bonds are augmented by 0.01), the diagonal elements of the Burden matrix contain the atomic numbers Z_i of the atoms and all other matrix elements are set equal to 0.001. From Burden matrices, Burden eigenvalues are computed and are used in QSAR modelling. It is assumed that smallest eigenvalues contain contributions from all atoms of the molecule and therefore reflect the topology of the whole molecule and have high discrimination power.

Appendix Δ′

Additional results for methodology of Chapter 4

In this Appendix additional results regarding the methodology of Chapter 4 are presented.

Δ′.1 SPIONs dataset

Table Δ′.1: Neighbours between training and test ENMs of the *SPIONs* read-across model built using the GA workflow.

	Training samples									
Test samples	11	10	1	9	15	4	6	5	0	14
2	0	0	1	0	0	1	0	0	0	0
3	0	0	1	0	0	1	0	0	0	0
8	0	1	0	0	1	0	0	1	0	1
12	0	0	0	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	1	0

Δ′.2 Gold ENMs dataset

Table Δ′.2: Neighbours between training and test ENMs of the *Gold ENMs dataset* read-across model built using the GA workflow for $wf_{OF} = 0.05$.

	Training samples											
Test samples	G15.NT@DCA	G15.MUTA	G60.DTNB	G15.CTAB	G30.TP	G15.DDT@ODA	G60.AUT	G15.LA	G60.PVA	G60.MUTA	G60.CIT	G15.AC
G15.Ala-SH	0	0	0	0	0	0	0	1	0	0	0	1
G15.Asn-SH	0	0	0	0	0	0	0	1	0	0	0	1
G15.AUT	0	0	0	0	0	1	1	0	0	0	0	0
G15.CIT	0	0	0	0	0	0	0	0	0	0	0	0
G15.MAA	0	0	0	0	0	0	0	1	0	0	0	1
G15.MBA	0	0	0	0	1	0	0	1	0	0	0	1
G15.MHA	0	0	0	0	1	0	0	1	0	0	0	1
G15.MPA	0	0	0	0	1	0	0	1	0	0	0	0
G15.MUA	0	0	0	0	1	0	0	1	0	0	0	1
G15.NT@PSMA-AAP	0	0	0	0	0	0	0	1	0	0	0	1
G15.NT@PSMA-EA	0	0	0	0	0	0	0	1	0	0	0	1
G15.NT@PSMA-Urea	0	0	0	0	0	0	0	1	0	0	0	1
G15.PEI-SH	0	0	0	0	0	0	0	0	0	0	0	0
G15.Phe	0	0	0	0	0	0	0	0	0	0	0	0
G15.Phe-SH	0	0	0	0	0	0	0	1	0	0	0	1
G15.Ser-SH	0	0	0	0	1	0	0	1	0	0	0	1
G15.Thr-SH	0	0	0	0	0	0	0	1	0	0	0	1
G15.Trp-SH	0	0	0	0	0	0	0	1	0	0	0	1
G30.AC	0	0	0	0	0	0	0	1	0	0	0	1
G30.AUT	0	0	0	0	0	1	1	0	0	0	0	0
G30.CALNN	0	0	1	0	0	0	0	1	0	0	0	1
G30.DDT@HDA	0	0	0	0	0	0	0	0	0	0	0	0
G30.Met-SH	0	0	0	0	1	0	0	1	0	0	0	1
G30.MHDA	0	0	0	0	1	0	0	0	0	0	0	0
G60.HDA	0	0	0	0	0	1	1	0	0	0	0	0
G60.MPA	0	0	0	0	1	0	0	0	0	0	0	0
G60.Phe-SH	0	0	0	0	0	0	0	1	0	0	0	1
G60.Ser-SH	0	0	1	0	0	0	0	0	0	0	0	0
G60.SPP	0	0	1	0	0	0	0	0	0	0	0	0

(*continued from Table Δ′.2*)

	Training samples									
Test samples	G15.NT@PSMA-EDA	G15.PAH-SH	G15.MES	G60.DDT@BDHDA	G15.DDT@SDS	G60.MBA	G15.CALNN	G15.T20	G15.SA	G15.DDT@BDHDA
G15.Ala-SH	0	0	0	0	0	0	0	0	0	0
G15.Asn-SH	0	0	0	0	0	0	0	0	0	0
G15.AUT	0	1	0	0	0	0	0	0	0	0
G15.CIT	0	0	0	0	0	0	0	0	1	0
G15.MAA	0	0	0	0	0	0	1	0	0	0
G15.MBA	0	0	0	0	0	1	0	0	0	0
G15.MHA	0	0	0	0	0	1	0	0	0	0
G15.MPA	0	0	0	0	0	1	0	0	0	0
G15.MUA	0	0	0	0	0	1	1	0	0	0
G15.NT@PSMA-AAP	0	0	0	0	0	0	0	0	0	0
G15.NT@PSMA-EA	0	0	0	0	0	0	1	0	0	0
G15.NT@PSMA-Urea	0	0	0	0	0	0	0	0	0	0
G15.PEI-SH	0	1	0	0	0	0	0	0	0	0
G15.Phe	0	0	0	0	0	0	0	0	1	0
G15.Phe-SH	0	0	0	0	0	0	0	0	0	0
G15.Ser-SH	0	0	0	0	0	0	0	0	0	0
G15.Thr-SH	0	0	0	0	0	0	0	0	0	0
G15.Trp-SH	0	0	0	0	0	0	0	0	0	0
G30.AC	0	0	0	0	0	0	0	0	0	0
G30.AUT	0	0	0	0	0	0	0	0	0	0
G30.CALNN	0	0	0	0	0	0	1	0	0	0
G30.DDT@HDA	0	0	0	1	0	0	0	0	0	0
G30.Met-SH	0	0	0	0	0	0	0	0	0	0
G30.MHDA	0	0	0	0	0	0	0	0	0	0
G60.HDA	0	0	0	0	0	0	0	0	0	0
G60.MPA	0	0	0	0	0	1	0	0	0	0
G60.Phe-SH	0	0	0	0	0	1	0	0	0	0
G60.Ser-SH	0	0	0	0	0	1	0	0	0	0
G60.SPP	0	0	0	0	0	0	0	0	0	0

(*continued from Table Δ′.2*)

	Training samples											
Test samples	G15.PLL-SH	G60.CVVIT	G15.DDT@SA	G60.CTAB	G15.HDA	G30.Thr-SH	G15.MSA	G15.AHT	G30.DDT@DOTAP	G15.SPP	G15.PVP	G60.ODA
G15.Ala-SH	0	0	0	0	0	0	0	0	0	1	0	0
G15.Asn-SH	0	0	0	0	0	0	0	0	0	0	0	0
G15.AUT	1	0	0	0	0	0	0	1	0	0	0	0
G15.CIT	0	0	0	0	0	0	0	0	0	0	0	0
G15.MAA	0	1	0	0	0	1	0	0	0	1	0	0
G15.MBA	0	0	0	0	0	1	1	0	0	0	0	0
G15.MHA	0	0	0	0	0	1	1	0	0	0	0	0
G15.MPA	0	0	0	0	0	1	1	0	0	0	0	0
G15.MUA	0	0	0	0	0	1	1	0	0	0	0	0
G15.NT@PSMA-AAP	0	0	0	0	0	0	0	0	0	0	0	0
G15.NT@PSMA-EA	0	0	0	0	0	0	0	0	0	0	0	0
G15.NT@PSMA-Urea	0	0	0	0	0	0	1	0	0	0	0	0
G15.PEI-SH	1	0	0	0	0	0	0	1	0	0	0	0
G15.Phe	0	0	0	1	0	0	0	0	0	0	0	0
G15.Phe-SH	0	0	0	0	0	0	0	0	0	1	0	0
G15.Ser-SH	0	0	0	0	0	1	1	0	0	1	0	0
G15.Thr-SH	0	0	0	0	0	0	0	0	0	1	0	0
G15.Trp-SH	0	0	0	0	0	0	0	0	0	1	0	0
G30.AC	0	0	0	0	0	0	0	0	0	0	0	0
G30.AUT	0	0	0	0	0	0	0	1	0	0	0	0
G30.CALNN	0	1	0	0	0	1	0	0	0	1	0	0
G30.DDT@HDA	0	0	0	0	0	0	0	0	0	0	0	1
G30.Met-SH	0	1	0	0	0	1	0	0	0	1	0	0
G30.MHDA	0	0	0	0	0	0	0	0	0	0	0	0
G60.HDA	0	0	0	0	0	0	0	0	0	0	0	1
a G60.MPA	0	1	0	0	0	0	0	0	0	0	0	0
G60.Phe-SH	0	1	0	0	0	1	1	0	0	0	0	0
G60.Ser-SH	0	1	0	0	0	1	0	0	0	0	0	0
G60.SPP	0	1	0	0	0	0	0	0	0	0	0	0

(*continued from Table Δ′.2*)

	Training samples										
Test samples	G30.MAA	G15.DTNB	G15.DDT@CTAB	G15.Gly-SH	G30.PAH-SH	G30.CFGAILS	G30.LA	G15.ODA	G15.TP	G30.MUA	G15.MHDA
G15.Ala-SH	0	0	0	1	0	0	0	0	0	0	0
G15.Asn-SH	0	0	0	1	0	0	1	0	0	0	0
G15.AUT	0	0	0	0	0	0	0	0	0	0	0
G15.CIT	0	0	0	0	0	1	0	0	0	0	0
G15.MAA	1	1	0	1	0	1	0	0	0	0	0
G15.MBA	1	0	0	1	0	0	0	0	1	0	1
G15.MHA	0	0	0	1	0	0	0	0	1	1	1
G15.MPA	1	0	0	1	0	0	0	0	1	0	1
G15.MUA	0	0	0	1	0	0	0	0	1	1	1
G15.NT@PSMA-AAP	0	0	0	1	0	0	0	0	0	0	1
G15.NT@PSMA-EA	0	0	0	1	0	0	0	0	0	0	1
G15.NT@PSMA-Urea	0	0	0	1	0	0	0	0	0	0	1
G15.PEI-SH	0	0	0	0	0	0	0	0	0	0	0
G15.Phe	0	0	0	0	0	1	0	0	0	0	0
G15.Phe-SH	0	1	0	1	0	0	0	0	0	0	0
G15.Ser-SH	1	0	0	1	0	0	1	0	1	1	1
G15.Thr-SH	0	0	0	1	0	0	0	0	0	0	0
G15.Trp-SH	0	1	0	1	0	0	0	0	0	0	0
G30.AC	0	0	0	0	0	0	1	0	1	1	0
G30.AUT	0	0	0	0	1	0	0	0	0	0	0
G30.CALNN	1	1	0	0	0	1	1	0	0	0	0
G30.DDT@HDA	0	0	0	0	0	0	0	0	0	0	0
G30.Met-SH	1	1	0	1	0	0	1	0	1	0	1
G30.MHDA	0	0	0	0	0	0	0	0	1	1	0
G60.HDA	0	0	0	0	0	0	0	0	0	0	0
G60.MPA	0	0	0	0	0	0	0	0	1	1	1
G60.Phe-SH	1	0	0	0	0	0	0	0	0	0	0
G60.Ser-SH	1	0	0	0	0	0	1	0	0	0	0
G60.SPP	0	0	0	0	0	1	1	0	0	0	0

(*continued from Table Δ′.2*)

	Training samples									
Test samples	G60.NT@PSMA-AP	G30.DDT@BDHDA	G15.F127	G30.DDT@CTAB	G60.Trp-SH	G15.PVA	G15.Met-SH	G60.DDT@DOTAP	G15.DDT@DOTAP	G30.MUTA
G15.Ala-SH	0	0	0	0	0	0	1	0	0	0
G15.Asn-SH	0	0	0	0	0	0	1	0	0	0
G15.AUT	0	0	0	0	0	0	0	0	0	0
G15.CIT	0	0	1	0	0	0	0	0	0	0
G15.MAA	0	0	0	0	0	0	1	0	0	0
G15.MBA	1	0	0	0	0	0	1	0	0	0
G15.MHA	1	0	0	0	0	0	0	0	0	0
G15.MPA	0	0	0	0	0	0	0	0	0	0
G15.MUA	0	0	0	0	0	0	0	0	0	0
G15.NT@PSMA-AAP	0	0	0	0	0	0	1	0	0	0
G15.NT@PSMA-EA	1	0	0	0	0	0	1	0	0	0
G15.NT@PSMA-Urea	0	0	0	0	0	0	1	0	0	0
G15.PEI-SH	0	0	0	0	0	0	0	0	1	0
G15.Phe	0	0	0	0	0	0	0	0	0	0
G15.Phe-SH	0	0	0	0	0	0	1	0	0	0
G15.Ser-SH	1	0	0	0	0	0	1	0	0	0
G15.Thr-SH	0	0	0	0	0	0	1	0	0	0
G15.Trp-SH	0	0	0	0	0	0	1	0	0	0
G30.AC	1	0	0	0	1	0	1	0	0	0
G30.AUT	0	0	0	0	0	0	0	0	0	0
G30.CALNN	0	0	0	0	1	0	1	0	0	0
G30.DDT@HDA	0	0	0	1	0	0	0	0	0	0
G30.Met-SH	1	0	0	0	1	0	1	0	0	0
G30.MHDA	0	0	0	0	0	0	0	0	0	0
G60.HDA	0	0	0	0	0	0	0	0	0	0
G60.MPA	1	0	0	0	0	0	0	0	0	0
G60.Phe-SH	1	0	0	0	1	0	0	0	0	0
G60.Ser-SH	1	0	0	0	1	0	0	0	0	0
G60.SPP	0	0	0	0	1	0	0	0	0	0

Δ′.3 MWCNTs [a] dataset

Table Δ′.3: Neighbours between training and test ENMs of the *MWCNTs [a]* read-across model built using the GA workflow.

	Training samples						
Test samples	1-methylnaphthalene	4-fluorophenol	ethylbenzene	phenethyl alcohol	nitrobenzene	3-bromophenol	methyl 2-methylbenzoate
p-xylene	0	0	1	0	0	0	0
phenol	0	1	0	0	0	0	0
benzonitrile	0	0	0	0	1	0	0
3-methylbenzyl alcohol	0	0	0	1	0	0	0
4-ethylphenol	0	0	0	0	0	0	0
ethyl benzoate	0	0	0	0	0	0	1
3-chlorophenol	0	0	0	0	0	1	0

continued from Table Δ′.3

	Training samples						
Test samples	3,5-dimethylphenol	iodobenzene	4-chloroanisole	acetophenone	bromobenzene	benzyl alcohol	4-chloroacetophenone
p-xylene	0	0	0	1	0	0	0
phenol	0	0	0	0	0	0	0
benzonitrile	0	0	0	0	1	0	0
3-methylbenzyl alcohol	0	0	0	0	0	0	0
4-ethylphenol	1	0	0	0	0	0	0
ethyl benzoate	0	0	0	0	0	0	0
3-chlorophenol	0	0	0	0	0	0	0

continued from Table Δ′.3

	Training samples						
Test samples	3-methylphenol	naphthalene	propylbenzene	4-nitrotoluene	4-chlorotoluene	methyl benzoate	chlorobenzene
p-xylene	0	0	0	0	1	1	0
phenol	0	0	0	0	0	0	0
benzonitrile	0	0	0	0	1	0	1
3-methylbenzyl alcohol	0	0	0	0	0	0	0
4-ethylphenol	0	0	0	0	0	0	0
ethyl benzoate	0	0	1	0	0	0	0
3-chlorophenol	1	0	0	0	0	0	0

Δ′.4 MeOx ENMs [b] dataset

Table Δ′.4: Neighbours between training and test ENMs of the *MeOx ENMs [b]* read-across model built using the GA workflow using $wf_{OF} = 0.05$.

	Training samples														
Test samples	WO_3	Y_2O_3	ZnO	SiO_2	Sb_2O_3	Cr_2O_3	CuO	HfO_2	TiO_2	Co_3O_4	La_2O_3	In_2O_3	Al_2O_3	NiO	Ni_2O_3
Mn_2O_3	0	0	1	0	0	1	1	0	1	1	0	0	0	1	1
CoO	0	0	1	0	0	1	1	0	1	1	0	0	0	1	1
Gd_2O_3	1	0	0	0	1	0	0	1	0	0	1	1	0	0	0
CeO_2	1	0	0	0	1	0	0	1	0	0	1	1	0	0	0
SnO_2	0	1	1	0	1	0	0	0	0	0	1	1	0	0	0
ZrO_2	0	1	1	0	1	1	0	0	1	0	1	1	0	1	1
Fe_2O_3	0	0	1	0	0	1	1	0	1	1	0	0	0	1	1
Yb_2O_3	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0

Appendix E′

Results availability

The results, the data and the developed software of the present Thesis are freely-available to any interested researcher. In this part the readers are provided with a concise list of the results relevant to the present Thesis.

E′.1 Read-across models using the genetic algorithms scheme/Apellis, Chapter 4

https://github.com/DemetraDanae/optimized-read-across
https://apellis.jaqpot.org/
hub.docker.com/r/demetradanae/apellis

E′.2 Grouping/read-across methodology using the feature space/vythos, Chapter 5

https://github.com/NikiKou/MILP-read-across
https://vythos.jaqpot.org/
hub.docker.com/r/demetradanae/vythos

E′.3 *k*NN/read-across models, Chapter 7

http://enaloscloud.novamechanics.com/EnalosWebApps/CNT/
http://enaloscloud.novamechanics.com/EnalosWebApps/ZetaPotential/
https://drive.google.com/drive/folders/1f-Kkm2Jh-wzHpZy_yVQiHNWCZeSxsAwR

Appendix ς΄

Λεξικό όρων στα ελληνικά

» *τι με κοιτάζεις Ρόζα μουδιασμένο*
συγχώρα με που δεν καταλαβαίνω
τι λένε τα κομπιούτερς κι οι αριθμοί

- Άλκης Αλκαίος

Στην ενότητα αυτή της Διατριβής αναφέρονται συνοπτικά οι κυριότερες έννοιες της Εργασίας στα ελληνικά. Θεώρησα σκόπιμο να επεξηγηθούν αυτές οι έννοιες στα ελληνικά αφενός λόγω της αγάπης μου για τη μητρική μου γλώσσα και για την προσωπική μου ανάγκη να κατανοώ επιστημονικές έννοιες σε αυτή. Αφετέρου, ο τρόπος παρουσίασής τους -απλά και χωρίς πολλές λεπτομέρειες- εξυπηρετεί άμεσα την ανάγκη μου να εξηγήσω στην παρέα μου -της οποίας τα περισσότερα μέλη δεν ασχολούνται με την νανοπληροφορική- τι ακριβώς έχω κάνει τα τελευταία χρόνια. Ελπίζω να τα κατάφερα!

ς΄.1 Αλγόριθμος των k πλησιέστερων γειτόνων

Μια από τις συχνά χρησιμοποιούμενες μεθόδους ταξινόμησης και παλινδρόμησης στην παρούσα Διατριβή, είναι ο αλγόριθμος των k πλησιέστερων γειτόνων (k-Nearest Neighbours, kNN) (Κεφ. 7). Ο συγκεκριμένος αλγόριθμος εντάσσεται στις μεθοδολογίες που βασίζονται σε παραδείγματα ή στους «οκνηρούς» ταξινομητές: από το στάδιο της εκπαίδευσης δεν προκύπτει κάποιο γενικευμένο μοντέλο, αντίθετα η ταξινόμηση ή η παλινδρόμηση πραγματοποιείται τη στιγμή που εξετάζεται μια νέα παρατήρηση (νέο δείγμα). Στην ουσία, η μεθοδολογία βασίζεται στην αναλογία μιας παρατήρησης με τις ήδη υπάρχουσες του συνόλου εκπαίδευσης, για το λόγο αυτό αποτελεί μια μεθοδολογία που ανήκει εν δυνάμει στις συγκριτικές μεθόδους read-across. Οι ίδιες οι παρατηρήσεις σε συνδυασμό με τον υπολογισμό της απόστασης χαράσσουν τα όρια που διακρίνουν τις κλάσεις μεταξύ τους, πραγματοποιώντας σχετικά άμεσα μια ομαδοποίηση των δεδομένων.

Αναλυτικότερα, θεωρούμε ένα σύνολο παρατηρήσεων (σύνολο εκπαίδευσης) με N μεταβλητές. Οι τιμές των μεταβλητών αυτών λειτουργούν σαν συντεταγμένες και τοποθετούν τις παρατηρήσεις σε συγκεκριμένα σημεία του πολυδιάστατου χώρου. Όταν προκύψει μια νέα προς εξέταση παρατήρηση, τοποθετείται στο χώρο με βάση τις τιμές των N μεταβλητών και υπολογίζονται οι αποστάσεις της (π.χ. ευκλείδεια απόσταση) από όλες τις παρατηρήσεις του συνόλου εκπαίδευσης. Στη συνέχεια εντοπίζονται οι k πλησιέστερες παρατηρήσεις (γείτονες) του συνόλου εκπαίδευσης και η νέα παρατήρηση καταχωρείται στην πλειοψηφική κλάση μεταξύ των γειτόνων στην περίπτωση της ταξινόμησης, ή λαμβάνει τιμή ίση με τον μέσο όρο των τιμών της υπό εξέταση ιδιότητας (εξαρτημένη μεταβλητή)των γειτόνων στην περίπτωση της παλινδρόμησης. Για να παραχθούν πιο ευαίσθητες προβλέψεις, είναι

σύνηθες οι γείτονες να συμμετέχουν με μεγαλύτερο ή μικρότερο συντελεστή βαρύτητας στην παραγωγή των προβλέψεων ανάλογα με την απόστασή τους από την υπό εξέταση παρατήρηση.

Κατά την εφαρμογή του εν λόγω αλγορίθμου θα πρέπει να ληφθεί αρχικά υπόψιν ότι μεταβλητές που λαμβάνουν τιμές σε διαφορετικά εύρη συμμετέχουν με διαφορετικό βαθμό βαρύτητας στον υπολογισμό των αποστάσεων, χωρίς όμως αυτό να σημαίνει απαραίτητα ότι οι μεταβλητές που λαμβάνουν υψηλότερες τιμές είναι σημαντικότερες από αυτές που λαμβάνουν χαμηλότερες τιμές. Για το λόγο αυτό θα πρέπει αφενός να πραγματοποιηθεί μια επιλογή μεταβλητών πριν τον υπολογισμό των αποστάσεων για να απομακρυνθούν μεταβλητές που προσθέτουν θόρυβο στην ανάλυση και αφετέρου, να εφαρμοστεί μια κανονικοποίηση στις τιμές των επιλεγμένων μεταβλητών, ώστε να συμμετέχουν ισοδύναμα στον υπολογισμό των αποστάσεων.

Η επιλογή της τιμής της παραμέτρου k είναι ένα ζήτημα που πρέπει να επιστήσουμε την προσοχή, καθώς εξαρτάται από αυτή η απόδοση του μοντέλου. Μια σχετικά μικρή της παραμέτρου οδηγεί στην παραγωγή προβλέψεων σχετικά γρήγορα, καθώς συμμετέχουν λιγότερες παρατηρήσεις στην πρόβλεψη. Ωστόσο, ανάλογα με την «πυκνότητα» των παρατηρήσεων στον πολυδιάστατο χώρο, μπορεί να οδηγήσει στην παραγωγή προβλέψεων-εγκλωβισμένων στα τοπικά χαρακτηριστικά των δεδομένων, ενώ μια μεγαλύτερη τιμή της παραμέτρου k μπορεί να λαμβάνει υπόψιν μια γενικότερη συμπεριφορά των δεδομένων που ανήκουν σε μια κατηγορία και να παράγει πιο αξιόπιστα αποτελέσματα (σε συνδυασμό με ένα καλώς ορισμένο πεδίο εφαρμογής). Για τους λόγους αυτούς, η επιλογή της παραμέτρου k πρέπει να ρυθμίζεται προσεκτικά ανάλογα με τα εκάστοτε διαθέσιμα δεδομένα.

ς΄.2 Αλγόριθμος των κατωφλιών

Μια παρεμφερής μεθοδολογία αυτής των k πλησιέστερων γειτόνων, είναι η μεθοδολογία των κατωφλιών που είναι επίσης μια μεθοδολογία που βασίζεται στην αναλογία νέων παρατηρήσεων και παρατηρήσεων που ανήκουν σε ένα σύνολο εκπαίδευσης με γνωστές ιδιότητες. Σε αυτή την περίπτωση όμως η επιλογή των γειτόνων για κάθε νέα παρατήρηση που εισέρχεται στον πολυδιάστατο χώρο δεν πραγματοποιείται με βάση τις k πλησιέστερες παρατηρήσεις, αλλά με βάση ένα κατώφλι (threshold) -μια αριθμητική τιμή μέγιστης απόστασης που επιτρέπεται να απέχουν δυο παρατηρήσεις για να θεωρηθούν γείτονες. Η προσέγγιση αυτή αναλύθηκε διεξοδικά στα Κεφ. 1.3 και 4 της παρούσας Διατριβής.

Η μεθοδολογία των κατωφλιών έχει κοινό σημείο εκκίνησης με τη μεθοδολογία των k πλησιέστερων γειτόνων: ενδείκνυται μια προεπεξεργασία των δεδομένων (επιλογή μεταβλητών ή μείωση των διαστάσεων, κανονικοποίηση) και ακολουθεί μια τοποθέτηση των παρατηρήσεων του συνόλου εκπαίδευσης στον πολυδιάστατο χώρο, με βάση τις συντεταγμένες που ορίζουν οι τιμές των μεταβλητών τους. Για κάθε νέα παρατήρηση που εισέρχεται στον πολυδιάστατο χώρο, υπολογίζονται οι αποστάσεις από όλες τις παρατηρήσεις του συνόλου εκπαίδευσης. Στη συνέχεια με βάση την τιμή ενός κατωφλιού, γείτονες θεωρούνται όσες παρατηρήσεις απέχουν μικρότερη απόσταση από την τιμή του κατωφλιού. Υπάρχει δυνατότητα για χρήση περισσότερων κατωφλιών ομοιότητας και σε αυτή την περίπτωση για να θεωρηθούν γείτονες δυο παρατηρήσεις θα πρέπει να ικανοποιούν τα όρια όλων των κατωφλιών.

Η πρόβλεψη της κλάσης ή της τιμής της ιδιότητας-εξόδου υπολογίζεται με βάση την πλειοψηφία των κλάσεων των γειτόνων ή τον μέσο όρο της ιδιότητας-εξόδου αντίστοιχα, και συχνά οι γείτονες συμμετέχουν στην πρόβλεψη με κάποιο συντελεστή βαρύτητας ανάλογα με την απόστασή τους από την υπό εξέταση παρατήρηση.

ς΄.3 Ασφάλεια από το στάδιο του σχεδιασμού

Ο όρος της ασφάλειας από το στάδιο του σχεδιασμού (safety-by-design) εμπεριέχει την ευρύτερη έννοια της ανάλυσης των κινδύνων σε όλα τα στάδια του κύκλου ζωής ενός προϊόντος, προκειμένου να διασφαλιστούν η δημόσια υγεία, και η ασφάλεια και υγεία στην εργασία. Ειδικότερα στον τομέα της νανοτεχνολογίας τα περιεχόμενα του όρου περιγράφηκαν με ακρίβεια από το ευρωπαϊκό έργο NanoReg2. Η ασφάλεια από το στάδιο του σχεδιασμού περιλαμβάνει την παραγωγή ασφαλών νανοπροϊόντων, την ασφαλή χρήση τους και την ασφαλή βιομηχανική τους παραγωγή.

Η παραγωγή ασφαλών προϊόντων, συνήθως μέρος του τομέα έρευνας και ανάπτυξης (research and development, R&D), περιλαμβάνει το σχεδιασμό λιγότερο επικίνδυνων νανομορφών με βάση τις ιδιότητες τους (π.χ. δομικές, φυσικοχημικές κ.λπ.). Σε αυτό το σκοπό συντρέχουν η ανάπτυξη υπολογιστικών εργαλείων και μεθόδων ομαδοποίησης που δεν απαιτούν παραγωγή και χρήση νανοϋλικών για να εφαρμοστούν, σε συνδυασμό με μικρής κλίμακας *in vivo* και *in vitro* πειραματικές μελέτες που μπορούν να πραγματοποιηθούν σε επιλεγμένες ομάδες νανοϋλικών ή που ήδη υπάρχουν στη Βιβλιογραφία. Η ανάλυση επικινδυνότητας και η προσπάθεια σχεδιασμού ασφαλών νανοπροϊόντων αναφέρεται σε όλο τον κύκλο ζωής του εκάστοτε προϊόντος.

Η ασφάλεια κατά τη χρήση του προϊόντος αποτελεί επίσης μέρος της ασφάλειας από το στάδιο του σχεδιασμού και αναφέρεται στους πιθανούς κινδύνους που προκύπτουν από τη χρήση του προϊόντος για τους καταναλωτές και το περιβάλλον. Από τις αντίστοιχες μελέτες που πραγματοποιούνται, προκύπτουν οδηγίες και περιορισμοί της χρήσης του εκάστοτε προϊόντος.

Τέλος, μέσω μελετών για την ασφάλεια από το στάδιο του σχεδιασμού εξασφαλίζεται ότι οι συνθήκες παραγωγής, αποθήκευσης και μεταφοράς των νανοπροϊόντων θα ικανοποιούν ορισμένες προδιαγραφές βιομηχανικής ασφαλείας, ώστε οι εργαζόμενοι σε όλα τα στάδια της παραγωγικής και εφοδιαστικής αλυσίδας να μην εκτεθούν σε πιθανούς κινδύνους που σχετίζονται με τα παραγόμενα νανοπροϊόντα.

ς΄.4 Γενετικοί αλγόριθμοι

Οι γενετικοί αλγόριθμοι (genetic algorithms) αποτελούν μια υποκατηγορία των εξελικτικών αλγορίθμων (evolutionary algorithms) για την επίλυση προβλημάτων μαθηματικής βελτιστοποίησης. Οι αλγόριθμοι αυτοί είναι ευρετικοί διότι, δεν βασίζονται στην εύρεση μιας αναλυτικής λύσης του υπό εξέταση προβλήματος, αλλά αναζητούν το βέλτιστο συνδυασμό μεταβλητών που ικανοποιούν την αντικειμενική συνάρτηση μέσω μιας διαδικασίας που μιμείται τις εξελικτικές διεργασίες που τελούνται στη φύση (§4.3.1).

Αναλυτικότερα, κάθε πιθανή λύση κωδικοποιείται σε ένα χρωμόσωμα (chromosome) που αποτελείται από γονίδια (genes) τα οποία αντιστοιχίζονται μοναδικά σε μια από τις μεταβλητές του προβλήματος. Η κωδικοποίηση εξαρτάται αφενός από τη φύση του προβλήματος βελτιστοποίησης, αφετέρου από τη μορφή των μεταβλητών που συμμετέχουν στη λύση. Κάθε λύση αξιολογείται ως προς την «ποιότητά» της μέσω μιας συνάρτησης καταλληλότητας (fitness function), η οποία στην ουσία είναι η αντικειμενική συνάρτηση του προβλήματος. Αρχικά δημιουργείται (συνήθως τυχαία) ένας αρχικός πληθυσμός από χρωμοσώματα στα οποία αποδίδεται και η αντίστοιχη τιμή καταλληλότητας. Στη συνέχεια εφαρμόζονται στον πληθυσμό οι τελεστές (operators) της επιλογής (selection), της διασταύρωσης (crossover) και της μετάλλαξης (mutation). Οι γενετικοί τελεστές της διασταύρωσης και της μετάλλαξης συνήθως συνοδεύονται και από μια πιθανότητα που ελέγχει το κατά πόσο θα εφαρμοστούν στα επιλεγμένα χρωμοσώματα.

Με βάση την τιμή της καταλληλότητας επιλέγονται ανά δυο χρωμοσώματα-γονείς (parent chromosomes) τα οποία είναι ικανά να επιζήσουν και να μεταφέρουν τα γονίδιά τους στις

επόμενες γενιές (generations). Τα χρωμοσώματα-γονείς διασταυρώνονται σε τυχαία σημεία ώστε να αναμειχθούν τα γονίδια που περιέχουν και τέλος στα προκύπτοντα χρωμοσώματα εφαρμόζεται η μετάλλαξη, δηλαδή μια μεταβολή στις τιμές από ορισμένα γονίδια, ώστε να εξασφαλιστεί η ποικιλομορφία των χρωμοσωμάτων και να διερευνηθεί πληρέστερα το πεδίο των λύσεων. Η διαδικασία αυτή επαναλαμβάνεται επιλέγοντας χρωμοσώματα από τον αρχικό πληθυσμό και παράγοντας νέα χρωμοσώματα μέχρις ότου να δημιουργηθεί ένας νέος πληθυσμός που περιέχει τόσα χρωμοσώματα όσα και ο αρχικός. Τα χρωμοσώματα του πληθυσμού αξιολογούνται ως προς την καταλληλότητά τους και η όλη διαδικασία επαναλαμβάνεται για έναν αριθμό επαναλήψεων-γενεών, στο τέλος των οποίων έχουν επιζήσει τα χρωμοσώματα με τα καλύτερα χαρακτηριστικά, δηλαδή αυτά που ικανοποιούν καλύτερα την αντικειμενική συνάρτηση. Στη Βιβλιογραφία, απαντώνται εναλλακτικοί τρόποι επιλογής, διασταύρωσης και μετάλλαξης των λύσεων, όπως και εναλλακτικά κριτήρια τερματισμού.

Οι γενετικοί αλγόριθμοι είναι κατάλληλοι για την επίλυση πολύπλοκων προβλημάτων όπου η επίλυση με αναλυτικό τρόπο είναι δύσκολη έως αδύνατη. Ωστόσο, λόγω του στοχαστικού χαρακτήρα τους (τυχαιότητα) πολλές φορές συγκλίνουν σε διαφορετικές λύσεις, που μάλιστα δεν είναι απαραίτητα οι βέλτιστες (παγκόσμιο άριστο). Οι γενετικοί αλγόριθμοι δηλαδή είναι επιρρεπείς στο να εγκλωβιστούν σε τοπικά βέλτιστα αντί να κατευθυνθούν προς τη γενική βέλτιστη λύση. Τέλος ανάλογα με το μέγεθος του προβλήματος και των γενετικών τελεστών, η εξεύρεση της λύσης, μπορεί να γίνει μια πολύ χρονοβόρα διαδικασία.

ς΄.5 Επιλογή μεταβλητών

Μια από τις πιο σημαντικές διεργασίες που συναντήσαμε σε όλα τα κομμάτια της Διατριβής, ήταν η επιλογή ενός υποσυνόλου από ένα σύνολο διαθέσιμων μεταβλητών (variable/feature selection) για την ανάπτυξη ενός μοντέλου πρόβλεψης. Η διεργασία αυτή, εμφανίστηκε είτε ως ανεξάρτητη δραστηριότητα και στάδιο προεπεξεργασίας των δεδομένων πριν την μοντελοποίηση (Κεφ. 7), είτε ως αναπόσπαστο μέρος μιας μεθοδολογίας πρόβλεψης (Κεφ. 4, 5 και 6).

Μέχρι σήμερα έχουν αναπτυχθεί και ελεγχθεί ως προς την ορθότητά τους πολλές μέθοδοι επιλογής μεταβλητών, μερικές από τις οποίες μάλιστα εξαρτώνται και από το είδος της επακόλουθης μοντελοποίησης. Ο αναγνώστης μπορεί να αναζητήσει μερικές από αυτές -αν δεν έχουν αναφερθεί στο κυρίως μέρος της Διατριβής- στο σύγγραμμα των I. H. Witten, E. Frank, M. A. Hall και C. J. Pal (Morgan Kaufmann, 2016) ή στο διαδίκτυο με μια απλή αναζήτηση.

Ανεξάρτητα από τη χρησιμοποιούμενη μέθοδο, η επιλογή μεταβλητών εξασφαλίζει την απομάκρυνση μεταβλητών που προσθέτουν «θόρυβο» στο παραγόμενο μοντέλο, και κατά συνέπεια οι αλγόριθμοι γίνονται αποδοτικότεροι ως προς την ποιότητα των προβλέψεων που παράγουν (αποφυγή υπερπροσασμογής του μοντέλου στα δεδομένα εκπαίδευσης), ενώ το στάδιο της εκπαίδευσής τους διαρκεί λιγότερο χρόνο. Το αποτέλεσμα αυτό είναι εμφανές, στις περιπτώσεις που χρησιμοποιούνται «μεγάλα» σύνολα δεδομένων. Πέρα όμως από την εκπαίδευση, κατά τη χρήση ενός μοντέλου που απαιτεί λίγες μεταβλητές χρειάζεται πρακτικά να γίνει συνοπτικότερη αναζήτηση είτε βιβλιογραφική είτε πειραματική σε ιδιότητες, επομένως επιτυγχάνεται εξοικονόμηση χρόνου και πόρων. Τέλος, μοντέλα με λιγότερες μεταβλητές από τις διαθέσιμες είναι περισσότερο κατανοητά από τους ερευνητές ή τους αντίστοιχους χρήστες τους.

ς΄.6 Μαθηματική αριστοποίηση

Ο όρος της μαθηματικής αριστοποίησης ή βελτιστοποίησης (optimisation) ή του μαθηματικού προγραμματισμού (mathematical programming) αποτελεί ένα από τα σημαντικότερα εργαλεία λήψης αποφάσεων και αναφέρεται στην αναζήτηση ενός συνδυασμού παραμέτρων (λύση ενός προβλήματος) από τις διάφορες εναλλακτικές ή πιθανές λύσεις που ελαχιστοποιούν ή μεγιστοποιούν την τιμή μιας αντικειμενικής συνάρτησης (objective function) και ταυτόχρονα ικανοποιούν μια σειρά περιορισμών.

Ανάλογα με τη διατύπωση του προβλήματος με μαθηματικούς όρους και το είδος των μεταβλητών που συμμετέχουν σε αυτό διακρίνονται διάφορες οικογένειες προβλημάτων μαθηματικού προγραμματισμού όπως τα προβλήματα γραμμικού προγραμματισμού, μη-γραμμικού προγραμματισμού, ακεραίου γραμμικού προγραμματισμού, μικτού ακεραίου γραμμικού προγραμματισμού κ.λπ. και συνοδεύονται και από τις αντίστοιχες μεθόδους επίλυσής τους.

ς΄.7 Μοντέλα τύπου (Q)SAR

Η εύρεση συσχετίσεων, υπό την μορφή υπολογιστικών μοντέλων, που να συνδέουν την μοριακή δομή των χημικών ενώσεων με την παρατηρούμενη δραστικότητά τους ή με άλλες ιδιότητες, θεωρείται ως μια αξιόπιστη εναλλακτική οδός για την εκτίμηση της επικινδυνότητας των ουσιών[1]. Τέτοια μοντέλα αναφέρονται συχνά ως ποιοτικά/ποσοτικά μοντέλα συσχέτισης δομής-ιδιοτήτων/δραστικότητας (Qualitative/Quantitative Structure–Property/Activity Relationships, QSPR ή QSAR αντίστοιχα) και αναπτύσσονται βάσει τεχνικών της στατιστικής ή της εξόρυξης δεδομένων (data mining).

Τα μοντέλα αυτά επιτρέπουν, πέρα από την πρόβλεψη ιδιοτήτων, να κατανοήσουμε την επίδραση ορισμένων δομικών ιδιοτήτων στην «συμπεριφορά» της ουσίας όταν εισέλθει σε ένα βιολογικό περιβάλλον. Η μοριακή δομή μιας ουσίας κωδικοποιείται στις τιμές συγκεκριμένων μεταβλητών οι οποίες ονομάζονται «μοριακοί περιγραφείς» (molecular descriptors). Η ιδιότητα ενδιαφέροντος μπορεί να έχει τη μορφή μιας αριθμητικής τιμής (ποσοτικά μοντέλα) ή τη μορφή μιας κατηγορικής τιμής (ποιοτικά μοντέλα). Τα μοντέλα αυτά εκπαιδεύονται χρησιμοποιώντας ήδη γνωστά δεδομένα και αξιολογούνται ως προς την ευρωστία τους, ενώ παράλληλα καθορίζεται και το πεδίο εφαρμογής τους ώστε να εξασφαλιστεί η παραγωγή αξιόπιστων προβλέψεων υπό πραγματικές συνθήκες.

Μοντέλα τύπου (Q)SAR χρησιμοποιούνται επιτυχώς στον τομέα της Χημειοπληροφορικής για την ανάλυση της επικινδυνότητας χημικών ουσιών που χρησιμοποιούνται σε καταναλωτικά προϊόντα, για την οργάνωση της προτεραιότητας της πειραματικής μελέτης ουσιών ή για την εύρεση δραστικών ουσιών για την ανάπτυξη νέων φαρμάκων. Λόγω της επιτυχίας αυτής, ορισμένες ερευνητικές ομάδες πρότειναν τη μετάβαση της χρήσης των μοντέλων τύπου (Q)SAR στον τομέα της Νάνοπληροφορικής με τη δημιουργία μοντέλων nano(Q)SAR ή (Q)NAR (Quantitative Nanostructure-Activity Relationships, ποιοτικά/ποσοτικά μοντέλα συσχέτισης νανοδομής-δραστικότητας) που να συσχετίζουν τις φυσικοχημικές ιδιότητες των νανοϋλικών με τη βιολογική τους δραστηριότητα. Ωστόσο η μετάβαση αυτή δεν είναι τόσο άμεση και αυτονόητη, καθώς πρέπει να ληφθεί υπόψιν μια σειρά περιορισμών που προκύπτουν από τη φύση των υλικών σε νανοκλίμακα.

Σε αντίθεση με τις χημικές ουσίες, η μέτρηση των φυσικοχημικών ιδιοτήτων των νανοϋλικών δεν είναι άμεση και ακριβής, λόγω της ετερογενούς φύσης τους. Κατά συνέπεια τα πειραματικά δεδομένα ενδέχεται να περιέχουν σημαντικά ποσοστά σφαλμάτων και να είναι ακατάλληλα για μοντελοποίηση. Στη δημιουργία σφαλμάτων συντρέχει και το ίδιο το

[1]Ο όρος «ουσίες» αναφέρεται τόσο σε χημικές ουσίες όσο και σε νανοϋλικα, καθώς αυτή είναι και η ορολογία που χρησιμοποιείται στους ευρωπαϊκούς κανονισμούς.

βιολογικό μέσον στο οποίο γίνεται η μέτρηση, π.χ. με τη δημιουργία συσσωματωμάτων (agglomerates). Επίσης τα ανεπαρκή σύνολα πειραματικών δεδομένων (έλλειψη πειραματικών μελετών τοξικότητας, αδυναμία συγχώνευσης δεδομένων που αφορούν σε διαφορετικούς τύπους νανοϋλικών, π.χ. μεταλλικά οξείδια, νανοσωλήνες άνθρακα κ.λπ. λόγω διαφορετικών μηχανισμών τοξικότητας που ακολουθούν) δεν επιτρέπουν την επιλογή αντιπροσωπευτικών και «αρκούντως μεγάλων» συνόλων εκπαίδευσης και ελέγχου για την εφαρμογή εξωτερικής επικύρωσης των μοντέλων, με κίνδυνο να αναπτύσσονται μοντέλα υπερπροσαρμοσμένα στα δεδομένα (over-fitting).

ς΄.8 Πεδίο εφαρμογής μοντέλων

Θεωρείται προφανές ότι ένα οποιοδήποτε μοντέλο τύπου SAR που έχει αναπτυχθεί για την πρόβλεψη μιας ιδιότητας είναι αδύνατο να προβλέψει με αξιοπιστία την ιδιότητα αυτή για οποιαδήποτε δεδομένα εισόδου. Προκειμένου λοιπόν να ενισχυθεί η εμπιστοσύνη του τελικού χρήστη στις προβλέψεις του μοντέλου πρέπει να οριστεί το πεδίο εφαρμογής του (domain of applicability, §2.4), το οποίο θα επιτρέψει να γίνεται χρήση του σε πραγματικές συνθήκες.

Το πεδίο εφαρμογής ενός μοντέλου εντοπίζεται στον πολυδιάστατο χώρο που οριοθετείται από τις ιδιότητες των δεδομένων εκπαίδευσης που χρησιμοποιήθηκαν για την ανάπτυξή του. Όταν εξετάζεται μια νέα ουσία, ερευνάται η σχετική της θέση στον χώρο που ορίζουν τα δεδομένα εκπαίδευσης και στην περίπτωση που η ουσία βρίσκεται εντός του πεδίου εφαρμογής, η πρόβλεψη που παράγεται από το μοντέλο θεωρείται αξιόπιστη, ενώ σε αντίθετη περίπτωση, θεωρείται μη αξιόπιστη. Στη Βιβλιογραφία απαντώνται πολλοί τρόποι καθορισμού του πεδίου εφαρμογής ενός μοντέλου (μέθοδος της μόχλευσης (leverage), μέθοδοι υπολογισμού αποστάσεων ή ομοιοτήτων κ.α.).

ς΄.9 Πλαίσιο συγκριτικών μεθόδων read-across

Ένα μεγάλο μέρος της Διατριβής αφιερώνεται στην ανάπτυξη μεθοδολογιών πρόβλεψης ανεπιθύμητων ιδιοτήτων νανοϋλικών που βασίζονται στο πλαίσιο read-across (§1.3). Το πλαίσιο αυτό καθορίστηκε από τον Ευρωπαϊκό Οργανισμό Χημικών Προϊόντων (European Chemicals Agency, ECHA) μέσω του «Read-Across Assessment Framework» που υπόκειται στο γενικότερο θεσμικό πλαίσιο του Κανονισμού για την Καταχώριση, Αξιολόγηση, Αδειοδότηση και τους Περιορισμούς των Χημικών Προϊόντων (Registration, Evaluation, Authorisation and Restriction of Chemicals, REACH).

Η κεντρική ιδέα του πλαισίου αυτού συνοψίζεται στο ότι η εκτίμηση ιδιοτήτων για μια ή περισσότερες ουσίες - «στόχους» (χημικές ουσίες ή νανοϋλικα) μπορεί να προκύψει από σχετική πληροφορία «ανάλογων» ουσιών. Ο όρος read-across, και όλες οι μεθοδολογίες που εντάσσονται σε αυτό το πλαίσιο, αποδίδεται πιο πιστά στα ελληνικά ως «συγκριτική προσέγγιση», ακριβώς επειδή οι προσεγγίσεις αυτές βασίζονται στην εξεύρεση σχετικών-παρόμοιων ουσιών, για κάθε υπό εξέταση ουσία. Ο όρος της συγκριτικής προσέγγισης συνοδεύεται συχνά και από τον όρο της ομαδοποίησης (grouping), καθώς συχνά η ομοιότητα των ουσιών οδηγεί στην δημιουργία ομάδων ουσιών και αντίστροφα· η ομοιότητα διαφόρων ουσιών αναζητείται στα πλαίσια γνωστών ομάδων.

Στο πλαίσιο των συγκριτικών μεθόδων υπάρχουν δυο κύριες προσεγγίσεις, η «προσέγγιση των αναλόγων» (analogue approach) και η «προσέγγιση των κατηγοριών» (category approach). Η διατύπωση των δυο προσεγγίσεων βασίζεται στο πλήθος των διαθέσιμων δεδομένων συνεπώς, τα όρια μεταξύ τους δεν είναι πάντοτε σαφή.

Στην προσέγγιση των αναλόγων η πρόβλεψη των ιδιοτήτων μιας ουσίας περιορίζεται σε μια μικρή περιοχή του πεδίου των διαθέσιμων δεδομένων κυρίως λόγω έλλειψης επαρκών

πληροφοριών. Πιο συγκεκριμένα μια ή περισσότερες ουσίες με γνωστές φυσικοχημικές, τοξικολογικές ή οικο-τοξικολογικές ιδιότητες (πηγές)[2] μπορούν να χρησιμοποιηθούν για την πρόβλεψη των ιδιοτήτων μιας ή περισσοτέρων υπό εξέταση ουσιών (στόχοι). Οι ουσίες πηγές και στόχοι έχουν παρεμφερείς δομικές ιδιότητες ωστόσο, δεν παρουσιάζονται εμφανείς τάσεις στις ιδιότητές τους, λόγω του μικρού αριθμού των διαθέσιμων δεδομένων. Για κάθε υπό εξέταση ουσία η εκτίμηση μιας ιδιότητας προκύπτει από τις παρόμοιες ουσίες-πηγές, εφαρμόζοντας τοπικά είτε μια απλή μεθοδολογία παλινδρόμησης είτε κάποιο πολυπλοκότερο μοντέλο. Στην χειρότερη περίπτωση, για μια ουσία-στόχο θα βρεθεί μόνο μια ουσία-πηγή, οπότε η εκτίμηση της υπό εξέταση ιδιότητας για την ουσία-στόχο θα συμπίπτει με την τιμή της ιδιότητας της ουσίας-πηγής.

Στην προσέγγιση των κατηγοριών, οι ουσίες οργανώνονται σε ομάδες βάσει των δομικών ιδιοτήτων τους και των επιτρεπτών τους διαφορών. Σε αυτή την περίπτωση οι ιδιότητες ενδιαφέροντος που αποτελούν αντικείμενο της πρόβλεψης και χαρακτηρίζουν μια κατηγορία, χαρακτηρίζουν και κάθε ουσία που ανήκει στην ίδια κατηγορία. Στην χειρότερη περίπτωση, η ισχύς της προβλεπόμενης ιδιότητας για μια υπό εξέταση ουσία θα είναι στην πραγματικότητα χαμηλότερη από την ισχύ της ίδιας ιδιότητας που χαρακτηρίζει την ομάδα. Η δομική ομοιότητα των ουσιών μπορεί να οφείλεται στην ύπαρξη κοινών λειτουργικών ομάδων, κοινών συστατικών, κοινών προδρόμων ουσιών ή ουσιών αποδόμησης, στη σταθερή τάση μεταβολής του μεγέθους των ιδιοτήτων κ.α. ή σε συνδυασμό των παραπάνω ωστόσο, πολλές φορές οι δομικές ιδιότητες από μόνες τους δεν επαρκούν για την αιτιολόγηση μιας ομαδοποίησης. Πλην των δομικών ομοιοτήτων, σε πολλές μελέτες αναζητούνται εναλλακτικοί τρόποι ομαδοποίησης χρησιμοποιώντας για παράδειγμα ανάλυση κυρίων συνιστωσών (principal component analysis, PCA), μεθόδους ιεραρχικής συσταδοποίησης (hierarchical clustering), αποστάσεις (π.χ. ευκλείδεια απόσταση), ακόμη και μελετώντας τους μηχανισμούς δράσης (mode-of-action) των ουσιών. Τέλος στα πλαίσια των σχηματιζόμενων ομάδων, είναι σύνηθες να εφαρμόζονται τεχνικές πρόβλεψης που ανήκουν στην προσέγγιση των αναλόγων, για την παραγωγή πιο αξιόπιστων εκτιμήσεων.

Τέλος οφείλουμε να σημειώσουμε ότι μερικές από τις μεθοδολογίες read-across που παρουσιάζονται στην παρούσα Διατριβή, αναπτύχθηκαν με γνώμονα τα δεδομένα και δεν επικεντρώθηκαν σε συγκεκριμένες ιδιότητες-εξόδους (Κεφ. 4, 5 και 6). Ο στόχος μας ήταν να βρεθούν αυτόματες και γενικευμένες διαδικασίες ομαδοποίησης και πρόβλεψης, που να μην εξαρτώνται από το είδος των ουσιών, από τα δομικά τους χαρακτηριστικά ως έχουν και, από την ιδιότητα προς πρόβλεψη, αλλά μόνο από μετρήσιμες ιδιότητες ή χαρακτηριστικά που μπορούσαν να κωδικοποιηθούν σε μια αριθμητική ή κατηγορική μεταβλητή. Οι μεθοδολογίες που προέκυψαν αποτελούν «υβρίδια» των δυο προαναφερθέντων προσεγγίσεων αφού επιτυγχάνουν την ομαδοποίηση ενός συνόλου ουσιών με γνωστές ιδιότητες, την κατάταξη κάθε νέας υπό εξέτασης ουσίας στις σχηματιζόμενες ομάδες και την παραγωγή προβλέψεων ξεχωριστά για κάθε νέα ουσία στα πλαίσια των ομάδων.

ς΄.10 Προεπεξεργασία δεδομένων

Η εξαγωγή αξιόπιστων προβλέψεων από δεδομένα, προϋποθέτει ότι τα δεδομένα που χρησιμοποιούνται εξαρχής για να αναπτυχθεί ένα κατάλληλο μοντέλο, να είναι «ποιοτικά», δηλαδή όσο το δυνατόν να έχουν συλλεχθεί επιμελώς βάσει πρωτοκόλλων ή κατάλληλου σχεδιασμού πειραμάτων και να περιέχουν ικανή «ποσότητα» χρήσιμης πληροφορίας για να βασιστεί σε αυτά η ανοικοδόμηση ενός αξιόπιστου μοντέλου. Ιδιαίτερα στον τομέα της Νανοπληροφορικής όπου μεγάλα σύνολα δεδομένων είναι σχετικά δυσεύρετα, είναι επιτακτική η ανάγκη συνεπούς συλλογής αξιόπιστων δεδομένων και προεπεξεργασίας

[2]Οι ουσίες-πηγές στο κυρίως μέρος της Διατριβής αναφέρονται συχνότερα ως «γείτονες» (neighbours) για λόγους συνέπειας με την ορολογία των αναπτυσσόμενων αλγορίθμων πρόβλεψης.

τους (data preprocessing ή data curation) πριν την εφαρμογή οποιασδήποτε μεθοδολογίας μοντελοποίησης (§2.1).

Αρχικά κατά την συλλογή των δεδομένων, εάν προέρχονται από διαφορετικές βιβλιογραφικές πηγές, είναι απαραίτητο να ελεγχθεί ότι οι καταχωρήσεις (τιμές) για κάθε μεταβλητή προέκυψαν ακολουθώντας ακριβώς την ίδια πειραματική μέθοδο υπό τις ίδιες συνθήκες ή έχουν υπολογιστεί με τον ίδιο τρόπο. Στη συνέχεια το σύνολο δεδομένων που έχει διαμορφωθεί, ελέγχεται για την ύπαρξη κενών τιμών (μεταβλητές που δεν έχουν συμπληρωθεί για όλα τα δείγματα του συνόλου). Ανάλογα με το είδος της μοντελοποίησης, με το είδους του προβλήματος και με την ευχέρεια των αναλυτών να «θυσιάσουν» μεταβλητές, οι κενές τιμές μπορούν να συμπληρωθούν με κάποια συγκριτική μεθοδολογία ή δραστικότερα, οι μεταβλητές που περιέχουν τις κενές τιμές δύναται να απομακρυνθούν.

Με τον ίδιο τρόπο, κατάφωρα ακραίες τιμές (outliers) που περιέχονται σε μεταβλητές επίσης απομακρύνονται, καθώς μπορούν να «παραπλανήσουν» το μοντέλο ώστε να παράγει μη αξιόπιστες προβλέψεις. Ακόμα, η ύπαρξη διπλότυπων καταχωρήσεων (δείγματα με ακριβώς τις ίδιες τιμές σε όλες τις ιδιότητες) μπορεί να λειτουργήσει μεροληπτικά ως προς συγκεκριμένες προβλέψεις συνεπώς θα πρέπει και οι αντίστοιχες καταχωρίσεις να απομακρύνονται. Φυσικά ο καθαρισμός των δεδομένων δεν περιορίζεται στα παραπάνω. Μάλιστα, πέρα από τα τετριμμένα σημεία που εξετάζονται στο στάδιο της προεπεξεργασίας των δεδομένων, μπορεί να πραγματοποιηθεί και χειροκίνητο φιλτράρισμα το οποίο εξαρτάται σε μεγάλο βαθμό από την εμπειρία του αναλυτή.

Τέλος, εκτός από τον καθαρισμό, η προεπεξεργασία των δεδομένων μπορεί να περιλαμβάνει διαδικασίες κανονικοποίησης (standarisation) ή ομαλοποίησης (normalisation) ώστε στη συνέχεια όλες οι μεταβλητές να συμμετέχουν ισοδύναμα στην μοντελοποίηση. Επίσης οι κατηγορικές μεταβλητές με περισσότερες από μια διακριτές τιμές -ανάλογα με το είδος της επακόλουθης μοντελοποίησης- ενδέχεται να χρειαστεί να χωριστούν σε επιμέρους μεταβλητές, μια για κάθε διακριτή τιμή, οι οποίες με τη σειρά τους παίρνουν δυαδικές τιμές.

Index

Bibliography

[1] D.-D. Varsou, A. Afantitis, A. Tsoumanis, G. Melagraki, H. Sarimveis, E. Valsami-Jones, and I. Lynch, "A safe-by-design tool for functionalised nanomaterials through the enalos nanoinformatics cloud platform", *Nanoscale Advances*, vol. 1, no. 2, pp. 706–718, 2019.

[2] D.-D. Varsou, A. Afantitis, G. Melagraki, and H. Sarimveis, "Read-across predictions of nanoparticle hazard endpoints: A mathematical optimization approach", *Nanoscale Advances*, vol. 1, no. 9, pp. 3485–3498, 2019.

[3] D.-D. Varsou, A. Afantitis, A. Tsoumanis, A. Papadiamantis, E. Valsami-Jones, I. Lynch, and G. Melagraki, "Zeta-potential read-across model utilizing nanodescriptors extracted via the nanoxtract image analysis tool available on the enalos nanoinformatics cloud platform", *Small*, p. 1 906 588, 2020.

[4] D.-D. Varsou and H. Sarimveis, "Apellis: An online tool for read-across model development", *Computational Toxicology*, p. 100 146, 2020.

[5] D.-D. Varsou, N.-M. Koutroumpa, and H. Sarimveis, "Automated grouping of nanomaterials and read-across prediction of their adverse effects based on mathematical optimization", *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 2766–2779, 2021.

[6] A. R. Leach and V. J. Gillet, *An introduction to chemoinformatics*. Springer Science & Business Media, 2007.

[7] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm", *npj Computational Materials*, vol. 6, no. 1, pp. 1–10, 2020.

[8] F. Fontana, P. Figueiredo, J. P. Martins, and H. A. Santos, "Requirements for animal experiments: Problems and challenges", *Small*, p. 2 004 182, 2020.

[9] S. Chibani and F.-X. Coudert, "Machine learning approaches for the prediction of materials properties", *APL Materials*, vol. 8, no. 8, p. 080 701, 2020.

[10] C. C. Hsu, M. J. Buehler, and A. Tarakanova, "The order-disorder continuum: Linking predictions of protein structure and disorder through molecular simulation", *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[11] S. Piana, P. Robustelli, D. Tan, S. Chen, and D. E. Shaw, "Development of a force field for the simulation of single-chain proteins and protein–protein complexes", *Journal of chemical theory and computation*, vol. 16, no. 4, pp. 2494–2507, 2020.

[12] P. D. Kolokathis and O. M. Braun, "Kobra: A rate constant method for prediction of the diffusion of sorbates inside nanoporous materials at different loadings", *Journal of computational chemistry*, vol. 40, no. 23, pp. 2053–2066, 2019.

[13] European Commission. (2011). Definition of a nanomaterial, [Online]. Available: `https://ec.europa.eu/environment/chemicals/nanotech/faq/definition_en.htm`. (accessed on: May 07, 2021).

[14] J. Jeevanandam, A. Barhoum, Y. S. Chan, A. Dufresne, and M. K. Danquah, "Review on nanoparticles and nanostructured materials: History, sources, toxicity and regulations", *Beilstein journal of nanotechnology*, vol. 9, no. 1, pp. 1050–1074, 2018.

[15] A. Ahamed, L. Liang, M. Y. Lee, J. Bobacka, and G. Lisak, "Too small to matter? physicochemical transformation and toxicity of engineered ntio2, nsio2, nzno, carbon nanotubes, and nag", *Journal of Hazardous Materials*, p. 124 107, 2020.

[16] W. Yang, L. Wang, E. M. Mettenbrink, P. L. DeAngelis, and S. Wilhelm, "Nanoparticle toxicology", *Annual Review of Pharmacology and Toxicology*, vol. 61, 2020.

[17] S. Halappanavar, P. Nymark, H. F. Krug, M. J. Clift, B. Rothen-Rutishauser, and U. Vogel, "Non-animal strategies for toxicity assessment of nanoscale materials: Role of adverse outcome pathways in the selection of endpoints", *Small*, p. 2 007 628, 2021.

[18] E. Valsami-Jones and I. Lynch, "How safe are nanomaterials?", *Science*, vol. 350, no. 6259, pp. 388–389, 2015.

[19] N. Sizochenko, A. Mikolajczyk, K. Jagiello, T. Puzyn, J. Leszczynski, and B. Rasulev, "How the toxicity of nanomaterials towards different species could be simultaneously evaluated: A novel multi-nano-read-across approach", *Nanoscale*, vol. 10, no. 2, pp. 582–591, 2018.

[20] A. Giusti, R. Atluri, R. Tsekovska, A. Gajewicz, M. D. Apostolova, C. L. Battistelli, E. A. Bleeker, C. Bossa, J. Bouillard, M. Dusinska, P. Gómez-Fernández, R. Grafström, M. Gromelski, Y. Handzhiyski, N. R. Jacobsen, P. Jantunen, K. A. Jensen, A. Mech, J. M. Navas, P. Nymark, A. G. Oomen, T. Puzyn, K. Rasmussen, C. Riebeling, I. Rodriguez-Llopis, S. Sabella, J. R. Sintes, B. Suarez-Merino, S. Tanasescu, H. Wallin, and A. Haase, "Nanomaterial grouping: Existing approaches and future recommendations", *NanoImpact*, p. 100 182, 2019.

[21] J. Njuguna, K. Pielichowski, and H. Zhu, *Health and environmental safety of nanomaterials: polymer nancomposites and other materials containing nanoparticles*. Elsevier, 2014.

[22] R. A. Yokel and R. C. MacPhail, "Engineered nanomaterials: Exposures, hazards, and risk prevention", *Journal of Occupational Medicine and Toxicology*, vol. 6, no. 1, pp. 1–27, 2011.

[23] C. Buzea, I. I. Pacheco, and K. Robbie, "Nanomaterials and nanoparticles: Sources and toxicity", *Biointerphases*, vol. 2, no. 4, MR17–MR71, 2007.

[24] T. X. Trinh, M. K. Ha, J. S. Choi, H. G. Byun, and T. H. Yoon, "Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles", *Environmental Science: Nano*, vol. 5, no. 8, pp. 1902–1910, 2018. DOI: 10.1039/c8en00061a.

[25] J. Y. Choi, G. Ramachandran, and M. Kandlikar, "The impact of toxicity testing costs on nanomaterial regulation", *Environmental Science and Technology*, vol. 43, no. 9, pp. 3030–3034, 2009, ISSN: 0013936X. DOI: 10.1021/es802388s.

[26] T. Hartung, "From alternative methods to a new toxicology", *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 77, no. 3, pp. 338–349, 2011. DOI: 10.1016/j.ejpb.2010.12.027.

[27] European Chemicals Agency, *Regulation (ec) no 1907/2006 of the european parliament and of the council of 18 december 2006*, ECHA, 2006. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02006R1907-20161011&from=EN, (accessed on: April 27, 2021).

[28] EU Directive, "63/eu of the european parliament and of the council of 22 september 2010 on the protection of animals used for scientific purposes", *Official Journal of the European Union*, vol. 276, pp. 33–74, 2010.

[29] D. A. Winkler, E. Mombelli, A. Pietroiusti, L. Tran, A. Worth, B. Fadeel, and M. J. McCall, "Applying quantitative structure–activity relationship approaches to nanotoxicology: Current status and future potential", *Toxicology*, vol. 313, no. 1, pp. 15–23, 2013.

[30] J. J. Villaverde, B. Sevilla-Morán, C. López-Goti, J. L. Alonso-Prados, and P. Sandín-España, "Considerations of nano-qsar/qspr models for nanopesticide risk assessment within the european legislative framework", *Science of The Total Environment*, vol. 634, pp. 1530–1539, 2018.

[31] L. Lamon, D. Asturiol, A. Vilchez, R. Ruperez-Illescas, J. Cabellos, A. Richarz, and A. Worth, "Computational models for the assessment of manufactured nanomaterials: Development of model reporting standards and mapping of the model landscape", *Computational Toxicology*, 2018.

[32] EU Science Hub, *Review of computational models for the safety assessment of nanomaterials*, 2017. [Online]. Available: `https://ec.europa.eu/jrc/en/science-update/review-computational-models-safety-assessment-nanomaterials`, (accessed on: January 08, 2019).

[33] A. Gajewicz, K. Jagiello, M. T. Cronin, J. Leszczynski, and T. Puzyn, "Addressing a bottle neck for regulation of nanomaterials: Quantitative read-across (nano-qra) algorithm for cases when only limited data is available", *Environmental Science: Nano*, vol. 4, no. 2, pp. 346–358, 2017.

[34] A. Gajewicz, "Development of valuable predictive read-across models based on "real-life"(sparse) nanotoxicity data", *Environmental Science: Nano*, vol. 4, no. 6, pp. 1389–1403, 2017.

[35] A. Gajewicz, M. T. Cronin, B. Rasulev, J. Leszczynski, and T. Puzyn, "Novel approach for efficient predictions properties of large pool of nanomaterials based on limited set of species: Nano-read-across", *Nanotechnology*, vol. 26, no. 1, p. 015 701, 2014.

[36] A. G. Oomen, E. A. Bleeker, P. M. Bos, F. van Broekhuizen, S. Gottardo, M. Groenewold, D. Hristozov, K. Hund-Rinke, M.-A. Irfan, and A. Marcomini, "Grouping and read-across approaches for risk assessment of nanomaterials", *International journal of environmental research and public health*, vol. 12, no. 10, pp. 13 415–13 434, 2015.

[37] L. Lamon, D. Asturiol, A. Richarz, E. Joossens, R. Graepel, K. Aschberger, and A. Worth, "Grouping of nanomaterials to read-across hazard endpoints: From data collection to assessment of the grouping hypothesis by application of chemoinformatic techniques", *Particle and fibre toxicology*, vol. 15, no. 1, pp. 1–17, 2018.

[38] European Chemicals Agency, *Read-across assessment framework (raaf)*, 2017. [Online]. Available: `https://op.europa.eu/en/publication-detail/-/publication/841c5a3a-2981-11e7-ab65-01aa75ed71a1`, (accessed on: November 24, 2020).

[39] R. Benigni, C. Laura Battistelli, C. Bossa, A. Giuliani, E. Fioravanzo, A. Bassan, M. Fuart Gatnik, J. Rathman, C. Yang, and O. Tcheremenskaia, "Evaluation of the applicability of existing (q) sar models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across", *EFSA Supporting Publications*, vol. 16, no. 3, 1598E, 2019.

[40] Y. K. Koleva, J. C. Madden, and M. T. Cronin, "Formation of categories from structure- activity relationships to allow read-across for risk assessment: Toxicity of α, β-unsaturated carbonyl compounds", *Chemical research in toxicology*, vol. 21, no. 12, pp. 2300–2312, 2008.

[41] A. Mech, K. Rasmussen, P. Jantunen, L. Aicher, M. Alessandrelli, U. Bernauer, E. A. J. Bleeker, J. Bouillard, P. D. P. Fanghella, R. Draisci, M. Dusinska, G. Encheva, G. Flament, A. Haase, Y. Handzhiyski, F. Herzberg, J. Huwyler, N. R. Jacobsen, V. Jeliazkov, N. Jeliazkova, P. Nymark, R. Grafström, A. G. Oomen, M. L. Polci, C. Riebeling, J. Sandström, B. Shivachev, S. Stateva, S. Tanasescu, R. Tsekovska, H. Wallin, M. F. Wilks, S. Zellmer, and M. D. Apostolova, "Insights into possibilities for grouping and read-across for nanomaterials in eu chemicals legislation", *Nanotoxicology*, vol. 13, no. 1, pp. 119–141, 2019.

[42] C. M. Sayes, P. A. Smith, and I. V. Ivanov, "A framework for grouping nanoparticles based on their measurable characteristics", *International Journal of Nanomedicine*, vol. 8, no. Suppl 1, p. 45, 2013.

[43] C. Helma, M. Rautenberg, and D. Gebele, "Nano-lazar: Read across predictions for nanoparticle toxicities with calculated and measured properties", *Frontiers in pharmacology*, vol. 8, p. 377, 2017.

[44] S. E. Escher, H. Kamp, S. H. Bennekou, A. Bitsch, C. Fisher, R. Graepel, J. G. Hengstler, M. Herzler, D. Knight, M. Leist, U. Norinder, G. Ouédraogo, M. Pastor, S. Stuard, A. White, B. Zdrazil, B. van de Water, and D. Kroese, "Towards grouping concepts based on new approach methodologies in chemical hazard assessment: The read-across approach of the eu-toxrisk project", *Archives of Toxicology*, vol. 93, no. 12, pp. 3643–3667, 2019.

[45] D.-D. Varsou, G. Tsiliki, P. Nymark, P. Kohonen, R. Grafström, and H. Sarimveis, "Toxflow: A web-based application for read-across toxicity prediction using omics and physicochemical data", *Journal of chemical information and modeling*, vol. 58, no. 3, pp. 543–549, 2017.

[46] North Carolina State University, *Comparative toxicogenomics database*, 2020. [Online]. Available: `http://ctdbase.org/detail.go?type=go&acc=GO%3a0003674&view=gene`, (accessed on: September 09, 2020).

[47] J. H. Arts, M. Hadi, M.-A. Irfan, A. M. Keene, R. Kreiling, D. Lyon, M. Maier, K. Michel, T. Petry, and U. G. Sauer, "A decision-making framework for the grouping and testing of nanomaterials (df4nanogrouping)", *Regulatory Toxicology and Pharmacology*, vol. 71, no. 2, S1–S27, 2015.

[48] European Chemicals Agency, *Guidance on information requirements and chemical safety assessment, appendix r.6-1 for nanomaterials applicable to the guidance on qsars and grouping of chemicals*, December 2019. [Online]. Available: `https://echa.europa.eu/documents/10162/23036412/appendix_r6_nanomaterials_en.pdf`, (accessed on: April 08, 2021).

[49] K. Aschberger, D. Asturiol, L. Lamon, A. Richarz, K. Gerloff, and A. Worth, "Grouping of multi-walled carbon nanotubes to read-across genotoxicity: A case study to evaluate the applicability of regulatory guidance", *Computational Toxicology*, vol. 9, pp. 22–35, 2019.

[50] Organization for Economic Cooperation & Development, *Case study on grouping and read-across for nanomaterials genotoxicity of nano-tio2*, September 2018. [Online]. Available: `http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=ENV/JM/MONO(2018)28&docLanguage=En`, (accessed on: February 21, 2019).

[51] M. Goodarzi, B. Dejaegher, and Y. V. Heyden, "Feature selection methods in qsar studies", *Journal of AOAC International*, vol. 95, no. 3, pp. 636–651, 2012.

[52] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2017.

[53] European Chemicals Agency, *Practical guide - how to use and report (q)sars*, ECHA, Helsinki, Finland, 2016. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/0bfe7b84-3386-11e6-969e-01aa75ed71a1/language-en.

[54] OECD, *Guidance document on the validation of (quantitative) structure-activity relationship [(q)sar] models*, Organisation for Economic Co-Operation and Development, Paris, France, 2007. [Online]. Available: https://www.oecd.org/env/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models-9789264085442-en.htm.

[55] T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T. P. Dasari, A. Michalkova, H.-M. Hwang, A. Toropov, D. Leszczynska, and J. Leszczynski, "Using nano-qsar to predict the cytotoxicity of metal oxide nanoparticles", *Nature nanotechnology*, vol. 6, no. 3, p. 175, 2011.

[56] G. Melagraki and A. Afantitis, "Enalos insiliconano platform: An online decision support tool for the design and virtual screening of nanoparticles", *RSC advances*, vol. 4, no. 92, pp. 50 713–50 725, 2014.

[57] K. P. Singh and S. Gupta, "Nano-qsar modeling for predicting biological activity of diverse nanomaterials", *RSC Advances*, vol. 4, no. 26, pp. 13 215–13 230, 2014.

[58] D. Fourches, D. Pu, C. Tassa, R. Weissleder, S. Y. Shaw, R. J. Mumper, and A. Tropsha, "Quantitative nanostructure- activity relationship modeling", *ACS nano*, vol. 4, no. 10, pp. 5703–5712, 2010.

[59] I. Furxhi, F. Murphy, M. Mullins, A. Arvanitis, and C. A. Poland, "Practices and trends of machine learning application in nanotoxicology", *Nanomaterials*, vol. 10, no. 1, p. 116, 2020.

[60] J. Han, M. Kamber, and J. Pei, "3 - data preprocessing", in *Data Mining (Third Edition)*, ser. The Morgan Kaufmann Series in Data Management Systems, J. Han, M. Kamber, and J. Pei, Eds., Third Edition, Boston: Morgan Kaufmann, 2012, pp. 83–124, ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00003-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123814791000034.

[61] A. V. Singh, M. H. D. Ansari, D. Rosenkranz, R. S. Maharjan, F. L. Kriegel, K. Gandhi, A. Kanase, R. Singh, P. Laux, and A. Luch, "Artificial intelligence and machine learning in computational nanotoxicology: Unlocking and empowering nanomedicine", *Advanced Healthcare Materials*, vol. 9, no. 17, p. 1 901 862, 2020.

[62] European Chemicals Agency, *Practical guide - how to use alternatives to animal testing to fulfil your information requirements for reach registration*, ECHA, Helsinki, Finland, 2016. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/e54fc06b-9ffd-11e6-868c-01aa75ed71a1.

[63] A. Golbraikh and A. Tropsha, "Beware of q2!", *Journal of molecular graphics and modelling*, vol. 20, no. 4, pp. 269–276, 2002.

[64] A. Tropsha, P. Gramatica, and V. K. Gombar, "The importance of being earnest: Validation is the absolute essential for successful application and interpretation of qspr models", *QSAR & Combinatorial Science*, vol. 22, no. 1, pp. 69–77, 2003.

[65] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning", *Journal of Analysis and Testing*, vol. 2, no. 3, pp. 249–262, 2018.

[66] M. Daszykowski, B. Walczak, and D. Massart, "Representative subset selection", *Analytica chimica acta*, vol. 468, no. 1, pp. 91–103, 2002.

[67] R. W. Kennard and L. A. Stone, "Computer aided design of experiments", *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[68] M. Daszykowski. (2006). Kennard and stone uniform subset selection (matlab code), [Online]. Available: https://www.researchgate.net/publication/281175342_Kennard_and_Stone_uniform_subset_selection_Matlab_code. (accessed on: September 02, 2020).

[69] NovaMechanics Ltd. (2018). Enalos+ knime nodes - modelling nodes, [Online]. Available: http://www.enalosplus.novamechanics.com/index.php/enalosplusnodes/modelling/. (accessed on: September 02, 2020).

[70] A. Stevens and L. Ramirez-Lopez. (2020). Prospectr: Miscellaneous functions for processing and sample selection of spectroscopic data, [Online]. Available: https://cran.r-project.org/web/packages/prospectr/index.html. (accessed on: September 02, 2020).

[71] K. Roy, R. N. Das, P. Ambure, and R. B. Aher, "Be aware of error measures. further studies on validation of predictive qsar models", *Chemometrics and Intelligent Laboratory Systems*, vol. 152, pp. 18–33, 2016.

[72] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using matthews correlation coefficient metric", *PloS one*, vol. 12, no. 6, e0177678, 2017.

[73] A. Tropsha, "Best practices for qsar model development, validation, and exploitation", *Molecular informatics*, vol. 29, no. 6-7, pp. 476–488, 2010.

[74] K. Roy, "Advances in qsar modeling", *Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences; Springer: Cham, Switzerland*, vol. 555, p. 39, 2017.

[75] S. Zhang, A. Golbraikh, S. Oloff, H. Kohn, and A. Tropsha, "A novel automated lazy learning qsar (all-qsar) approach: Method development, applications, and virtual screening of chemical databases using validated all-qsar models", *Journal of chemical information and modeling*, vol. 46, no. 5, pp. 1984–1995, 2006.

[76] C. D. Walkey, J. B. Olsen, F. Song, R. Liu, H. Guo, D. W. H. Olsen, Y. Cohen, A. Emili, and W. C. Chan, "Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles", *ACS nano*, vol. 8, no. 3, pp. 2439–2455, 2014.

[77] R. Liu, W. Jiang, C. D. Walkey, W. C. Chan, and Y. Cohen, "Prediction of nanoparticles-cell association based on corona proteins and physicochemical properties", *Nanoscale*, vol. 7, no. 21, pp. 9664–9675, 2015.

[78] S. Palchetti, L. Digiacomo, D. Pozzi, G. Peruzzi, E. Micarelli, M. Mahmoudi, and G. Caracciolo, "Nanoparticles-cell association predicted by protein corona fingerprints", *Nanoscale*, vol. 8, no. 25, pp. 12 755–12 763, 2016.

[79] B. Kharazian, N. Hadipour, and M. Ejtehadi, "Understanding the nanoparticle–protein corona complexes using computational and experimental methods", *The international journal of biochemistry & cell biology*, vol. 75, pp. 162–174, 2016.

[80] D. Westmeier, S. K. Knauer, R. H. Stauber, and D. Docter, "Chapter 1 - bio–nano interactions", in *Adverse Effects of Engineered Nanomaterials (Second Edition)*, B. Fadeel, A. Pietroiusti, and A. A. Shvedova, Eds., Second Edition, Academic Press, 2017, pp. 1–12, ISBN: 978-0-12-809199-9. DOI: https://doi.org/10.1016/B978-0-12-809199-9.00001-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012809199900001X.

[81] S. Hänzelmann, R. Castelo, and J. Guinney, "Gsva: Gene set variation analysis for microarray and rna-seq data", *BMC bioinformatics*, vol. 14, no. 1, p. 7, 2013.

[82] V. Forest, J.-F. Hochepied, L. Leclerc, A. Trouvé, K. Abdelkebir, G. Sarry, V. Augusto, and J. Pourchez, "Towards an alternative to nano-qsar for nanoparticle toxicity ranking in case of small datasets", *Journal of Nanoparticle Research*, vol. 21, no. 5, pp. 1–14, 2019.

[83] H. Zhang, Z. Ji, T. Xia, H. Meng, C. Low-Kam, R. Liu, S. Pokhrel, S. Lin, X. Wang, Y.-P. Liao, *et al.*, "Use of metal oxide nanoparticle band gap to develop a predictive paradigm for oxidative stress and acute pulmonary inflammation", *ACS nano*, vol. 6, no. 5, pp. 4349–4368, 2012.

[84] R. Liu, H. Y. Zhang, Z. X. Ji, R. Rallo, T. Xia, C. H. Chang, A. Nel, and Y. Cohen, "Development of structure–activity relationship for metal oxide nanoparticles", *Nanoscale*, vol. 5, no. 12, pp. 5644–5653, 2013.

[85] X. R. Xia, N. A. Monteiro-Riviere, S. Mathur, X. Song, L. Xiao, S. J. Oldenberg, B. Fadeel, and J. E. Riviere, "Mapping the surface adsorption forces of nanomaterials in biological systems", *ACS nano*, vol. 5, no. 11, pp. 9074–9081, 2011.

[86] H. Zhou, Q. Mu, N. Gao, A. Liu, Y. Xing, S. Gao, Q. Zhang, G. Qu, Y. Chen, G. Liu, *et al.*, "A nano-combinatorial library strategy for the discovery of nanotubes with reduced protein-binding, cytotoxicity, and immune response", *Nano letters*, vol. 8, no. 3, pp. 859–865, 2008.

[87] D. Fourches, D. Pu, L. Li, H. Zhou, Q. Mu, G. Su, B. Yan, and A. Tropsha, "Computer-aided design of carbon nanotubes with the desired bioactivity and safety profiles", *Nanotoxicology*, vol. 10, no. 3, pp. 374–383, 2016.

[88] S. Kar, A. Gajewicz, T. Puzyn, and K. Roy, "Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells", *Toxicology in Vitro*, vol. 28, no. 4, pp. 600–606, 2014.

[89] H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, and W. Tong, "Mold2, molecular descriptors from 2d structures for chemoinformatics and toxicoinformatics", *Journal of chemical information and modeling*, vol. 48, no. 7, pp. 1337–1344, 2008.

[90] G. Melagraki and A. Afantitis, "Enalos knime nodes: Exploring corrosion inhibition of steel in acidic medium", *Chemometrics and Intelligent Laboratory Systems*, vol. 123, pp. 9–14, 2013.

[91] P. K. Ojha and K. Roy, "Comparative qsars for antimalarial endochins: Importance of descriptor-thinning and noise reduction prior to feature selection", *Chemometrics and Intelligent Laboratory Systems*, vol. 109, no. 2, pp. 146–161, 2011.

[92] NanoMILE. (2017). Nanomile - engineered nanomaterial mechanisms of interactions with living systems and the environment: A universal framework for samfe nanotechnology, [Online]. Available: `nanomile.eu-vri.eu/`. (accessed on: July 10, 2020).

[93] A. Mikolajczyk, A. Gajewicz, B. Rasulev, N. Schaeublin, E. Maurer-Gardner, S. Hussain, J. Leszczynski, and T. Puzyn, "Zeta potential for metal oxide nanoparticles: A predictive model developed by a nano-quantitative structure–property relationship approach", *Chemistry of Materials*, vol. 27, no. 7, pp. 2400–2407, 2015.

[94] A. R. Gliga, S. Skoglund, I. O. Wallinder, B. Fadeel, and H. L. Karlsson, "Size-dependent cytotoxicity of silver nanoparticles in human lung cells: The role of cellular uptake, agglomeration and ag release", *Particle and fibre toxicology*, vol. 11, no. 1, pp. 1–17, 2014.

[95] D. Lin, X. Tian, F. Wu, and B. Xing, "Fate and transport of engineered nanomaterials in the environment", *Journal of environmental quality*, vol. 39, no. 6, pp. 1896–1908, 2010.

[96] W.-S. Cho, R. Duffin, F. Thielbeer, M. Bradley, I. L. Megson, W. MacNee, C. A. Poland, C. L. Tran, and K. Donaldson, "Zeta potential and solubility to toxic ions as mechanisms of lung inflammation caused by metal/metal oxide nanoparticles", *Toxicological Sciences*, vol. 126, no. 2, pp. 469–477, 2012.

[97] J. M. Berg, A. Romoser, N. Banerjee, R. Zebda, and C. M. Sayes, "The relationship between ph and zeta potential of \30 nm metal oxide nanoparticle suspensions relevant to in vitro toxicological evaluations", *Nanotoxicology*, vol. 3, no. 4, pp. 276–283, 2009.

[98] R. R. Marín, F. Babick, and L. Hillemann, "Zeta potential measurements for non-spherical colloidal particles–practical issues of characterisation of interfacial properties of nanoparticles", *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, vol. 532, pp. 516–521, 2017.

[99] G. V. Lowry, R. J. Hill, S. Harper, A. F. Rawle, C. O. Hendren, F. Klaessig, U. Nobbmann, P. Sayre, and J. Rumble, "Guidance to improve the scientific value of zeta-potential measurements in nanoehs", *Environmental Science: Nano*, vol. 3, no. 5, pp. 953–965, 2016.

[100] B. Michen, C. Geers, D. Vanhecke, C. Endes, B. Rothen-Rutishauser, S. Balog, and A. Petri-Fink, "Avoiding drying-artifacts in transmission electron microscopy: Characterizing the size and colloidal state of nanoparticles", *Scientific reports*, vol. 5, p. 9793, 2015.

[101] M. I. Kotzabasaki, I. Sotiropoulos, and H. Sarimveis, "Qsar modeling of the toxicity classification of superparamagnetic iron oxide nanoparticles (spions) in stem-cell monitoring applications: An integrated study from data curation to model development", *RSC Advances*, vol. 10, no. 9, pp. 5385–5391, 2020.

[102] V. Forest, J.-F. Hochepied, and J. Pourchez, "Importance of choosing relevant biological end points to predict nanoparticle toxicity with computational approaches for human health risk assessment", *Chemical research in toxicology*, vol. 32, no. 7, pp. 1320–1326, 2019.

[103] UPCI.NTUA. (2019). Spions/data preprocess, [Online]. Available: `github.com/ntua-unit-of-control-and-informatics/SPIONs/blob/master/Data_PreProcess_CSV.ipynb`. (accessed on: July 09, 2020).

[104] C. A. Floudas and P. M. Pardalos, *Encyclopedia of optimization, 2nd edition*. Springer Science & Business Media, 2008.

[105] A. Alexandridis, P. Patrinos, H. Sarimveis, and G. Tsekouras, "A two-stage evolutionary algorithm for variable selection in the development of rbf neural network models", *Chemometrics and intelligent laboratory systems*, vol. 75, no. 2, pp. 149–162, 2005.

[106] J. H. Holland *et al.*, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[107] M. Mitchell, *An introduction to genetic algorithms*. MIT press, 1998.

[108] D. A. Coley, *An introduction to genetic algorithms for scientists and engineers*. World Scientific Publishing Company, 1999.

[109] H. A. Orr, "Fitness and its role in evolutionary genetics", *Nature Reviews Genetics*, vol. 10, no. 8, pp. 531–539, 2009.

[110] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, 6th edition*. Garland Science, 2015.

[111] Cambridge Dictionary. (2020). Elitism, [Online]. Available: dictionary.cambridge.org/dictionary/english/elitism. (accessed on: June 09, 2020).

[112] G. A. Chourdakis, "Study and design of data mining methods and applications to metabolomics problems", B.S. thesis, National Technical University of Athens, School of Chemical Engineering, Athens, Greece, 2014.

[113] M. I. Worldwide, "Dynamic light scattering common terms defined", *Inform white paper*, pp. 1–6, 2011.

[114] J. Lim, S. P. Yeap, H. X. Che, and S. C. Low, "Characterization of magnetic nanoparticle by dynamic light scattering", *Nanoscale research letters*, vol. 8, no. 1, p. 381, 2013.

[115] J. Stetefeld, S. A. McKenna, and T. R. Patel, "Dynamic light scattering: A practical guide and applications in biomedical sciences", *Biophysical reviews*, vol. 8, no. 4, pp. 409–427, 2016.

[116] A. Dhawan and V. Sharma, "Toxicity assessment of nanomaterials: Methods and challenges", *Analytical and bioanalytical chemistry*, vol. 398, no. 2, pp. 589–605, 2010.

[117] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens", *Annals of Statistics*, vol. 44, no. 2, pp. 813–852, 2016.

[118] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, "Mathematical programming for piecewise linear regression analysis", *Expert Systems With Applications*, vol. 44, pp. 156–167, 2016.

[119] J. Cardoso-Silva, G. Papadatos, L. G. Papageorgiou, and S. Tsoka, "Optimal piecewise linear regression algorithm for qsar modelling", *Molecular Informatics*, vol. 38, no. 3, p. 1 800 028, 2019.

[120] S. Majumdar and S. C. Basak, "Beware of external validation!-a comparative study of several validation techniques used in qsar modelling", *Current Computer-Aided Drug Design*, vol. 14, no. 4, pp. 284–291, 2018.

[121] J. Löfberg, "Yalmip : A toolbox for modeling and optimization in matlab", in *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004.

[122] J. Löfberg, "Automatic robust convex programming", *Optimization Methods and Software*, vol. 27, no. 1, pp. 115–129, 2012.

[123] Y. Jiang, Y. He, and H. Zhang, "Variable selection with prior information for generalized linear models via the prior lasso method", *Journal of the American Statistical Association*, vol. 111, no. 513, pp. 355–376, 2016.

[124] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[125] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[126] E. Benfenati, "Theory, guidance and applications on qsar and reach", *E-book available at: http://www. orchestra-qsar. eu/documents/333*, 2012.

[127] Y. T. Chau and C. W. Yap, "Quantitative nanostructure–activity relationship modelling of nanoparticles", *Rsc Advances*, vol. 2, no. 22, pp. 8489–8496, 2012.

[128] A. A. Toropov, A. P. Toropova, T. Puzyn, E. Benfenati, G. Gini, D. Leszczynska, and J. Leszczynski, "Qsar as a random event: Modeling of nanoparticles uptake in paca2 cancer cells", *Chemosphere*, vol. 92, no. 1, pp. 31–37, 2013.

[129] C.-Y. Shao, S.-Z. Chen, B.-H. Su, Y. J. Tseng, E. X. Esposito, and A. J. Hopfinger, "Dependence of qsar models on the selection of trial descriptor sets: A demonstration using nanotoxicity endpoints of decorated nanotubes", *Journal of chemical information and modeling*, vol. 53, no. 1, pp. 142–158, 2013.

[130] A. Assarsson, I. Pastoriza-Santos, and C. Cabaleiro-Lago, "Inactivation and adsorption of human carbonic anhydrase ii by nanoparticles", *Langmuir*, vol. 30, no. 31, pp. 9448–9456, 2014.

[131] R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*. John Wiley & Sons, 2009, vol. 41.

[132] M. Hao, Y. Li, Y. Wang, and S. Zhang, "Prediction of pkcθ inhibitory activity using the random forest algorithm", *International journal of molecular sciences*, vol. 11, no. 9, pp. 3413–3433, 2010.

[133] T. Arai and W. Norde, "The behavior of some model proteins at solid-liquid interfaces 1. adsorption from single protein solutions", *Colloids and surfaces*, vol. 51, pp. 1–15, 1990.

[134] Z. Peng, K. Hidajat, and M. Uddin, "Selective and sequential adsorption of bovine serum albumin and lysozyme from a binary mixture on nanosized magnetic particles", *Journal of colloid and interface science*, vol. 281, no. 1, pp. 11–17, 2005.

[135] C. He, Y. Hu, L. Yin, C. Tang, and C. Yin, "Effects of particle size and surface charge on cellular uptake and biodistribution of polymeric nanoparticles", *Biomaterials*, vol. 31, no. 13, pp. 3657–3666, 2010.

[136] M. K. Ha, T. X. Trinh, J. S. Choi, D. Maulina, H. G. Byun, and T. H. Yoon, "Toxicity classification of oxide nanomaterials: Effects of data gap filling and pchem score-based screening approaches", *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[137] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, *et al.*, "Pubchem substance and compound databases", *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2016.

[138] C. Sacchetti, K. Motamedchaboki, A. Magrini, G. Palmieri, M. Mattei, S. Bernardini, N. Rosato, N. Bottini, and M. Bottini, "Surface polyethylene glycol conformation influences the protein corona of polyethylene glycol-modified single-walled carbon nanotubes: Potential implications on biological performance", *ACS nano*, vol. 7, no. 3, pp. 1974–1989, 2013.

[139] S. Schöttler, G. Becker, S. Winzen, T. Steinbach, K. Mohr, K. Landfester, V. Mailänder, and F. R. Wurm, "Protein adsorption is required for stealth effect of poly (ethylene glycol)-and poly (phosphoester)-coated nanocarriers", *Nature nanotechnology*, vol. 11, no. 4, pp. 372–377, 2016.

[140] G. Maiorano, S. Sabella, B. Sorce, V. Brunetti, M. A. Malvindi, R. Cingolani, and P. P. Pompa, "Effects of cell culture media on the dynamic formation of protein-nanoparticle complexes and influence on the cellular response", *ACS nano*, vol. 4, no. 12, pp. 7481–7491, 2010.

[141] R. Vogel, A. K. Pal, S. Jambhrunkar, P. Patel, S. S. Thakur, E. Reátegui, H. S. Parekh, P. Saá, A. Stassinopoulos, and M. F. Broom, "High-resolution single particle zeta potential characterisation of biological nanoparticles using tunable resistive pulse sensing", *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017.

[142] S. Briffa, F. Nasser, E. Valsami-Jones, and I. Lynch, "Uptake and impacts of polyvinylpyrrolidone (pvp) capped metal oxide nanoparticles on daphnia magna: Role of core composition and acquired corona", *Environmental Science: Nano*, vol. 5, no. 7, pp. 1745–1756, 2018.

[143] A. B. Stefaniak, V. A. Hackley, G. Roebben, K. Ehara, S. Hankin, M. T. Postek, I. Lynch, W.-E. Fu, T. P. Linsinger, and A. F. Thünemann, "Nanoscale reference materials for environmental, health and safety measurements: Needs, gaps and opportunities", *Nanotoxicology*, vol. 7, no. 8, pp. 1325–1337, 2013.

[144] T. Wollmann, H. Erfle, R. Eils, K. Rohr, and M. Gunkel, "Workflows for microscopy image analysis and cellular phenotyping", *Journal of biotechnology*, vol. 261, pp. 70–75, 2017.

[145] C. Chomenidis, G. Drakakis, G. Tsiliki, E. Anagnostopoulou, A. Valsamis, P. Doganis, P. Sopasakis, and H. Sarimveis, "Jaqpot quattro: A novel computational web platform for modeling and analysis in nanoinformatics", *Journal of Chemical Information and Modeling*, vol. 57, no. 9, pp. 2161–2172, 2017.

[146] A. Hughes, Z. Liu, M. Raftari, and M. E. Reeves, "A workflow for characterizing nanoparticle monolayers for biosensors: Machine learning on real and artificial sem images", PeerJ PrePrints, Tech. Rep., 2014.

[147] S. Mondini, A. Ferretti, A. Puglisi, and A. Ponti, "Pebbles and pebblejuggler: Software for accurate, unbiased, and fast measurement and analysis of nanoparticle morphology from transmission electron microscopy (tem) micrographs", *Nanoscale*, vol. 4, no. 17, pp. 5356–5372, 2012.

[148] K. Odziomek, D. Ushizima, T. Puzyn, and M. Haranczyk, "Toward quantitative structure activity relationship (qsar) models for nanoparticles", Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2015.

[149] T. Wagner, *Imagej*, 2016. [Online]. Available: `https://imagej.net/Shape_Filter`, (accessed on: December 09, 2020).

[150] Z. Ji, X. Wang, H. Zhang, S. Lin, H. Meng, B. Sun, S. George, T. Xia, A. E. Nel, and J. I. Zink, "Designed synthesis of ceo2 nanorods and nanowires for studying toxicological effects of high aspect ratio nanomaterials", *ACS nano*, vol. 6, no. 6, pp. 5366–5380, 2012.

[151] R. Ghosh Chaudhuri and S. Paria, "Core/shell nanoparticles: Classes, properties, synthesis mechanisms, characterization, and applications", *Chemical reviews*, vol. 112, no. 4, pp. 2373–2433, 2012.

[152] S. Noventa, C. Hacker, D. Rowe, C. Elgy, and T. Galloway, "Dissolution and bandgap paradigms for predicting the toxicity of metal oxide nanoparticles in the marine environment: An in vivo study with oyster embryos", *Nanotoxicology*, vol. 12, no. 1, pp. 63–78, 2018.

[153] MathWorks. (2020). Matlab, the language of technical computing, [Online]. Available: `www.mathworks.com/help/matlab/index.html`. (accessed on: June 01, 2020).

[154] YALMIP. (2020). Solvers, [Online]. Available: `yalmip.github.io/allsolvers/`. (accessed on: June 01, 2020).

[155] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: `https://www.R-project.org`.

[156] V. S. Marwah, G. Scala, P. A. S. Kinaret, A. Serra, H. Alenius, V. Fortino, and D. Greco, "Eutopia: Solution for omics data preprocessing and analysis", *Source code for biology and medicine*, vol. 14, no. 1, p. 1, 2019.

[157] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, *et al.* (2020). Shiny: Web application framework for r, [Online]. Available: `cran.r-project.org/web/packages/shiny/index.html`. (accessed on: May 13, 2020).

[158] D. Attali. (2015). Building shiny apps - an interactive tutorial, [Online]. Available: deanattali.com/blog/building-shiny-apps-tutorial/. (accessed on: May 13, 2020).

[159] RStudio. (2017). Learn shiny, [Online]. Available: shiny.rstudio.com/tutorial/. (accessed on: May 13, 2020).

[160] D. Attali. (2016). Advanced shiny, [Online]. Available: github.com/daattali/advanced-shiny. (accessed on: May 13, 2020).

[161] Python Software Foundation. (2021). Python, [Online]. Available: python.org. (accessed on: January 26, 2021).

[162] J. Forrest, T. Ralphs, S. Vigerske, LouHafer, B. Kristjansson, jpfasano, EdwinStraver, M. Lubin, H. G. Santos, rlougee, and M. Saltzman, *Coin-or/cbc: Version 2.9.9*, version releases/2.9.9, Jul. 2018. DOI: 10.5281/zenodo.1317566. [Online]. Available: https://doi.org/10.5281/zenodo.1317566.

[163] KNIME AG. (2019). Knime analytics platform, [Online]. Available: www.knime.com/knime-analytics-platform. (accessed on: May 14, 2020).

[164] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "Knime-the konstanz information miner: Version 2.0 and beyond", *AcM SIGKDD explorations Newsletter*, vol. 11, no. 1, pp. 26–31, 2009.

[165] Docker Documentation. (2020). Docker overview, [Online]. Available: docs.docker.com/get-started/overview/. (accessed on: May 31, 2020).

[166] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, *et al.*, "Fiji: An open-source platform for biological-image analysis", *Nature methods*, vol. 9, no. 7, pp. 676–682, 2012.

[167] M.-N. Pons, H. Vivier, K. Belaroui, B. Bernard-Michel, F. Cordier, D. Oulhana, and J. A. Dodds, "Particle morphology: From visualisation to measurement", *Powder Technology*, vol. 103, no. 1, pp. 44–57, 1999.

[168] ImageJ, *Imagej*, October 14,2020. [Online]. Available: https://github.com/imagej/imagej-ops/find/master, (accessed on: December 12, 2020).

[169] P. Singh, R. Kumar, B. Sharma, and Y. Prabhakar, "Topological descriptors in modeling malonyl coenzyme a decarboxylase inhibitory activity: N-alkyl-n-(1, 1, 1, 3, 3, 3-hexafluoro-2-hydroxypropylphenyl) amide derivatives", *Journal of Enzyme Inhibition and Medicinal Chemistry*, vol. 24, no. 1, pp. 77–85, 2009.

[170] H. Hong, S. Slavov, W. Ge, F. Qian, Z. Su, H. Fang, Y. Cheng, R. Perkins, L. Shi, and W. Tong, "Mold2 molecular descriptors for qsar", *Statistical modelling of molecular descriptors in QSAR/QSPR*, vol. 2, pp. 65–109, 2012.